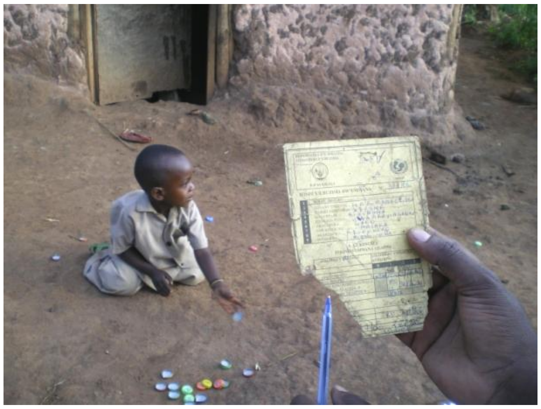
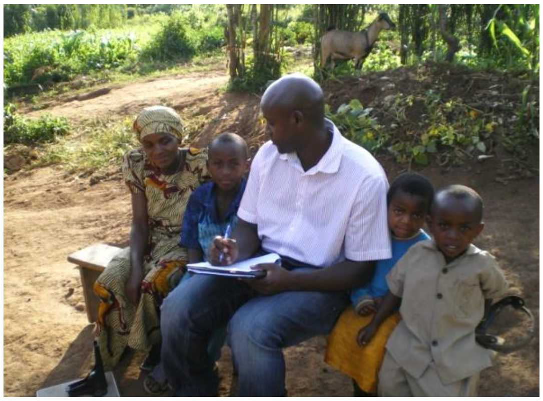
# Impact Evaluation Toolkit

## Measuring the Impact of Results-Based Financing on Maternal and Child Health

Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

Version 1.0. June 2012





www.worldbank.org/health/impactevaluationtoolkit

## Table of Contents

## Acknowledgements

The Impact Evaluation (IE) Toolkit for Results-based Financing (RBF) in Health is a product of the Health, Nutrition and Population Hub (HDNHE) in collaboration with regional RBF teams and the Development Economics Research Group (DECRG).

The authors of the Toolkit are Christel Vermeersch (World Bank TTL), Elisa Rothenbühler (World Bank) and Jennifer Sturdy (Millennium Challenge Corporation). The authors' views expressed in this publication do not necessarily reflect the views of the World Bank.

The authors wish to thank the numerous World Bank staff and consultants, as well as country counterparts, who contributed to the IE Toolkit for RBF in Health. Adam Ross (World Bank), Sebastian Martinez (Inter-American Development Bank) and Paul Gertler (University of California, Berkeley) contributed to initial conceptualization of the Toolkit. Regional Task Team Leaders and investigators, including Tekabe Belay (Afghanistan), Rafael Cortez (Argentina), Adam Ross (Argentina), Paula Giovagnoli (Argentina), Luis Perez (Argentina), Christophe Lemière (Benin), Gaston Sorgho and Paul Jacob Robyn (Cameroon/Central African Republic), Jean-Jacques Frère (Democratic Republic of Congo), Juliette Seban (Democratic Republic of Congo), Dinesh Nair (Nigeria), Paulin Basinga (Rwanda) and Ashis Das (Zambia) shared their regional IE experiences through country spotlights. In particular, we would like to thank Dr. Muhammad Pate, Minister of State for Health, Federal Republic of Nigeria, who gracefully agreed to share his views on how IE can make a difference for policy. Edit Velenyi contributed to interviews and write-up of country spotlights. Andres Martinez (University of Michigan) provided expert contributions on power calculations. Alvaro Canales, Victor Canales, Beatriz Godoy, Juan Muñoz and Mario Navarrete from Sistemas Integrales provided expert inputs into questionnaires, sampling and survey procedures and training materials, and developed the CS-Pro data entry programs. David Peters and Shivam Gupta led the Johns Hopkins University research team who advised on key maternal and child health indicators and measurement, and drafted an early version of the health facility questionnaires and indicators guide. Ashis Das, Jed Friedman, Paul Jacob Robyn, Gaston Sorgho, Monique Vledder, and Damien de Walque piloted the household and health facility questionnaires in Zambia, Cameroon and Central African Republic, and gave extremely valuable feedback to the team. Bianca Brijnath (Monash University) contributed to the development of terms of reference for qualitative work, Logan Brenzel to the creation of terms of reference for costing work and costing questionnaires. Joe Naimoli (USAID) contributed to the conceptual framework for monitoring and documenting RBF programs. Matthew Welch, Olivier Dupriez, Akiko Sagesaka and Mehmood Asghar assisted the team in setting up the RBF data catalog and data access guidelines. Petra Vergeer and Gyuri Fritsche advised on implementation of RBF programs. Alaka Holla, Tony Measham and Petra Vergeer commented earlier drafts of the Toolkit. Darren Dorkin and Rama Lakshminarayanan continuously supported the development of the Toolkit as the program managers for the Health Results Innovation Trust Fund (HRITF). Finally, Matthew Mulligan provided skilled research assistance to the team. The team is also very thankful to peer reviewers who provided insightful comments: Margaret Grosh Christophe Lemière, Mattias Lundberg, Adam Wagstaff, and Feng Zhao. This Toolkit was developed with generous funding from the multi-donor health Results Innovation Trust Fund (HRITF).

**Recommended citation:**

Vermeersch, C., Rothenbühler, E. and J.R. Sturdy, 2012. Impact Evaluation Toolkit: Measuring the impact of Results-based Financing on Maternal and Child Health. Version 1.0. The World Bank, Washington, DC.©

## List of Figures

## List of Tables

## List of Acronyms

| | |
|---|---|
| ANC | Antenatal Care |
| CAFE | Computer Assisted Field Edits |
| CAR | Central African Republic |
| CCT | Conditional Cash Transfer |
| Co-PI | Co-Princial Investigator |
| DC | Dublin Core |
| DDI | Dublin Data Initiative |
| DECRG | Development Economics Research Group (a World Bank unit) |
| DEO | Data Entry Operator |
| DHS | Demographic and Health Surveys |
| EC | Evaluation Coordinator |
| ERB | Ethics Review Board |
| HDNHE | Human Development Network Health Nutrition and Population (a World Bank unit) |
| HF | Health Facility |
| HH | Household |
| HIV/AIDS | Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome |
| HMIS | Health Management Information System |
| HRITF | Health Results Innovation Trust Fund |
| ICC | Intracluster Correlation |
| ID | Identification |
| IE | Impact Evaluation |
| IEC | Independent Ethics Committee |
| IRB | Institutional Review Board |
| MCH | Maternal and Child Health |
| MDG | Millennium Development Goals |
| MOU | Memorandum of Understanding |
| NGO | Non-governmental Organization |
| PBC | Performance Based Contracting |
| PBF | Performance Based Financing |
| PDF | Portable Document Format |
| PI | Principal Investigator |
| PMTCT | Prevention of Mother –to-Child Transmission |
| PPT | Powerpoint© |
| PSU | Primary Sampling Unit |
| RBF | Results-based Financing |

| | |
|---|---|
| STATA© | A statistical analysis program |
| TOR | Terms of Reference |
| TTL | Task Team Leader |
| WHO | World Health Organization |

# Getting Started: What is the Toolkit and How does it Work?

## Background

The Toolkit was developed with funding from the Health Results Innovation Trust Fund (HRITF). The objective of the HRITF is to design, implement and evaluate sustainable results-based financing (RBF) pilot programs that improve maternal and child health outcomes for accelerating progress towards reaching MDGs 1c, 4 & 5. A key element of this program is to ensure a rigorous and well designed impact evaluation is embedded in each country's RBF project in order to document the extent to which RBF programs are effective, operationally feasible, and under what circumstances. The evaluations are essential for generating new evidence that can inform and improve RBF, not only in the HRITF pilot countries, but also elsewhere. The HRITF finances grants for countries implementing RBF pilots, knowledge and learning activities, impact evaluations, as well as analytical work. [1]

The work program on impact evaluation for Results-based financing consists of three pillars:

- Conduct rigorous, prospective impact evaluations on the causal effects of health-related RBF interventions on the access to and quality of service delivery, health expenditures, and health outcomes. In addition, the evaluations may address both the cost-effectiveness and operational complexity of alternative RBF interventions.
- Coordinate and standardize to the extent possible the evaluation methodologies across multiple RBF interventions to facilitate the comparison of alternative approaches, assess the external validity of impacts, and assess the feasibility of similar interventions across different socio-economic and cultural settings.
- Summarize and disseminate the lessons learned in materials that are accessible and relevant to country policy makers and other stakeholders.

## What is the Impact Evaluation Toolkit?

The Impact Evaluation Toolkit is a hands-on guide on how-to design and implement impact evaluations. While many parts of the toolkit can apply to impact evaluation in general, the focus of the toolkit is to help evaluate the impact Results-Based Financing (RBF) projects in the health sector on maternal and child health. The toolkit is structured as follows (see Figure 1):

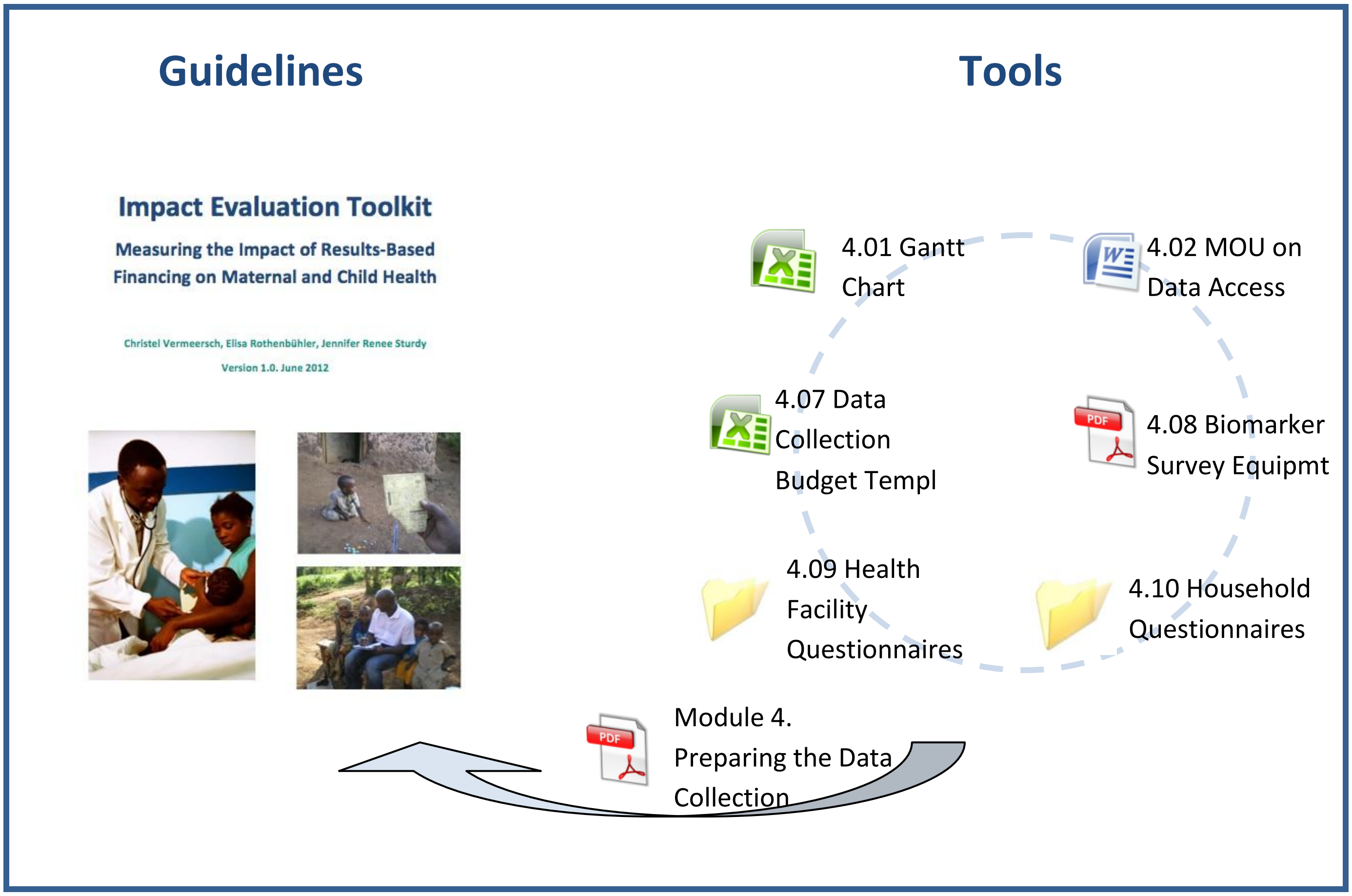
[1] An overview of analytical work is available in the **HRITF Analytical Work Program Overview**.

- For each stage of the impact evaluation (IE) cycle, the Toolkit outlines best-practice procedures in a guiding narrative called ***Guidelines***. Each stage corresponds to one module.
- In each module, the Toolkit provides technical ***Tools*** that can be used to implement the recommendations of the Guidelines. More than 50 tools are included, such as terms of reference for IE team members and survey firms, a list of Maternal and Child Health (MCH) indicators of interest, research protocols, questionnaires, enumerator training manuals and curricula, field work supervision materials, data analysis tools, etc. These standardized tools can facilitate cross-country comparisons of the results of RBF projects.

**If you want to fully use the potential of the Impact Evaluation Toolkit, you need to access, use and adapt the Tools in addition to the following Guidelines. To access both Guidelines and Tools, visit the Impact Evaluation Toolkit website:**

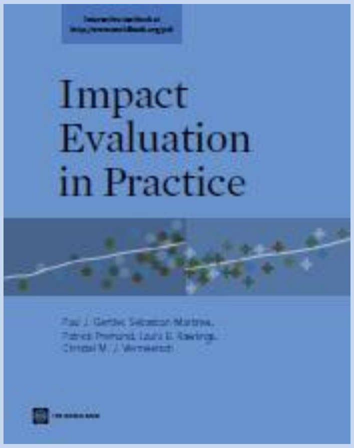
**www.worldbank.org/health/impactevaluationtoolkit**

Figure 1: The Basic Structure of the Toolkit


Guidelines
Tools
Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy
Version 1.0. June 2012
4.01 Gantt Chart
4.02 MOU on Data Access
4.07 Data Collection Budget Templ
4.08 Biomarker Survey Equipmt
4.09 Health Facility Questionnaires
4.10 Household Questionnaires
Module 4. Preparing the Data Collection

The Toolkit is the practical companion piece to the handbook *Impact Evaluation in Practice* (Gertler et al. 2011).

While the handbook delves deeper into the theory of impact evaluation, the Toolkit aims at providing practical guidance and tools for implementers of impact evaluation.



## Who is this Toolkit for?

The Toolkit is intended to support Task Team Leaders (TTLs), principal investigators (Principal Investigators), researchers, survey firms, government stakeholders and other in-country impact evaluation team members as they design and implement impact evaluations.

## What is the Scope of this Toolkit?

**Results-Based Financing and Maternal and Child Health**:

The Toolkit is geared primarily at impact evaluations of RBF projects that focus on improving maternal and child health. In this Toolkit, we use the HRITF definition of RBF "any program that rewards the delivery of one or more outputs or outcomes by one or more incentives, financial or otherwise, upon verification that the agreed-upon result has actually been delivered. Incentives may be directed to service providers (supply side), program beneficiaries (demand side) or both." At this time, the Toolkit does not discuss broader definitions of RBF, such as Diagnosis Related Groups.

While 25% of the content of this Toolkit are specific to RBF and/or maternal and child health, **75% are of general use**. Practitioners with a clear understanding of their field of interest can adapt the RBF and/or maternal and child health specific content to another topic.

**Prospective randomized impact evaluations**: The Toolkit is geared towards teams have already decided they want to implement a prospective impact evaluation. While other methods are available to measure impact (e.g. retrospective impact evaluation) and in certain contexts may turn out to be preferable, such methods are out of the scope of this Toolkit.

**Prospective evaluations**
Prospective evaluations are developed at the same time as the program is being designed and are built into program implementation. Baseline data are collected prior to program implementation for both treatment and comparison groups.

*Impact Evaluation in Practice*, Gertler et al. 2011

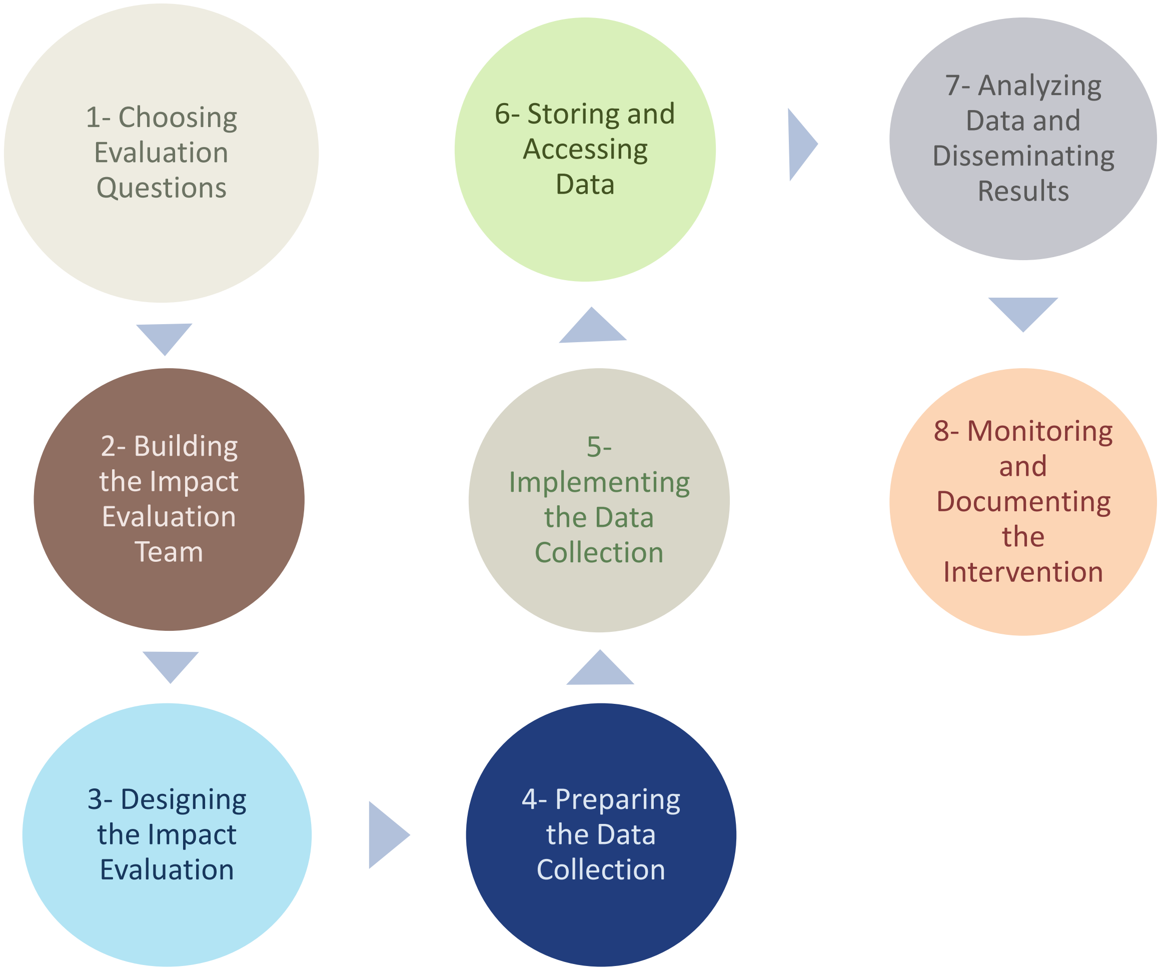
## Modules Overview

The Toolkit contains eight modules that address different stages of the impact evaluation cycle:

**Module 1. Choosing Evaluation Questions**. Each impact evaluation ultimately aims to **inform policy decisions** that will strengthen health systems and improve health status. Defining evaluation questions that are relevant to each country and contribute to the global evidence base on RBF is a key exercise in that regard. Policy questions should not only help understand (i) whether RBF works but also (ii) why RBF works. The theory of change for the RBF intervention frames the policy questions and will be used at the design stage of the impact evaluation.

**Module 2. Building the Impact Evaluation Team** with consideration to qualifications and time commitment. Each IE should be led by a committed and qualified Principal Investigator and either a Co-Principal Investigator or a strong Evaluation Coordinator. Partnering with local researchers can add cultural and institutional sensitivity, perspective and credibility to the analysis and presentation of the results. As an added bonus, these partnerships contribute to building local capacity for leading impact evaluation work in-country. Finally, investigators will need to assess (and if necessary, build) local data collection skill and capacity, as well as identify leads who can carry out complementary activities such as cost analysis and qualitative research activities.

Figure 2: The Eight Modules of the Toolkit


1- Choosing Evaluation Questions
2- Building the Impact Evaluation Team
3- Designing the Impact Evaluation
4- Preparing the Data Collection
5- Implementing the Data Collection
6- Storing and Accessing Data
7- Analyzing Data and Disseminating Results
8- Monitoring and Documenting the Intervention

**Module 3. Designing the IE**

- The IE team will learn how to **build a Results Chain for the RBF intervention** by: (i) identifying and outlining the country specific RBF intervention(s) that will be implemented; (ii) identifying the population that will be targeted by any pilot program; (iii) using the results chain framework to identify input, output, activity and outcome indicators that will be used to assess impact; and (iv) formulating the primary evaluation questions and hypotheses.
- The IE team will learn how to **develop an evaluation strategy**, captured in an Impact Evaluation Design Paper which rigorously identifies the causal impact of the intervention. This involves identifying a treatment and comparison group (or groups) and collecting both baseline and endline data on treatment and comparison groups, defining the inclusion criteria for the sampling frame, and conducting power calculations to identify the appropriate sample size for the study.

**Module 4. Preparing the Data Collection**

- The IE team will need to **develop the IE and Project Gantt Chart** to ensure proper coordination between the intervention and IE activities. The IE team should coordinate with the project design team to ensure that the operational design of the intervention and the evaluation are consistent, and the evaluation is in the context of the operational design. In addition, coordination is required to ensure that the baseline measurement of indicators is collected before the intervention is initiated, and sufficient exposure to the RBF intervention(s) is maintained.
- The IE team will learn about d**eveloping the Research Protocol and Ensuring Ethical Clearance** of the study according to local requirements.
- The IE team will find guidance on **hiring a Survey Firm** with the capacity and experience to manage large-scale, multi-site data collection activities.
- The IE team will learn how to **develop Survey Instruments and Field Procedures** to collect the data. The IE is only as good as the quality of the data collected; therefore survey instruments and field procedures are key factors for determining the quality of the data. Survey instruments have been developed by the HNP hub team for country teams in order to maximize coordination and standardization of measurement across countries. However, the instruments need to be adapted to local culture and institutional environments.

**Module 5. Implementing the Data Collection.** The IE team will learn how to **ensure proper delivery, supervision and reporting** of training, data collection and entry to ensure the survey firm adheres to agreed plans and protocols. In addition, the IE team will need to monitor the timeline to ensure that the IE adheres to the timeline agreed with the Government counterparts.

**Module 6. Storing and Accessing Data.** The IE team will learn about developing a **Data Documentation, Storage and Access Plan** with project design and IE teams in order to guarantee safety, confidentiality and documentation of the data.

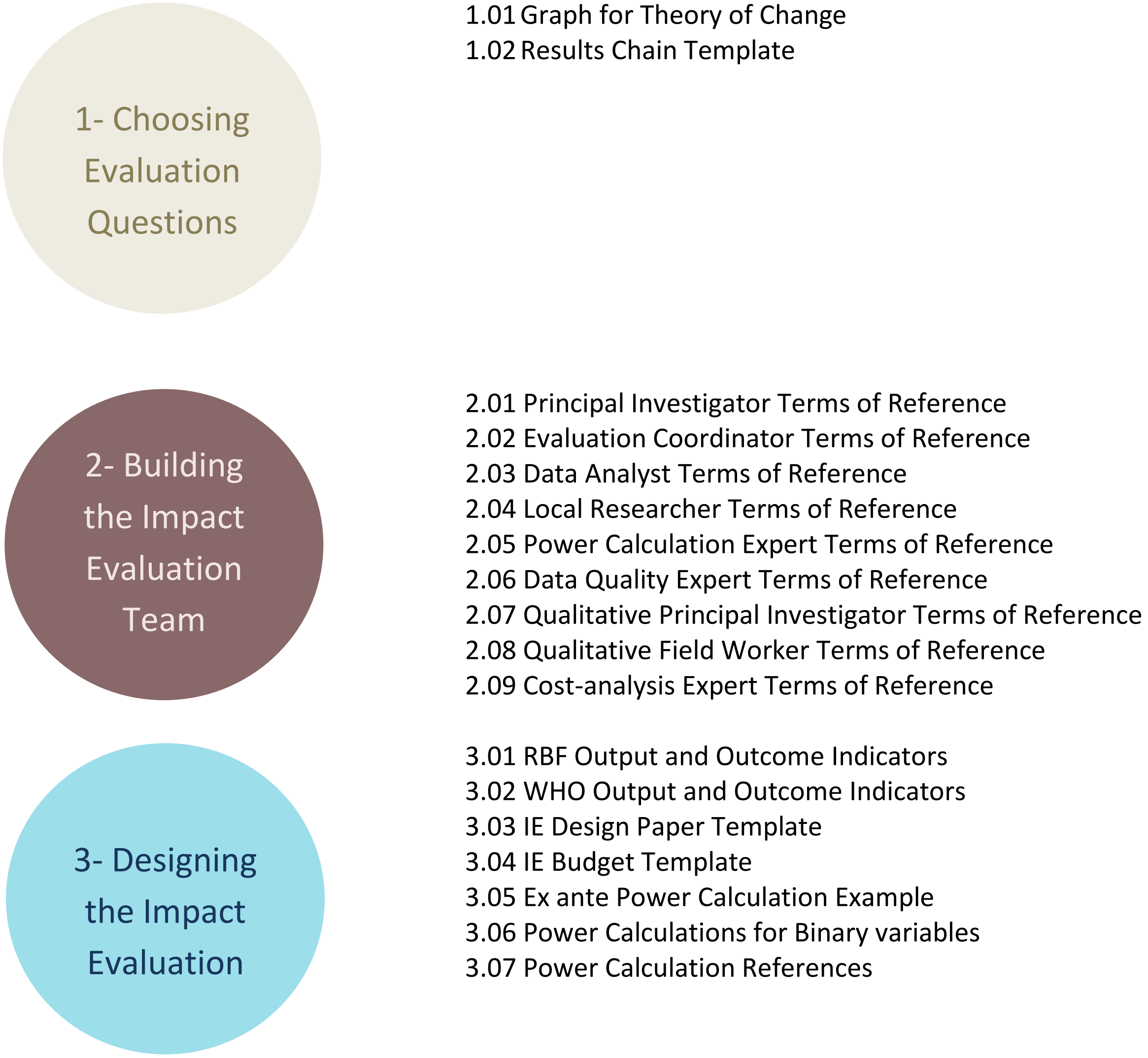
**Module 7. Analyzing Data and Disseminating Results.** The IE team will learn about developing a **Data Analysis and Dissemination Plan** in order to ensure timely dissemination of descriptive and analytical products.

**Module 8. Monitoring and Documenting the Intervention.** The team will learn about developing a **Monitoring and Documentation Plan** in order to monitor project implementation, adherence to evaluation design and assignment to treatment and comparison groups, as well as identify complementary data sources, such as Health Management Information Systems (HMIS), financial and administrative data.

## Tools Overview

Each module contains Tools to help implement the corresponding stage in the impact evaluation. When a tool is mentioned in the guidelines of the Toolkit, it is flagged with a **bold red font**.

### Figure 3: The Tools



1- Choosing Evaluation Questions

1.01 Graph for Theory of Change
1.02 Results Chain Template

2- Building the Impact Evaluation Team

2.01 Principal Investigator Terms of Reference
2.02 Evaluation Coordinator Terms of Reference
2.03 Data Analyst Terms of Reference
2.04 Local Researcher Terms of Reference
2.05 Power Calculation Expert Terms of Reference
2.06 Data Quality Expert Terms of Reference
2.07 Qualitative Principal Investigator Terms of Reference
2.08 Qualitative Field Worker Terms of Reference
2.09 Cost-analysis Expert Terms of Reference

3- Designing the Impact Evaluation

3.01 RBF Output and Outcome Indicators
3.02 WHO Output and Outcome Indicators
3.03 IE Design Paper Template
3.04 IE Budget Template
3.05 Ex ante Power Calculation Example
3.06 Power Calculations for Binary variables
3.07 Power Calculation References

## 4- Preparing the Data Collection

## 5- Implementing the Data Collection

## 6- Storing and Accessing Data

7- Analyzing Data and Disseminating Results

7.01 Household Baseline Report
7.01a Handbook Household Baseline Report
7.01b Rwanda Household Baseline Outcome Indicators
7.01c Rwanda Household Baseline Report Do-Files (STATA)
7.01d Rwanda Household Baseline Ex-post Power Calculations
7.02 Health Facility Baseline Report
7.02a Suggested detailed outline of health facility baseline report
7.03 Community Health Worker (CHW) Baseline Report
7.03a Rwanda CHW Baseline Report Do-Files (STATA)
7.04 STATA ado file for Baseline balance table
7.05 WHO Anthro calculation package
7.06 STATA training
7.07 STATA Training IE Design Validation

8- Monitoring and Documenting the Intervention

8.01 Monitoring Indicators Rwanda Example
8.02 Field Supervision Visit Templates Rwanda

## Drawing Experience from other Countries

Throughout the toolkit, **Country Spotlights** illustrate real challenges and lessons learned in actual impact evaluations of RBF programs. Most of the Country Spotlights originated in impact evaluations that were financed by the HRITF, though the toolkit also includes other interesting cases. The spotlights were developed in collaboration with project Task Team Leaders and impact evaluation teams. The toolkit guidelines only contain extracts of the Spotlights. The spotlights are featured in their entirety in the “Country Spotlights” section of the Toolkit website.

## Adapting Tools to Country Needs

**Country specific content** in the tools of the Toolkit is highlighted so IE teams can easily adapt their content. Country-specific content is flagged using either red font (in questionnaires) or yellow highlighted font (in most other tools).

## How to Prioritize the Recommendations of the Toolkit

The Toolkit hopes to set standards of quality and scientific rigor by providing a comprehensive set of recommendations and tools. However, real-life conditions, budgets, country dialogue and context influence the feasibility of certain recommendations. Table 1 summarizes the main recommendations highlighted throughout the Toolkit. It aims to help IE teams prioritize between what is:

- Critical: what we believe should really be included or considered in the design and implementation of impact evaluations
- Important: what should ideally be included or considered, but could be revised or adapted if necessary
- Nice to have: what we encourage IE teams to include, but could be omitted if necessary

**Table 1: List of Recommendations**

| Module | Recommendations | Critical | Important | Nice to have |
|---|---|---|---|---|
| 1 | - The relevance of the chosen policy/evaluation questions, both locally and globally, matters more than the number of questions addressed. Not all dimensions of RBF can be explored in a single impact evaluation, so team will need to prioritize questions. | ✓ | | |
| 1 | - Understanding whether RBF works is a first step. Understanding the reasons for failure or success of the RBF program is key to improving it and ensuring its sustainability. | | ✓ | |
| 2 | - Team member(s) primarily responsible for project design and implementation (e.g. the TTL) should not serve as the principal investigator. | ✓ | | |
| 2 | - The principal investigator and evaluation coordinator play a crucial role in supervising the survey firm(s). | ✓ | | |
| 2 | - Local research counterparts can greatly contribute to the success of the impact evaluation, because they can bring local knowledge and foster country ownership of the program. | | ✓ | |
| 2 | - Teams should assess local capacity to conduct surveys and identify whether any technical support will be needed to ensure the quality of survey data. | | ✓ | |
| 2 | - A data quality expert can help set up the right initial conditions for ensuring the quality of survey data before the survey firm goes into the field. A local supervisor can verify the data quality assurance processes during the implementation of the surveys. | | ✓ | |
| 2 | - Qualitative and cost effectiveness analysis can add great richness and granularity to the questions that the impact evaluation will answer. | | ✓ | |
| 2 | - Impact evaluations involve several rounds of sophisticated data – a good data analyst will help the team manage and analyze the data quickly and reliably. | | ✓ | |
| 2 | - While power calculations can be the responsibility of the principal investigator, a power calculation expert may have more time and expertise to dedicate to this task. | | | ✓ |
| 3 | - A prospective impact evaluation should be designed prior to or simultaneously with the intervention. | ✓ | | |
| 3 | - Teams should develop a results framework for the RBF project to identify the main pathway(s) by which the RBF program's activities will affect key outputs and outcomes. | ✓ | | |
| 3 | - The recommended identification strategy for the RBF Impact Evaluations is randomized assignment to | | ✓ | |

| Module | Recommendations | Critical | Important | Nice to have |
|---|---|---|---|---|
| | intervention(s) and comparison groups. | | | |
| 3 | • Teams should assess present and future threats to the internal validity of the evaluation (e.g. contamination, lack of power) and monitor them over time. | ✓ | | |
| 3 | • Power calculations are an important part of the design of an impact evaluation. Without sufficient power, the impact evaluation may not be able to answer key policy questions. The sample size must allow for sufficient power. | ✓ | | |
| 3 | • The sample must be representative of the population that will ultimately benefit from the program. | ✓ | | |
| 3 | • When country counterparts buy into the concept of the impact evaluation and understand the importance of respecting the arms of the study, it will be easier to successfully keep treatment and comparison groups intact until the follow-up survey. | ✓ | | |
| 3 | • The choice of indicators for the study is critical – each indicator should be measurable with the chosen data collection instruments. | ✓ | | |
| 3 | • Teams can refer to *Impact Evaluation in Practice* (Gertler et al. 2011) for in depth discussion on appropriate identification strategies for impact evaluation. | | ✓ | |
| 3 | • When deciding on the unit of randomization, teams are balancing the power of the impact evaluation and the risk of contamination across randomization units. | | ✓ | |
| 4 | • The research protocol should contain all relevant information related to the protection of human subjects, including specific sampling criteria, informed consent and data confidentiality protocols. | ✓ | | |
| 4 | • The impact evaluation must be approved by an Institutional Board: the Principal Investigator should plan for contracting this board to conduct the ethical review and approve the research prior to the beginning of field activities. | ✓ | | |
| 4 | • A Project/impact evaluation Gantt Chart can help teams coordinate activities and timelines from the project and from the impact evaluation. | | ✓ | |
| 4 | • The impact evaluation team should agree with Government counterparts what will be the policy of accessing the data from the impact evaluation. A written Memorandum of Understanding can help prevent misunderstandings. | | ✓ | |
| 4 | • The decision between CAFE and field-based data entry has major implications for the selection of the survey firm and should be decided in advance of survey firm procurement. | | ✓ | |

| Module | Recommendations | Critical | Important | Nice to have |
|---|---|---|---|---|
| 4 | • Hiring a survey firm is a time intensive process, which typically requires 3-6 months and should be initiated in the early stages of project planning. | ✓ | | |
| 4 | • Depending on the situation and expertise in country, it may be preferable to hire one survey firm that would conduct both health facility and household surveys, or for two separate firms. As a general rule, we recommend that teams use a competitive selection process. | | ✓ | |
| 4 | • The survey management team should include a Project Manager, a Field Manager and a Data Manager during the full duration of the preparation and implementation of the data collection. | ✓ | | |
| 4 | • Negotiations with the survey firm require a clear understanding of budget and time constraints, which have implications for field team composition and survey duration. | ✓ | | |
| 4 | • The survey firm should be supported from the early stages of survey preparation by a data quality expert, especially in local survey firms with limited capacity. | | ✓ | |
| 4 | • The structure and quality of the survey instruments are crucial for data quality and comparability of results across countries. We recommend that project teams use the RBF Facility and Household questionnaires as a basis. The Principal Investigator of the evaluation should determine which modules are appropriate and which are not, and ensure key outcomes of interest can be calculated from the questionnaires. Teams should feel free to make the adjustments that they deem necessary. | | ✓ | |
| 4 | • The toolkit questionnaires are meant to be comprehensive – teams may want to limit the number of modules to limit the cost and time requirement for administering the questionnaires. | | ✓ | |
| 4 | • Community surveys can allow measuring infrastructures and existing support networks within the community. They can also be used as a complement to household surveys, especially when household surveys need to be drastically shortened. | | | ✓ |
| 5 | • The impact evaluation team and survey firm should define the protocol for uniquely identifying observations in the data bases, as well as linking across databases. | ✓ | | |
| 5 | • The impact evaluation team should define the protocol for identifying the treatment and comparison areas within the databases. | ✓ | | |
| 5 | • The quality and duration of the training of field teams are key to the success of data collection. | ✓ | | |
| 5 | • While survey firms are in charge of data collection, the impact evaluation team should work with the survey firm to ensure appropriate and timely reporting on field work. | ✓ | | |

| Module | Recommendations | Critical | Important | Nice to have |
|---|---|---|---|---|
| 5 | • The research protocol and survey manuals should contain all the information needed by the survey firms to ensure data collection is conducted ethically and according to plans. | ✓ | | |
| 5 | • The safety and confidentiality of the data collected should be safeguarded carefully during data collection and entry. Field teams should report any logistical or security challenge. | ✓ | | |
| 5 | • The impact evaluation team should closely monitor the quality of data collection and data entry, and may want to hire a data quality expert to help in this process. | | ✓ | |
| 5 | • Local survey firms may have limited capacity in data entry programming, entry and management. The Toolkit contains data entry forms for CS-Pro software that correspond to the household and health facility questionnaires in the toolkit. | | ✓ | |
| 5 | • It is preferable to enter the data concurrently with field work, rather than after its completion. | | | ✓ |
| 6 | • The TTL should plan for and coordinate comprehensive and complete documentation of impact evaluation activities.<br>▸Include updated Concept Note, Research Protocol, Questionnaires, Training Manuals, etc.<br>▸Decide on what information needs to be removed for respondent confidentiality. | ✓ | | |
| 6 | • The Principal Investigator should prepare (a) separate ID control file(s) that establishes the link between the geographical ID codes and the field ID codes. | ✓ | | |
| 6 | • The Principal Investigator should decide on any variables that cannot be released publicly (e.g. sensitive personal information). | ✓ | | |
| 6 | • Confidential files (ID control file and other non publicly available data) should be stored in a secure location, preferably a data enclave. | | ✓ | |
| 6 | • Impact evaluation teams should allocate sufficient time for documenting and uploading the data, in order to guarantee data access continuity within the team, ease future data sharing and analysis process | | ✓ | |
| 6 | • Impact evaluation teams should refer to the Memorandum of Understanding (or other data sharing agreement) when documenting, storing and sharing the data. | | ✓ | |
| 7 | • Data analysts should keep a record of any alteration and statistical analysis performed on the data. | ✓ | | |
| 7 | • The original data must absolutely be kept intact. Any alteration must be saved as a different dataset. | ✓ | | |
| 7 | • Prior to baseline data analysis, the data analyst should refer to international and national guidelines on | | ✓ | |

| Module | Recommendations | Critical | Important | Nice to have |
|---|---|---|---|---|
| | how to calculate indicators. (eg. WHO) | | | |
| 7 | • The data analyst can help identify errors that occurred during baseline data collection or entry. This can then allow for adjustments in training and supervision during future rounds of data collection. | | | ✓ |
| 7 | • Data cleaning, analysis and dissemination of results take time. It helps to plan ahead in terms of manpower and funds. | | | ✓ |
| 7 | • Ex-post power calculations are a part of the internal validity checks of the impact evaluation. If need be, they can recommend ways to increase power at follow-up. | | | ✓ |
| 7 | • The analysis should be developed keeping in mind the best way of ultimately disseminating results and informing policymakers. | | | ✓ |
| 7 | • Impact evaluation data are typically very rich: while analyzing the impact of RBF may be the primary goal, other analyses can be conducted to inform policymaking. | | | ✓ |
| 8 | • Monitoring and documenting project activities are a crucial complement to the impact evaluation because they provide information on the actual interventions on the ground, and therefore, on the intervention that is being evaluated. | | ✓ | |
| 8 | • Impact evaluation teams will want the program to identify two major risks to the impact evaluation: (1) compensation of the comparison group through an alternative intervention or program; and (2) imitation of the treatment by the comparison group. | | ✓ | |

# Module 1

# Choosing Evaluation Questions

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

# Module 1. Choosing Evaluation Questions

| Main Recommendations and Available Tools for this Module | | | |
|---|---|---|---|
| **Recommendations** | **Critical** | **Important** | **Nice to have** |
| • The relevance of the chosen policy/evaluation questions, both locally and globally, matters more than the number of questions addressed. Not all dimensions of RBF can be explored in a single impact evaluation, so team will need to prioritize questions. | ✓ | | |
| • Understanding whether RBF works is a first step. Understanding the reasons for failure or success of the RBF program is key to improving it and ensuring its sustainability. | | ✓ | |

| **Tools** |
|---|
| • 1.01 Editable Graph for Theory of Change<br>• 1.02 Results Chain Template |

## Module Contents

## The Promise of Results-based Financing and the Evidence Gap

The World Bank's 2007 Health, Nutrition and Population (HNP) Strategy renewed the World Bank's focus on results and on strengthening health systems. A key objective in the HNP Strategy is to tighten the links between lending and results through increased use of Results Based Financing (RBF).

Musgrove (2010) defines RBF for Health as "*any program that rewards the delivery of one or more outputs or outcomes by one or more incentives, financial or otherwise, upon verification that the agreed-upon result has actually been delivered. Incentives may be directed to service providers (supply side), program beneficiaries (demand side) or both […] Verification that results were actually obtained is an essential feature. The ideal is perhaps for verification to be undertaken by a neutral third party, even if the principal pays the corresponding costs, but many arrangements are possible. Ex ante verification (before payment) can be complemented by ex-post assessment.*"

RBF in the health sector is viewed as a powerful, yet still largely unproven, tool to strengthen health systems and accelerate progress towards the health MDGs (Levine and Eichler 2009, Cochrane 2012). While there is a strong base of evidence on the positive impacts of Conditional Cash Transfer (CCT) programs on human development outcomes (Fiszbein and Schady 2009) and quite some evidence on the impact of demand-side vouchers for health services (Meyer et al. 2011), there is very little evidence available on the impact of supply-side RBF interventions or non-CCT demand-side interventions on health indicators in low-income countries.[2] In fact, to date we are aware of only a few case-controlled impact evaluations of programs that provide financial incentives to health care providers in low and middle income countries, though a number of other studies present promising non-experimental results. The boxes below present the abstracts from a few of those impact evaluations and evidence reviews.

Even though evidence is lacking for some types of RBF interventions, in the last five years, numerous countries have started or scaled up these interventions. This offers a unique opportunity to invest in rigorous and well-designed impact evaluations that document the extent to which health-related RBF policies are effective, are operationally feasible, and under what circumstances. With a well-coordinated RBF impact evaluation agenda, the evidence generated can be used by countries and donors to make well-informed policy decisions.

[2] There is, however, a growing literature on P4P for medical care in the U.S. and the U.K. See for example Fleetcroft et al (2012), Jha et al (2012), Lindenauer (2007), Doran et al (2006), and Perterson et al (2006).

**Paying for performance to improve the delivery of health interventions in low- and middle-income countries: a Cochrane Review**

***"Background*** *There is a growing interest in paying for performance as a means to align the incentives of health workers and health providers with public health goals. However, there is currently a lack of rigorous evidence on the effectiveness of these strategies in improving health care and health, particularly in low- and middle-income countries. Moreover, paying for performance is a complex intervention with uncertain benefits and potential harms. A review of evidence on effectiveness is therefore timely, especially as this is an area of growing interest for funders and governments.*

***Objectives*** *To assess the current evidence for the effects of paying for performance on the provision of health care and health outcomes in low and middle-income countries.*

*[…]*

***Authors' conclusions*** *The current evidence base is too weak to draw general conclusions; more robust and also comprehensive studies are needed. Performance based funding is not a uniform intervention, but rather a range of approaches. Its effects depend on the interaction of several variables, including the design of the intervention (e.g. who receives payments, the magnitude of the incentives, the targets and how they are measured), the amount of additional funding, other ancillary components such as technical support, and contextual factors, including the organisational context in which it is implemented."*

Witter et al.(2012)

### Performance Based Financing, Evidence from Rwanda

*“**Background** Evidence about the best methods with which to accelerate progress towards achieving the Millennium Development Goals is urgently needed. We assessed the effect of performance-based payment of health-care providers (payment for performance; P4P) on use and quality of child and maternal care services in health-care facilities in Rwanda.*

***Methods** 166 facilities were randomly assigned at the district level either to begin P4P funding between June, 2006, and October, 2006 (intervention group; n=80), or to continue with the traditional input-based funding until 23 months after study baseline (control group; n=86). Randomisation was done by coin toss. We surveyed facilities and 2158 households at baseline and after 23 months. The main outcome measures were prenatal care visits and institutional deliveries, quality of prenatal care, and child preventive care visits and immunisation. We isolated the incentive effect from the resource effect by increasing comparison facilities’ input-based budgets by the average P4P payments made to the treatment facilities. We estimated a multivariate regression specification of the difference-in-difference model in which an individual’s outcome is regressed against a dummy variable, indicating whether the facility received P4P that year, a facility-fixed effect, a year indicator, and a series of individual and household characteristics. Findings Our model estimated that facilities in the intervention group had a 23% increase in the number of institutional deliveries and increases in the number of preventive care visits by children aged 23 months or younger (56%) and aged between 24 months and 59 months (132%). No improvements were seen in the number of women completing four prenatal care visits or of children receiving full immunisation schedules. We also estimate an increase of 0·157 standard deviations (95% CI 0·026–0·289) in prenatal quality as measured by compliance with Rwandan prenatal care clinical practice guidelines.*

***Interpretation** The P4P scheme in Rwanda had the greatest effect on those services that had the highest payment rates and needed the least effort from the service provider. P4P financial performance incentives can improve both the use and quality of maternal and child health services, and could be a useful intervention to accelerate progress towards Millennium Development Goals for maternal and child health.”*

Basinga et al. (2011)

### Performance-based financing – Evidence from Rwanda (II)

*“This study examines the impact of performance incentives for health care providers in Rwanda on child health outcomes using a prospective quasi-experimental design that was nested into the program roll-out. We find that the P4P scheme had a large and significant effect on the weight-for-age of children 0-11 months and on the height-for-age of children 24-49 months (0.53 and 0.25 std dev respectively). We attribute this improvement to increases in the quantity of well-child care as well as improvements in the quality of prenatal care. Consistent with economic theory, we find larger effects in aspects of service that are in the control of providers, and in those where the monetary rewards were higher. We argue that changes in provider effort were the main driver of the observed impacts. We find a 20 percent reduction in the knowledge to practice efficiency gap for prenatal care. Finally, we find evidence of a strong complementarity between the P4P scheme and the presence of high-skill health workers in the health centers.”*

Gertler and Vermeersch (2012)

**Contracting for Health – Evidence from Cambodia**

*"In 1999, Cambodia contracted out management of government health services to NGOs in five districts that had been randomly made eligible for contracting. The contracts specified targets for maternal and child health service improvement. Targeted outcomes improved by about 0.5 standard deviations relative to comparison districts. Changes in non-targeted outcomes were small. The program increased the availability of 24-hour service, reduced provider absence, and increased supervisory visits. There is some evidence it improved health. The program involved increased public health funding, but led to roughly offsetting reductions in private expenditure as residents in treated districts switched from unlicensed drug sellers and traditional healers to government clinics."*

Bloom et al. (2006)

**Incentivizing villages to improve health and education – Evidence from Indonesia**

*"This paper reports an experiment in over 3,000 Indonesian villages designed to test the role of performance incentives in improving the efficacy of aid programs. Villages in a randomly-chosen one-third of subdistricts received a block grant to improve 12 maternal and child health and education indicators, with the size of the subsequent year's block grant depending on performance relative to other villages in the subdistrict. Villages in remaining subdistricts were randomly assigned to either an otherwise identical block grant program with no financial link to performance, or to a pure control group. We find that the incentivized villages performed better on health than the non-incentivized villages, particularly in less developed provinces, but found no impact of incentives on education. We find no evidence of negative spillovers from the incentives on untargeted outcomes. Incentives led to what appear to be more efficient use of block grants, and led to an increase in labor from health providers, who are partially paid fee-for-service, but not teachers. On net, between 50-75% of the total impact of the block grant program on health indicators can be attributed to the performance incentives."*

Olken et al. (2011)

**Incentives tied to provider performance – Evidence from the Philippines**

*"The merits of using financial incentives to improve clinical quality have much appeal, yet few studies have rigorously assessed the potential benefits. The uncertainty surrounding assessments of quality can lead to poor policy decisions, possibly resulting in increased cost with little or no quality improvement, or missed opportunities to improve care. We conducted an experiment involving physicians in thirty Philippine hospitals that overcomes many of the limitations of previous studies. We measured clinical performance and then examined whether modest bonuses equal to about 5 percent of a physician's salary, as well as system-level incentives that increased compensation to hospitals and across groups of physicians, led to improvements in the quality of care. We found that both the bonus and system-level incentives improved scores in a quality measurement system used in our study by ten percentage points. Our findings suggest that when careful measurement is combined with the types of incentives we studied, there may be a larger impact on quality than previously recognized."*

Peabody et al. (2011)

**The impact of vouchers on the use and quality of health goods and services in developing countries: A systematic review**

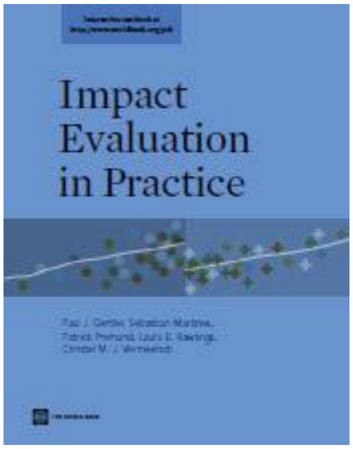
***"Background:*** *One approach to delivering health assistance to developing countries is the use of health voucher programmes, where vouchers are distributed to a targeted population for free or subsidised health goods/services. Theoretically, vouchers are expected to successfully target specific populations, increase utilisation, improve quality, enhance efficiency, and ultimately improve the health of populations. Objectives: The primary objective of this systematic review is to assess whether voucher programmes thus far have been successful in achieving these desired outcomes. Methods: Using explicit inclusion/exclusion criteria, a search of bibliographic databases, key journals, and organisational websites were conducted in September – October 2010. Other search strategies used include bibliographic backreferencing, supplemental keyword searches using specific programme information, and contacting key experts in the field. A narrative synthesis approach was taken to qualitatively summarise the identified quantitative outcome variables in five categories (targeting, utilisation, efficiency, quality, and health impact). Using the direction of effect of outcome variables and the confidence in the study findings, the findings for each category of outcomes were aggregated and assigned to one of five pre-established conclusion categories: (1) insufficient evidence; (2) evidence of no effect; (3) conflicting evidence; (4) modest evidence of effect; or (5) robust evidence of effect. Sub-group and sensitivity analyses were also performed. A quantitative meta-analysis was not conducted due to the heterogeneous natures of the outcome variables reviewed.*

***Results:*** *A total of 24 studies evaluating 16 different health voucher programmes were identified in this review. The findings from 64 outcome variables informed five main conclusions: (1) there is modest evidence that voucher programmes effectively target voucher for health goods/services to specific populations (based on four programmes); (2) there is insufficient evidence to determine whether voucher programmes deliver health goods/services more efficiently than competing health financing strategies (based on one programme); (3) there is robust evidence that voucher programmes increase utilisation of health goods/services (based on 13 programmes); (4) there is modest evidence that voucher programmes improve the quality of health services (based on three programmes); and (5) the evidence indicates that voucher programmes do not have an impact on the health of populations (based on six programmes); however, this last conclusion was found to be unstable in a sensitivity analysis.*

***Conclusions:*** *The evidence indicates that health voucher programmes have been successful in increasing utilisation of health goods/services, targeting specific populations, and improving the quality of services. While these results are encouraging, the subsequent link that voucher programmes improve the health of the population is not evident in the data analysed in this review. The methodology used in this analysis allows policy-makers to synthesise evidence from heterogeneous studies and therefore include more data than could be used in a standard meta-analysis. However, vouchers are still relatively new and the number of published studies evaluating vouchers is a limitation. Future reviews using this methodology can compare health voucher programmes to competing financing techniques and incorporate new evidence on voucher programmes for evaluations currently underway; however, the synthesis tools used in this review should be validated."*

Meyer et al. (2011)

## What is Impact Evaluation?[3]



Impact Evaluations are part of a broader agenda of evidence-based policy making. In a context in which policy makers, donors and civil society are demanding results and accountability from public programs, impact evaluation can provider robust and credible evidence on performance and, crucially, on whether a particular program achieved its desired outcomes. Globally, impact evaluations help build knowledge on the effectiveness of programs.

Impact evaluation is one among a range of methods that support evidence-based policy. Other methods include monitoring, process evaluations, qualitative assessments and costing. Impact evaluation is particular in that it seeks to assess the changes in well-being that can be attributed or are caused by a particular program or policy. Unlike monitoring and evaluation, impact evaluation is generally structured around one type of question *What is the impact (or causal effect) of a program on an outcome of interest?* In contrast to before/after comparisons and simple end-user satisfaction surveys, impact evaluation aims to isolate the impact of the program from other confounding factors.

## Why Evaluate RBF Programs?

Impact evaluations are especially useful when countries test out innovative, new interventions that seem promising in theory but for which we have little hard evidence. Policy makers who want to use evidence to back their policies need information on a variety of questions, such as *Is this program effective compared to the current situation? Of the many ways in which an RBF program can be implemented, which one is the most effective one?*

An impact evaluation of a country RBF program provides evidence on whether that particular intervention worked in that particular country context. Taken together, evidence from impact evaluations that examine various RBF mechanisms in various countries can inform Governments and partners how to effectively design and use RBF mechanisms to improve health system functioning and health outcomes in a range of contexts. In addition, IEs can help determine whether RBF has any unintended consequences, such as encouraging providers to shift their attention away from delivering services that are not included in the RBF mechanism. Finally, IEs can help document the costs associated with administering payment systems that are based on results.

[3] This section is based heavily on chapter 1 of *Impact Evaluation in Practice* (Gertler et al. 2011). Please refer to this manual for a more extensive discussion. The book can be downloaded at www.worldbank.org/ieinpractice free of charge.

### Country Spotlight: Motivation for Impact Evaluation
### Nigeria: Showing Results to Leverage Funding[4]

**Dr. Pate:** In my previous office, as Executive Director of the National Primary Health Care Development Agency (NPHCDA) –which is a federal parastatal agency responsible for primary care delivery across all 36 states in Nigeria– it was clear that more resources and innovation will be required to put the health MDG targets back on track in a country that has a population of close to 150 million people and has some of the worst MCH indicators. Since the NPHCDA is mandated to provide coverage to everyone for all essential care and the health MDG targets are heavily driven by the strength of the primary delivery system, we knew that innovations, which would lead us to more effective care and efficient use of resources were needed to propel the country toward better maternal and child health outcomes. However, no matter if we sought support for increased domestic budget allocations to the primary sector or sought funding from development agencies, we faced the same questions – *Could we show results? Could we show impact? Could we prove that we were getting good value for the money, whether it is from a domestic or international source?* We soon realized that we needed credible results for government budget allocations and official development assistance, which included loans and grants. The need for solid evidence and results has increased for both governments and donor agencies because of budget pressures and fiscal strains during the economic crisis.

## Determining Evaluation Questions

The initial step in setting up any evaluation is to establish the type of question to be answered by the evaluation, constructing a theory of change that outlines how the project is supposed to achieve the intended results, developing a results chain, formulating hypotheses to be tested by the evaluation, and selecting performance indicators (Gertler et al. 2011).

A theory of change is a description of how an intervention is supposed to deliver the desired results. It describes the causal logic of how and why a particular project, program, or policy will reach its intended outcomes. A theory of change is a key underpinning of any impact evaluation, given the cause-and-effect focus of the research. As one of the first steps in the evaluation design, a theory of change can help specify the research questions. Theories of change depict a sequence of events leading to outcomes. They explore the conditions and assumptions needed for the change to take place, make explicit the causal logic behind the program, and map the program interventions along logical causal pathways.

Gertler et al. 2011

[4] For the full interview, please see country spotlight M1_Nigeria_Motivation for IE, an interview with Dr. Pate.

## Theory of Change for Results-based Financing in Health

RBF interventions work within the country's health system, and therefore they will vary according to country circumstances. Accordingly, impact evaluations need to be tailored to the particular intervention and start from "inside the RBF black box" of each country's program. Country project design and IE teams should work together to identify the design elements of the RBF intervention(s), the policy questions that can be answered through an impact evaluation, as well as the country's priorities among those questions, and how the IE can contribute to the current international knowledge gap on RBF.

In this toolkit, we outline some possible "theories of change" for RBF. Following Musgrove's glossary, we distinguish between Performance-Based Financing (PBF), Performance-Based Contracting (PBC), Conditional Cash Transfer (CCT), In-kind Transfers, and Vouchers. We will only discuss theories of change for these RBF interventions, rather than Output-based Aid or Cash on Delivery.

The following are key design elements that will determine the theory of change:

- Is the intervention on the supply or demand side?
- Are payments to providers or to households?
- Are payments made to individuals or to groups of individuals?
- Is performance measured at the group level or at the individual level?
- Are payments to providers linked to the quantity of services provided?
- Are payments to households linked to utilization of health services?
- Are payments to providers linked to the quality of services provided?
- Does the RBF mechanism introduce or strengthen supervision or monitoring of, or feedback to service providers?
- Does the RBF mechanism increase autonomy of decision-making at the level of the provider or at any other level? Does RBF increase the total amount of resources available to service providers (in supply side interventions)? Does it increase the total amount of resources available at other levels?
- How is performance measured and verified?
- Who is the purchaser of services?
- How high is the financing for performance to providers or households?
- How are the beneficiary providers and/or household selected? What are the criteria?
- Are there any parallel interventions being introduced at the same time as performance-based financing, such as training of providers or information to communities?

The exact theory of change will depend on the key design elements of each program. Below we outline various aspects of a theory of change for the Rwanda PBF program, which is a supply-side RBF program that pays health centers bonuses that depend on the quantity and quality of care provided. We first outline a model of how providers may react to a payment formula that contains various quantity indicators and a quality indicator. We then discuss a model of how to measure the efficiency gap between knowledge and practice of care. Finally, we provide a graphical depiction of what a theory of change may look like for linking provider payment to quantity. Similar graphs could be made to outline the theory of change for other key design elements in the program.[5] A simpler way to depict the theory of change would be through the use of a simple results chain that links inputs and activities with outputs, intermediate and final outcomes. [6]

2012/06/13

**Country Spotlight: Theory of Change**
**Rwanda PBF (P4P) Program**

**Adapted from Gertler and Vermeersch (2012)**

***Payment scheme:***

The [Rwanda PBF] scheme pays for 14 maternal and child healthcare services conditioned on an overall facility quality assessment score. The formula used for payment to facility *i* in month *t* is:

$$Payment_{it} = \left( \sum_{j} P_j U_{jit} \right) \times Q_{it} \quad \text{with} \quad 0 \leq Q_{it} \leq 1,$$

where $P_j$ is the payment per service unit *j* (e.g. institutional delivery or child preventive care visit), $U_{ijt}$ is the number of patients using service j in facility *i* in period *t*, and $Q_{it}$ is the overall quality index of facility *i* in period *t*.

[…]

***Behavioral model:***

[…] we use a simple behavioral model to hypothesize how the introduction of P4P would likely affect medical care provider behavior. We have in mind a rural clinic that is staffed with 4 to 6 medical providers with no close substitutes locally. We assume for simplicity that a facility acts as one single decision-maker that we call the provider. Key to this discussion is the provider's objective function. We assume that medical care providers typically value their patients' health as well as the income they earn from the services they provide to treat patients. We take into account this ethical aspect of preferences by assuming that providers treat all patients who show up for care and provide them with at least a minimum level of care as defined by their ethical standards.

We begin by considering the case where the facility is paid a fixed amount for staff costs and has a fixed budget for non-personnel costs, and assume that the non-personnel budget cannot be reallocated for staff costs. In this case, seeing more patients and providing them with better care does not affect the provider's income. Hence, the provider treats all patients who show up and provides them with the minimum level of care.

The P4P scheme introduces a new dimension to the provider's optimization problem by linking part of the facility's income to the provision of certain services and to quality of care. For simplicity, we assume that the provider allocates effort to two types of patient services (e.g. prenatal care and delivery) and quality of care. Taking into account the basic structure of the P4P formula, we can write the new profit function as

$$V = I + \left[ P_1 U_1(\varepsilon_1) + P_2 U_2(\varepsilon_2) \right] Q(\varepsilon_q) - C(\varepsilon) \tag{3}$$

[5] An editable version of the graph is available in tool **1.01 Graph for Theory of Change**

[6] More information can be found in the tool **3.01a RBF Indicators.** A Powerpoint© template for a results chain is provided in tool **1.02 Results Chain Template.**

where *I* is the fixed salary, $P_i$ is the P4P payment for service *i*, $U_i$ is the total quantity of service *i* provided to patients, *Q* is the overall quality of care, and *C*(∗) is the cost of effort. Recall the the $U_i$'s are listed in Table 1 and *Q* is an index constructed based on the items in Table 2.

The provider chooses effort levels $\varepsilon_1$ and $\varepsilon_2$, to increase the quantity of services provided above the minimum levels necessary to treat patients who show up, as well as effort $\varepsilon_q$, to improve the quality of care above the minimum ethical standards.[7] The service production functions $U_i(.)$ and the quality production function *Q(.)* are increasing in effort, but at a decreasing rate. Finally, the effort cost function *C(.)* is a function of total effort (i.e., $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_q$) and is convex.

The provider then chooses effort levels to maximize income subject to effort levels being weakly positive. In the case of an interior solution, effort is allocated in such a way that marginal revenue of effort is equalized across the three types of effort and that it is equal to the marginal cost of effort:

$$P_1U_1'(\varepsilon_1) = P_2U_2'(\varepsilon_2) = \left[P_1U_1(\varepsilon_1) + P_2U_2(\varepsilon_2)\right]Q'(\varepsilon_q) = C'(\varepsilon) \tag{4}$$

Note that the marginal return to effort supplied to each service depends not just on its own price but also on the price of the other service, as does the marginal return to effort supplied to quality depends on both prices. Hence, an increase in any of the two prices always raises the return to effort supplied to quality. Effort supplied to anything raises the marginal cost of effort because the cost of effort is a function of total effort.

The relative amount of effort allocated to the two types of services satisfies the following condition:

$$\frac{U_2'(\varepsilon_2)}{U_1'(\varepsilon_1)} = \frac{P_1}{P_2} \tag{5}$$

i.e. the ratio of the marginal returns to effort in delivering the services should equal the ratio of the payment rates for those services. Hence, more effort will be allocated to the service that has the highest price and the higher marginal productivity of effort.

***Economic Predictions:***

We can discuss the likely effects of introducing P4P in terms of a comparative static of price increases, whereby the original level of *P* and $\varepsilon$ are close to zero. Consider an increase in $P_1$, the payment for service 1. This will raise the marginal revenue from supplying effort to service 1 and to the provision of quality, and therefore is an incentive to supply more effort to that service and quality. Because the increased effort raises the marginal cost of total effort, the provider will reduce effort to service 2. As a result, the increase in effort for service 1 and for quality comes at the cost of both reduced effort for the other service and reduced leisure. Hence, while the total amount of effort increases, the relative allocation of effort increases to service 1 and quality and falls to service 2. If the price increase is large enough, the optimal effort allocated to service 2 will fall below the minimum ethical constraint and, as a result, the constraint will bind.

However, the comparative static analysis of a single price change is not exactly applicable to the introduction of a P4P scheme as the P4P scheme changes all prices simultaneously. Before the price increase, all effort levels are at the minimum ethical constraint. Increases in the prices of the services will increase the allocation of effort to quality because increases in any and all prices raise the marginal return to supplying effort to quality. The largest allocations of effort to a service will be to those services for which the relative price increases are the largest and the marginal productivity of effort is the highest. Analogously, the smallest allocations of effort will be to those services that get the smallest relative price increase and have the lowest marginal return to effort. In fact, if for a particular service the relative price increase is small enough and the marginal productivity of effort low enough, the provider will not supply any more effort to that service despite the absolute increase in price. In this case, the supply of effort will remain at the minimum ethical bound.

Hence, the effect of the introduction of the P4P payments depends not only on the relative payment rates, but also on how hard it is to increase the levels of services. In general, we argue that it takes more work to increase services that depend on patient choices than services that are completely in the provider's control. For example, it takes more work

[7] In this way, we effectively normalize the minimum effort levels to zero.

to convince a pregnant woman to come to the clinic for prenatal care than give the women a tetanus shot once she is there. Hence, even if payments were equal for an additional patient visit as for a tetanus shot, one would expect to see larger increases in the number of tetanus shots (which is under the control of the provider) than in the number of visits to the facility (which is largely under the control of the patients). Moreover, we argue that initiation of care takes more effort than its continuation. For example, it will take a provider substantial amounts of effort to go out to the community to find pregnant women, especially in the first trimester of pregnancy, and bring them in for prenatal care. By contrast, it is a relatively easier task to use an existing prenatal care visit to lobby women already in prenatal care to deliver in the facility.

The previous discussion assumes that the prices of the services enter in the profit function in a simple linear fashion as presented in equation 2. In reality, the payment scheme is more complicated and the services listed in Table 1 are made up of both primary reasons to visit a clinic as well as services provided conditional on such a visit. While they are all $U_i$'s, the services provided during the visits also enter the quality index $Q$. Moreover, the payment $P$ for seeing a patient depends on the services provided during that visit. Consider the payment for prenatal care. Providers receive \$0.18 for every pregnant women who starts prenatal care, an additional \$0.37 if the women completes at least 4 visits, an additional \$0.92 if they give the patient a tetanus shot and malaria prophylaxis during a prenatal care visit, and an additional \$1.83 if they assess the delivery to likely be risky and refer the mother to deliver at the district hospital. Hence, payments for prenatal care depends not only on the number of pregnant women coming for care and the number of times they visit, but also on the content of care provider during those visits.

In fact, payment rates for visits are much higher if the provider supplies better content of care. As we discussed, a provider will receive \$0.55 for four prenatal care visits of low quality versus \$1.47 for providing high quality. If the provider detects a high-risk pregnancy and refers the woman to the hospital for delivery, payments for this high-quality care even increase to \$3.30. In the case of growth monitoring, the payment to the provider is \$0.18 per visit plus an additional \$1.83 if the child is malnourished and she refers her to the hospital for treatment. Since 45 percent of Rwandan children under age five have moderate chronic malnutrition, and 19 percent have severe chronic malnutrition,[8] (Institut National de la Statistique du Rwanda and ORC Macro 2006), the expected payment for a high quality growth-monitoring visit is quite high. Overall, the incentive structure focuses not just on treating more patients, but on providing more patients with higher quality of care; this happens through both the multiplicative scaling factor $Q$ and by direct payment for content of care services in the $U_i$'s.

***Empirical predictions:***

This discussion provides us with a number of empirical predictions. First, increases in payments will be more effective for services for which the relative price increase is highest and for those that have the highest relative marginal return to effort. Second, increases in payments will not necessary increase all services. There may be no effect on services for which payment rates and the marginal return to effort is low. Third, payment rate for a service depends not only on the number of patients treated, but also the content of care provided during a visit and it is this payment rate that matters for the allocation of effort. Finally, we expect the introduction of P4P to increase quality $Q$, the multiplicative factor in the payment formula.

[8] Moderate (severe) chronic malnutrition corresponds to height-for-age below -2 (resp. -3) standard deviations from the median of the reference population. (Institut National de la Statistique du Rwanda and ORC Macro 2006)

## Country Spotlight: Efficiency Gaps
## Rwanda PBF (P4P) Program

**Adapted from Gertler and Vermeersch (2012)**

Another interpretation of how P4P works is based on the idea that providers are not delivering services up their full ability (knowledge). There is indeed evidence of this efficiency argument as provider deliver of clinical services during prenatal care is substantially lower than their knowledge of appropriate clinical procedures. Recall that providers on average know 63 percent of appropriate procedures, but deliver only 45 percent. This leaves an 18 percentage point difference between knowledge and practice. If we consider a provider's knowledge as their production possibilities frontier, then one can interpret the gap between knowledge and practice as a measure of technical inefficiency. The P4P incentives are intended to reduce technical inefficiency.

We present the efficiency gap in figure where skill is represented on the horizontal access as the share of prenatal CPG recommended clinical services that the provider knows and the vertical access represents quality delivered as the share of prenatal CPG recommended clinical services actually provided. The 45° line is the production possibility frontier (PPF) where providers deliver clinical quality care to the best of their knowledge. If providers deliver a quality of care below their level of knowledge, then they would be performing inside the PPF. The vertical distance between the frontier and the performance point is a measure of technical inefficiency.

We also included in figure A the actual performance curves of the providers in our data set. The curves are bivariate nonparametric regressions of quality against knowledge separately for treatment and comparison groups at endline. Notice that both lines are well inside the PPF implying substantial levels of technical inefficiency at all skill levels. In addition, while the performance curves are upwards sloping, they are flatter than the PPF. This implies that while knowledge improves performance, the efficiency gap increases with knowledge. Finally, the performance curve for the treatment group is above and steeper sloped than the curve for the comparison group. This implies that P4P reduced the efficiency gap and reduces it more for more skilled providers.

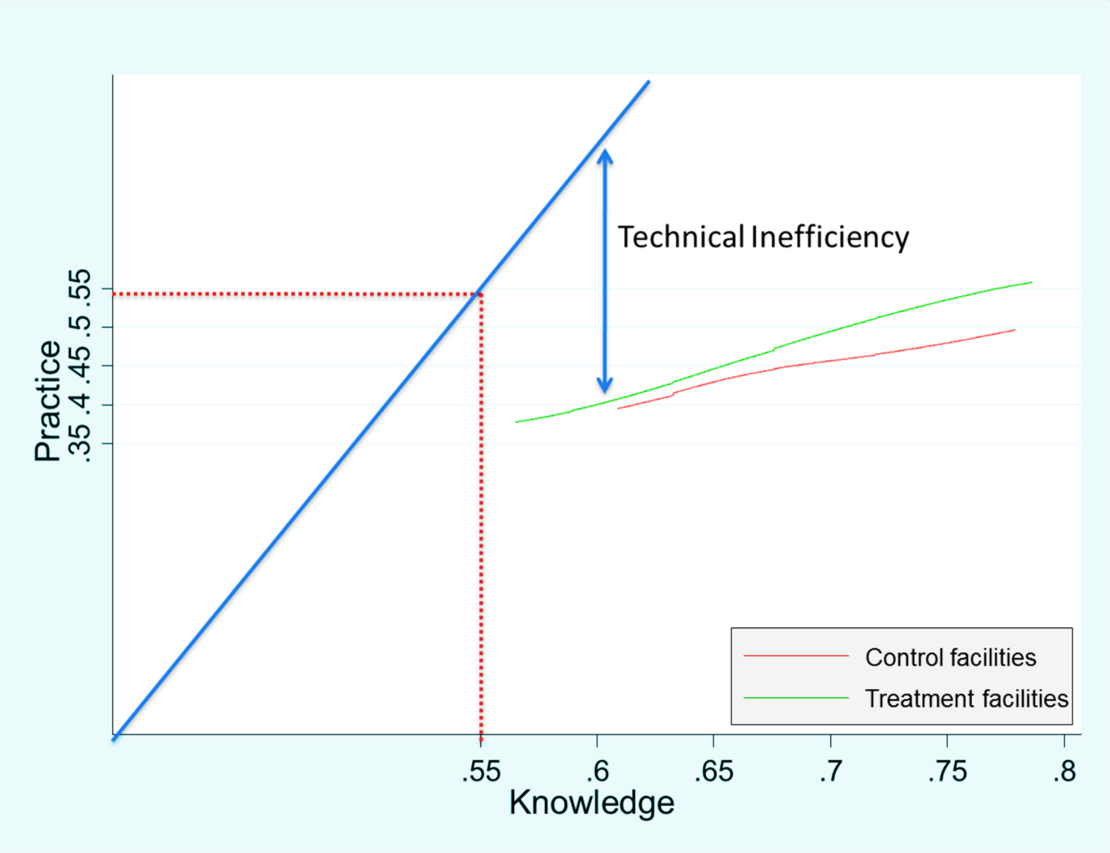
We now estimate the order of magnitude of the impact of P4P on the efficiency gap. We measure the efficiency gap as the share of CPG clinical services the provider knows minus the share of CPG clinical services delivered. We find that P4P reduces the efficiency gap by 3.5 percentage points or about 20 percent of the gap on average (Table 9 Model 1). When we control for provider knowledge, the effect of P4P on efficiency increases slightly to 4 percentage points (Table 9 Model 2). In this model higher knowledge is actually associated with a larger efficiency gap. In other words, while increases in provider knowledge improve the quality of care, the improvement in quality is less than the improvement in knowledge. Finally, we estimate that P4P has a much larger effect on efficiency for more knowledgeable providers. We find no increase in efficiency for providers below the knowledge median, but we find a 6 percentage point improvement among providers above the knowledge median (Table 9 Model 3).

Table 9: Impact of P4P on Efficiency Gap (Knowledge – Quality)

| | β | P-Value | β | P-Value | β | P-Value |
|---|---|---|---|---|---|---|
| P4P (=1) | -0.035 | 0.00 | -0.04 | 0.03 | -0.02 | 0.24 |
| Knowledge Z-Score | | | 0.16 | 0.00 | 0.21 | 0.00 |
| P4P * Knowledge in Top 50% | | | | | -0.06 | 0.01 |
| N Observations | 3709 | | 3709 | | 3709 | |

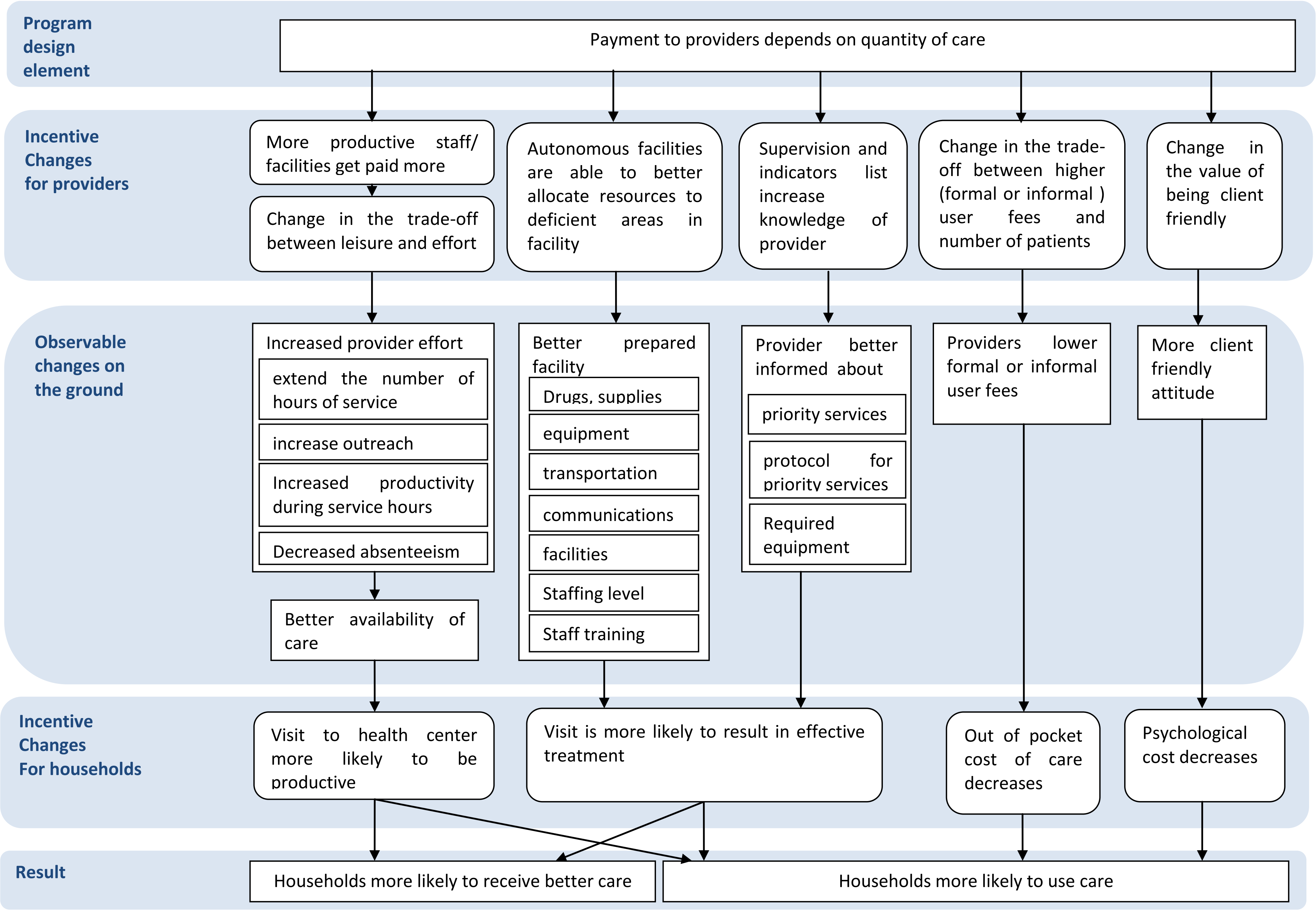
Notes: P-Values are for one-sided tests of the null hypothesis that β = 0 and are calculated based on a WILD bootstrap with 999 draws.

Figure A: The Knowledge-Practice Efficiency Gap for Prenatal Care in 2008 Follow-up



Notes: The horizontal axis is Knowledge expressed as the percentage of protocol items correctly identified by the provider during the administration of the vignette. The vertical axis is the percentage of protocol items that were delivered during prenatal care, as reported in patient exit interviews and in household surveys.

Figure 4: An Example of a Theory of Change for Payment to Providers


Program design element
Payment to providers depends on quantity of care
Incentive Changes for providers
More productive staff/ facilities get paid more
Change in the trade-off between leisure and effort
Autonomous facilities are able to better allocate resources to deficient areas in facility
Supervision and indicators list increase knowledge of provider
Change in the trade-off between higher (formal or informal ) user fees and number of patients
Change in the value of being client friendly
Observable changes on the ground
Increased provider effort
extend the number of hours of service
increase outreach
Increased productivity during service hours
Decreased absenteeism
Better availability of care
Better prepared facility
Drugs, supplies
equipment
transportation
communications
facilities
Staffing level
Staff training
Provider better informed about
priority services
protocol for priority services
Required equipment
Providers lower formal or informal user fees
More client friendly attitude
Incentive Changes For households
Visit to health center more likely to be productive
Visit is more likely to result in effective treatment
Out of pocket cost of care decreases
Psychological cost decreases
Result
Households more likely to receive better care
Households more likely to use care

### Country Spotlight: A behavioral model for the provision of targeted versus non-targeted outcomes

### Cambodia Contracting Approach

The Cambodia model of contracting out health services linked incentives to 8 targeted outcomes. Bloom et al. (2006) outline a model of how such a contract will affect provision of the targeted services, and the provision of the non-targeted services.

*"A Holmstrom-Milgrom (1991) framework suggests that contracts linking incentives to the 8 targeted outcomes will lead to better performance on those measures, but how it affects other outcomes depends on whether effort directed at those non-targeted outcomes is a complement or substitute with the targeted outcomes. Either scenario is plausible. For example, it could be that the incentives provided to the contractor cause contractors to create incentives for health workers to reduce absence from the facilities, and that this is complementary with providing other types of care. On the other hand, facilities might shift resources away from unmeasured care to targeted outcomes.*

*We will formalize this idea in a simplified Holmstrom-Milgrom (1991) framework. Suppose there are two health outcomes. The agent has control over two kinds of effort that are costly to exert. Suppose only one of the outcomes is contractible. Denote the outcomes C and NC and the effort types $e_1$ and $e_2$ and let them be produced as follows*

$C = f(e_1,e_2) + \varepsilon$

$NC = g(e_1,e_2) + \eta$

*The agent cares about compensation w as well as the cost of exerting effort,*

$u(w,e_1,e_2) = w - c(e_1,e_2)$

*Agents are paid a linear wage in the amount of the contracted outcome produced*

$w = \alpha + B.C$

*The agent's first order conditions are*

$$\frac{dc}{de_1} = B\frac{df}{de_1};\frac{dc}{de_2} = B\frac{df}{de_2}$$

*Note that the function $g(e_1,e_2)$ does not appear in the first order conditions. The agent chooses effort only according to the tradeoff between the cost of effort and the marginal increase in C output that results from effort. Increasing B will typically increase C, but may increase or decrease NC."*

Bloom, E. et al (2006)

## Impact Evaluation Questions

Impact evaluation questions follow directly from the theory of change that is associated with a particular intervention. While each RBF intervention is somewhat different, there are nonetheless a number of evaluation questions that are being addressed in a number of different evaluations in the HRITF-financed impact evaluation program. In the following sections, classify those questions as "first" and "second" generation, though these are not necessarily sequential. When choosing impact evaluation questions, it's important to keep in mind the following: not every IE needs to address every evaluation question – what is more important is whether the chosen questions are relevant, both locally and globally.

## First Generation Questions: Does RBF work?

As mentioned below, impact evaluation purports to answer the question: "What is the impact (or causal effect) of the intervention on outcomes of interest?" The first generation of policy questions to be addressed by IEs relate to determining *whether or not health-related RBF works, to what degree and in what contexts*, or in short "Does RBF work?" In this case, there are several sets of **outcomes of interest**:

**Quantity of health services delivered:** Most existing RBF interventions are designed to increase utilization of key health services for maternal and child services by providing additional bonus payments to providers and/or users. These services typically include preventive health care, such as immunizations, pre-natal care, institutional delivery, or bed-net distribution. We can measure service delivery at the provider-level, as well as at the population-level, independently of whether we are evaluating supply-side or demand-side interventions. For example, we can measure indicators such as the number of prenatal care visits, institutional deliveries, and growth monitoring visits using health facility survey data. On the side of the user, we can compute the probability that a woman will have 4 prenatal care visits, the probability that she will deliver in a health facility, or the probability that a child's growth has been monitored in the last 6 months through household survey data.

**Quality of the services provided:** There is a concern that bonus payments to providers to increase quantity of services provided will lead to a decrease in the quality of services provided, particularly in rural areas with limited human and capital resources. For this reason, bonus payments are typically tied not only to the quantity of services, but the quality of services as well. Whether or not payments are tied to quality, it is crucial for the IE to measure whether the RBF mechanism affects quality, either positively or negatively. Globally, this type of evidence will help us understand how to increase both the utilization and the quality of key services.

**Health status of the population:** The final objective of any RBF mechanism is not only to increase quantity and/or quality of services, but more importantly to improve the health status of the population. Most existing RBF interventions are intended to have a direct impact on the child and maternal health status of populations. As highlighted above, while it may not be possible to measure MMR or IMR, outcome indicators such as nutritional status are observable through anthropometric measurements and/or anemia testing.

**Resource management at the health center:** RBF is typically featured as a measure for health systems strengthening. For this reason, one key policy question is how the RBF intervention(s) impact financial, human resources, equipment and drug supply management at various levels in the health system.

**Non-RBF services delivered:** It is also important to measure any externalities, positive or negative, associated with the RBF intervention(s). There is concern that providers will shift their provision of care to RBF services in order to increase the RBF payment, at the expense of non-RBF services. For this reason, the IEs should capture information on non-RBF services to identify if there is any shift in quality and quantity of non-RBF services as a result of the RBF intervention(s).

**Equity of service delivery and utilization:** There are several potential ways through which RBF may affect the equity of service delivery. For example, RBF in rural or remote areas may target disadvantaged populations and may or may not be able to increase the accessibility and affordability of care for those populations. RBF may have downstream effects on out-of-pocket payments (formal or informal), which could affect the type of population using services. A number of RBF programs include differential higher payments for services provided to the poor and/or remote populations, and it would be important to know whether such payments are successful in overcompensating providers for the costs of reaching them. Wherever possible, evaluations should check whether the RBF program disproportionately benefits the poor.

**"Market" effects:** In many instances, households have a choice as to which provider to use for care. This is often the case in urban areas, where households have a choice between public and private providers. But even in rural areas, in many countries households can choose which public provider to attend, or they may be able to use a private provider. In addition to the public-private dimension, there may be different types of providers, such as doctors, pharmacies, drug stalls, traditional healers, community health workers, etc. This is important in the analysis of the impact of RBF, for several reasons:

- Take for example the case of a supply-side RBF program that explicitly rewards public providers for the services they provide, or the case of a demand-side program that gives women vouchers to attend public providers. In both cases, an increase in the quantity of services provided by those public providers is not sufficient to prove that service levels overall have increased. Patients may have switched from private providers to public providers. One could even imagine situations where overall service provision goes down – say for example, if private providers go out of business and the public providers do not fully take over the patient loads from the private providers. Household surveys allow us to measure whether service provision overall went up if they ask for utilization of services from all types of providers.
- Market effects may also have an impact on the overall quality of care that is provided. Imagine a situation where there are public and private providers, and where private providers provide better quality care than public providers. If (demand or supply-side) RBF makes patients switch from private providers to public providers, and the quality of care does not change, then on average patients will receive worse quality of care with RBF than without RBF. To measure whether this happens, one would need some measure of quality of care at the population level. This could be done through household surveys, though measures of quality of care from household surveys may suffer from recall bias. Alternatively, one could field a survey to measure quality in a representative sample of public and private facilities; however, one would also need to have a measure of patient loads in both types of facilities in order to estimate the average level of quality for the population. Measurement can get very complicated if there are many different types of providers.

**Income effect or incentive effect?** Paying performance-based payments to health care provider or behavior-dependent transfers to households can potentially have two effects. The first effect is the so-called *resource* effect[9] that comes from the fact that the payments increase the provider's or the household's resources. The second effect is the *incentive effect* which stems from linking the payments to behavior or performance, as opposed to lump sum or unconditional payments. The relative size of the resource and incentive effects is important for policy making: if the resource effect is very large compared to the incentive effect, then it would probably be cheaper to increase the amount of resources without linking them to performance or behaviors. This way, one could avoid the often expensive verification activities that are necessary when making payments based on performance or behaviors.

The existence of the resource and incentive effects has implications for impact evaluation. Say for example that the treatment group receives performance-based payments while the control group receives nothing. By comparing those two groups, the impact evaluation estimates the resource effect and the incentive effect *together*. It is not possible to know what the incentive effect alone amounts to. Therefore it's difficult to know whether the cost of verification was a worthwhile expense, or whether it would have been better to distribute the resources without putting performance conditions. By contrast, take an evaluation where the treatment group receives performance-based payments while the control group receives the same amount of money (on average), but not linked to performance. In other words, on average the control group gets "compensated" with the same amount of money as the treatment group but the amount does not depend on the performance of the control group. In that case, the treatment group and the comparison group have the same amount of resources on average. Therefore differences between the two groups in terms of outputs or outcomes cannot be due to the resource effect, but rather they must be due to the incentive effect.

2012/06/13

**Country Spotlight: Resource versus incentive effect**
**Rwanda PBF (P4P) Program**

*"Because our aim was to assess the effect of the incentive-based bonus (P4P) scheme separately from the effect of an increase in financial resources, the amount of resources for the intervention and comparison facilities had to be held constant. Traditional input-based budgets allocated to the facilities in the control group were increased by the average amount of P4P payments that facilities in the intervention group received every 3 months during the 23-month assessment window."*

Basinga et al. (2011)

In addition to answering the above listed questions, a crucial element of determining the effects of RBF is to disaggregate impacts by the characteristics of providers and beneficiaries. These include:

[9] The resource effect is sometimes also called the income effect.

**Provider characteristics**: We would also like to disaggregate the impacts of the RBF intervention(s) based on the provider's training and knowledge levels, autonomy, type of ownership (public/private), etc. This allows us to determine if the impacts of the RBF intervention(s) are greater for:

- Highly skilled staff versus lesser skilled staff
- Providers with more autonomy versus less autonomy
- Public facilities versus private facilities

There are various reasons why RBF programs impacts may be different for different types of providers. For example, it may be that older providers have lower (or higher) baseline measures of knowledge, and that providers with less knowledge do not respond as much to the incentive. In this case, one may want to complement the RBF intervention with some kind of continuing medical education for the low skill providers.

**Population characteristics**: We would like to disaggregate the impacts of RBF interventions on the population by age, gender, poverty level, rural/urban. Through this, we can determine if the impacts of the RBF intervention(s) are greater for:

- Younger women versus older women
- Younger children versus older children
- Wealthier households versus poorer households
- Rural households versus urban households

## Second Generation Questions: How can RBF work better?

As RBF is introduced in more settings, a number of common design and implementation challenges confront Governments, international agencies and implementing partners. Stakeholders are finding it is not enough to know whether or not RBF works, but also how to maximize the impacts of RBF. Impact evaluations can be designed to address some of these core questions related to RBF design, including:

**What are the right levels of rewards?** What type of reward should be introduced (cash vs. in-kind)? What amount of reward is most cost-effective at improving outcomes? What are the right reward levels for each indicator selected? What are the right indicators to trigger rewards? Can we come up with a formula to determine the level of payments per service?

**Who should be incentivized in supply-side interventions?** Should the payments be introduced at the national level or the sub-national level? Should payments be made at the facility or provider level? Should payments be introduced at the hospital level or at the primary health care level?

**Who should be incentivized in demand-side interventions?** Should the payments/rewards be targeted according to socio-economic criteria? Are payments best made to household heads, women, men, children? Should monetary rewards be distributed in cash or through bank accounts?

**How do we reduce reporting errors and corruption?** What is the optimal intensity and frequency of data verification and data counter-verification? What are the most effective sanctions against incorrect reporting or corruption?

**How does provider knowledge affect their reaction to performance-based rewards?** Do higher skilled providers respond better than lower skilled providers? Will capacity building (such as training activities) improve provider response? How much capacity building or re-training is optimal?

**What are the key organizational building blocks to make RBF work?** What is the right level of autonomy over use of funds, hiring, procurement, etc.? What is the most effective ownership structure (public vs. private vs. NGO)?

## The HRITF-funded Evaluation Portfolio

**Country research questions on RBF contribute to global knowledge.** The HRITF finances impact evaluations are designed around a common research agenda and methodology, while still being tailored to each country's specificities, operational objectives and policy interests. The combination of results from various countries and RBF approaches will create a unique, comprehensive assessment of RBF that explores multiple dimensions regarding what RBF is, how it is implemented, and what behavior and outcomes it triggers.

**Impact evaluation questions.** While all impact evaluations aim to identify the basic question of what is the impact of RBF on common service and health outcome indicators, each evaluation also provides additional insight into a specific dimension of RBF or into a specific type of RBF intervention. Some countries evaluate the impact of supply versus demand-side payments; the impact of differential incentive levels; the equity aspects of RBF; etc. This will contribute to bridging the global knowledge gap, not only on whether RBF works, but also on why RBF works or does not, and what the drivers of RBF success (or failure) are. Table 2 indicates the focus of each country's evaluation.

**Outcomes of interest.** Most impact evaluations financed by HRITF have a common focus on maternal and child health. Within this umbrella, countries focus on specific aspects, such as family planning, Prevention-of-Mother-To-Child-Transmission (PMTCT) of HIV/AIDS, or cross-sectional issues such as out-of-pocket payments for healthcare or staff motivation. A few of the operations being supported look at RBF and NCDs in an effort to find lessons that are applicable to MCH and nutrition. For example, the Karnataka evaluation focuses on RBF payments for the treatment of cardiovascular and cancer conditions at the tertiary hospital level. In addition, the Turkey program assessment focuses on detection and control of diabetes and hypertension at the family practice level.

Table 3 below presents the variety of outcomes of interest by country.

### Table 2: Interventions Evaluated, by Country

| Evaluate the impact of… | Countries |
|---|---|
| **Supply-side RBF payments** | Afghanistan, Argentina, Benin, Burkina Faso, Cameroon, CAR, DRC, India, Kyrgyz, Lesotho, Nigeria, Rwanda, Turkey, Zambia, Zimbabwe |
| **RBF and training of providers** | Zimbabwe |
| **Additional financing** | Zambia, Zimbabwe |
| **RBF for quality of care** | Afghanistan, Benin, Cameroon, Kyrgyz, Nigeria, Zambia |
| **Differential incentive levels** | Argentina |
| **Enhanced monitoring and supervision** | Argentina, Kyrgyz Republic, Cameroon, CAR |
| **RBF for hospitals** | Kyrgyz, Argentina, India |
| **Demand-side RBF payments** | Rwanda |
| **Community-Based RBF** | India, Rwanda |
| **Other or TBD** | Burkina Faso, Lao, Liberia, Sri Lanka, Tajikistan |

### Table 3: Outcomes of Interest, by Country

| Outcomes of interest | Countries |
|---|---|
| **Maternal Care (Quality/Utilization)** | Afghanistan, Argentina, Benin, Burundi, Cameroon, CAR, DRC, Kyrgyz Republic, Lesotho, Liberia, Nigeria, Rwanda, Tajikistan, Zambia, Zimbabwe |
| **Family Planning** | Afghanistan, Cameroon, CAR, Ethiopia, Lesotho, Rwanda, Zambia, Zimbabwe |
| **Child Health Care (Quality/Utilization)** | Afghanistan, Argentina, Benin, Burundi, Cameroon, CAR, DRC, Kyrgyz Republic, Lesotho, Liberia, Nigeria, Rwanda, Tajikistan, Zambia, Zimbabwe |
| **Quality of Care** | Afghanistan, Benin, Burundi, Kyrgyz Republic, Lesotho, Liberia, Zimbabwe |
| **Out-of-pocket Payments** | Afghanistan, Benin, DRC, India, Tajikistan, Zimbabwe |
| **Tuberculosis, Malaria, HIV/AIDS** | Afghanistan, Benin, Liberia, Nigeria, Zambia, Zimbabwe |
| **Staff Motivation** | Benin, DRC |
| **Non-communicable diseases** | India (tertiary care), Turkey (prevention) |

# Module 2

# Building the Impact Evaluation Team

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

## Module 2. Building the Impact Evaluation team

| Main Recommendations and Available Tools for this Module | | | |
|---|---|---|---|
| **Recommendations** | **Critical** | **Important** | **Nice to have** |
| • Team member(s) primarily responsible for project design and implementation (e.g. the TTL) should not serve as the principal investigator. | ✓ | | |
| • The principal investigator and evaluation coordinator play a crucial role in supervising the survey firm(s). | ✓ | | |
| • Local research counterparts can greatly contribute to the success of the impact evaluation, because they can bring local knowledge and foster country ownership of the program. | | ✓ | |
| • Teams should assess local capacity to conduct surveys and identify whether any technical support will be needed to ensure the quality of survey data. | | ✓ | |
| • A data quality expert can help set up the right initial conditions for ensuring the quality of survey data before the survey firm goes into the field. A local supervisor can verify the data quality assurance processes during the implementation of the surveys. | | ✓ | |
| • Qualitative and cost effectiveness analysis can add great richness and granularity to the questions that the impact evaluation will answer. | | ✓ | |
| • Impact evaluations involve several rounds of sophisticated data – a good data analyst will help the team manage and analyze the data quickly and reliably. | | ✓ | |
| • While power calculations can be the responsibility of the principal investigator, a power calculation expert may have more time and expertise to dedicate to this task. | | | ✓ |

**Tools**

- 2.01 Principal Investigator TOR
- 2.02 Evaluation Coordinator TOR
- 2.03 Data Analyst TOR
- 2.04 Local Researcher TOR
- 2.05 Power Calculation Expert TOR
- 2.06 Data Quality Expert TOR
- 2.07 Qualitative Principal Investigator TOR
- 2.08 Qualitative Field Worker TOR
- 2.09 Cost-analysis Expert TOR

## Module Contents

An important attribute for a credible evaluation is that there are no conflicts of interest for the evaluators. In other words, the evaluators must be sufficiently separated from the program implementers. However, it is often difficult for an impact evaluation to be completely divorced from the operational rules of the program, because it is the rules of the program that determine, among other things, where the comparison group is going to come from.

In light of this difficulty, we recommend that the design and implementation of the *impact evaluation* and the analysis of data should be conducted by a team that is sufficiently separate from the team that is responsible for the design and implementation of the *project*. However, these teams will still need to work together in order to ensure that:

- The priority policy questions for the respective country are integrated into the impact evaluation
- The IE and project activities are properly timed (e.g. the baseline should be completed before the intervention starts)
- The implementation of the intervention concurs with the selection of treatment and comparison groups for the impact evaluation strategy.

An IE team typically consists of a combination of full-time and part-time staff based both locally and internationally. The IE team usually consists of the following members:

- Principal investigator (PI) and (if relevant) Co-Principal Investigator (Co-PI)
- Evaluation Coordinator (EC)
- Data analyst
- Data Quality Expert(s) and potentially External Supervisor
- Power Calculations Expert

The IE Team for the project may also include:

- Qualitative Research Expert
- Cost Analysis Expert

**Terms of Reference** for all these team members are provided in the Toolkit.

It is good practice to include local collaborators in the IE team where capacity exists, or where it can be build. For example, local academics may be interested in participating as co-Principal Investigators in the evaluation, and gradually increasing their skills. Local participation can increase country buy-in, local knowledge and ownership of the program, and result in a win-win situation.

It is important to emphasize that communication among team members and the coordination between project and impact evaluation teams via the TTL is crucial for the success of the IE.

2012/03/08

**Country Spotlight: Implications of team building and reporting on the implementation of impact evaluation**
**Democratic Republic of Congo Health Sector Rehabilitation and Support Project (HSRSP)**

At the early stages of the impact evaluation and for a significant portion of baseline data collection, the Principal Investigator (PI) of the IE team and the project Task Team Leader (TTL) were not based in DRC, (…) international consultants were in charge of leading the sampling and randomization on the one hand, and the preparation and implementation of data collection on the other hand, with only a few missions on the ground and no contacts with the PIU and the World Bank Team in DRC. Outside from the team, the communication between provincial authorities and the project and IE teams was also lacking. Because of the team turnover and lack of presence on the ground, the survey firm lacked training, supervision, verification and quality control during fieldwork.

[To respond to these issues], after the completion of baseline data collection, (…) the new TTL of the project appointed a new Principal Investigator and co-Principal Investigator in order to analyze the baseline data and prepare for the follow-up survey. The Principal Investigator and co-Principal Investigator were not based in DRC, but they appointed a research assistant (RA), based in Lubumbashi full time, to handle day to day activities of the IE and understand practical challenges to address in a post-conflict setting. The RA was the focal point of activities on the ground and gave regular feedback to IE team members off the ground. The RA was also able to understand several challenges linked to program implementation. The whole team emphasized communication between team members.

***Local continued presence of IE team members is key*** *to the success of the IE. Having at least one co-Principal Investigator, evaluation coordinator or research assistant on the ground, especially during data collection activities, is an extremely valuable strategy.*
*A common issue encountered during impact evaluations is the lack of communication between IE and project teams. This can clearly jeopardize the validity of the IE.* ***World Bank Task Team Leaders must*** *make a point in bridging the information gap between both teams, and* ***facilitate collaboration between operational and IE teams****.*
*It is important to clearly define the role of each team member.* ***Terms of reference should include all activities a team member is expected to endorse, and reporting modalities*** *to the TTL and/or the team. (…) TTLs are in charge of ensuring each team member has the capacities to fulfill those terms of reference, and training them if not.*

Full story available: see Country Spotlights section of the Toolkit.

In the following section, we detail the roles and responsibilities that we believe the different possible members of the IE team should have:

## Principal Investigator

The role of the Principal Investigator is to provide technical leadership on the IE design, methodology and analysis, as well as overall management of the study. The Principal Investigator tailors the evaluation to country-specific conditions, while keeping in mind the objectives of the global RBF IE program.

**The Principal Investigator works with the project TTL and Government counterparts in order to ensure that the IE design and implementation are integrated with the roll-out of the RBF intervention. As discussed above, for an evaluation to be credible, the evaluators must be sufficiently separated from the program implementers. Therefore, and barring truly exceptional circumstances, we highly recommend that the TTL of the project or other team member(s) primarily responsible for project design and supervision should not serve as the principal investigator for the impact evaluation.**

Table 4 outlines the estimated time commitment for a principal investigator.

**Table 4: Estimated Time Commitment for a Principal Investigator**

| Activities | Working Days |
|---|---|
| Impact Evaluation Design and Discussions with Key Counterparts (includes at least one mission in country) | 30 |
| Baseline Data Collection Supervision and Management | 15 |
| Baseline Data Analysis and Dissemination | 20 |
| Monitoring and Follow Up | 15 |
| Endline Data Collection Supervisions and Management (includes at least one mission in country) | 15 |
| Impact Analysis and Dissemination | 30 |
| TOTAL | **125 days** |

This time commitment estimate is based on the assumption the Principal Investigator will collaborate with an Evaluation Coordinator (see below) to provide day-to-day assistance to the survey firm during data collection preparation, field work, entry and analysis. In some cases, Principal Investigators may not be able to commit sufficient time and attention to an evaluation on their own, and will need to partner with another investigator as a co-Principal Investigator (co-PI).

We recommend that the Principal Investigator should have at least the following qualifications:

- PhD in relevant field, preferably economics or health policy.
- Minimum 5 years of project impact evaluation experience
- Minimum 5 years experience in designing and implementing quantitative impact evaluations using randomized or otherwise controlled designs
- Relevant experience in measurement of health outcomes through household surveys
- Relevant experience designing and coordinating field work for large household surveys and health facility surveys
- Relevant experience analyzing quantitative data (household and facilities) using statistical analysis software (preferably STATA)
- Relevant experience in coordinating implementation of impact evaluation field work
- Excellent written English communication skills, with focus on research protocols, research papers and descriptive reports for diverse audience
- Ability to facilitate communication between various levels of management and work independently in order to meet deadlines
- Ideally, the Principal Investigator should have published evaluations in peer reviewed journals.

## Evaluation Coordinator

The Evaluation Coordinator manages the day-to-day activities related to the design of the impact evaluation, data collection and analysis. This typically requires a substantial time commitment, as estimated in Table 5.

### Table 5: Estimated Time Commitment for an Evaluation Coordinator

| Activities | Working Days |
|---|---|
| Impact Evaluation Design and Discussions with Key Counterparts (includes at least one mission in country) | 25 |
| Baseline Data Collection Preparation Supervision and Management (in country for the preparation and the full duration of the survey) | 125 |
| Baseline Data Analysis and Dissemination | 50 |
| Monitoring and Follow Up between the baseline and endline surveys | 60 |
| Endline Data Collection Supervisions and Management (in country for the preparation and the full duration of the survey) | 125 |
| Impact Analysis and Dissemination | 75 |
| TOTAL | **460 days** |

We further recommend that the Evaluation Coordinator should have at least the following qualifications:

- Master's level degree or equivalent in relevant field, such as health, public health, or economics
- Experience with statistical analysis software (STATA)
- Relevant experience conducting, managing and designing field work and data collection for empirical research
- Excellent written English communication skills, with focus on research protocols, research papers and descriptive reports for diverse audience
- Fluency in local language preferable
- Exceptional organizational skills, ability to facilitate communication between various levels of management and work independently in order to meet deadlines
- Previous experience with project impact evaluation in developing countries is highly desirable

## Data Analyst

The data analyst is responsible for helping the Principal Investigator to completing the analysis of the baseline and endline datasets in a timely manner. Analysis of the baseline data is required to validate the evaluation design, provide the project team and partners with a descriptive report of the data and recommendations for midline (if applicable) and endline rounds. For HRITF-funded evaluations, initial analysis of the baseline data to validate evaluation design is a milestone to release the second tranche of funding. In addition, there will be considerable pressure to produce the impact analysis after endline data is collected. While the principal investigator and evaluation coordinator may have the skills to conduct this analysis, they may not have sufficient time available to clean and document the data, run the analyses and write-up the results. Table 6 outlines the estimated time commitment for a data analyst.

Table 6: Estimated Time Commitment for a Data Analyst

| Activities | Working Days |
|---|---|
| Baseline Data Analysis and Dissemination | 50 |
| Impact Analysis and Dissemination | 75 |
| TOTAL | **125 days** |

## Power Calculation Expert

A power calculations expert determines the sample size required or the minimum detectable treatment effect to answer the proposed evaluation questions. Put simply, this expert will estimate the minimum sample size needed to detect a meaningful difference in results between the treatment and comparison groups. For studies with a given sample size this expert will estimate the smallest treatment effect that can be statistically detected. The time commitment of a power calculation-sampling expert is minimal compared to other roles: if the required data are already available, 3-5 days should be sufficient.

The role of the power calculation expert can be assumed by the Principal Investigator. However given the high technicality of this task and the limited number of work days required, we recommend to hire a specialist.

## Data Quality Expert

A major challenge when implementing surveys is to ensure sufficient quality of the data. In some countries, there is strong local capacity to plan and implement good quality data collection, while in others there is limited capacity. Experience has shown that, in countries where there is limited local capacity, a competitive bid is typically awarded to an international firm to collect data. In such cases, the following issues have arisen: (i) international firms hire local subcontractors and/or employees to perform the data collection; therefore, the quality of the data depends on the quality of training and supervision provided by the international firm; (ii) international firms do not necessarily have experience in the particular country where the survey is taking place, which can lead to sub-standard results; and (iii) international firms have an incentive to limit the time of international staff to be in country in order to reduce costs, resulting in reduced supervision and technical support.

For the above reasons, we recommend that the IE team include a consultant (either an individual or a firm hired specifically for this purpose) to provide technical assistance on in-country data quality assurance for maximizing data quality during the study, including:

**Reviewing Survey Firm Technical and Financial Proposals**. A data quality expert can review and comment on survey firm technical and financial proposals in order to assess if the firm is proposing an appropriate methodology, field team composition and work plan, and the budget is adequate given the proposed data collection, entry and management requirements. It is preferable to have these proposals reviewed prior to selection and contract execution, as it is usually very difficult to modify or extend a budget once the contract is signed. Proper review of technical and financial proposals helps mitigate future challenges such as under-estimating time and budget requirements. Issues related to hiring the survey firm(s) are further discussed in Module 4.

**Designing, Adapting and Pre-Testing Survey Instruments**. The first tool to ensure data quality is a well designed survey instrument with appropriate content and formatting. While standardized survey instruments for evaluating RBF have been developed, teams typically underestimate the amount of time that is required to adapt and pilot survey instruments in the country context. If the Principal Investigator and Evaluation Coordinator do not have sufficient time available, the data quality expert can advise on the development and piloting of key in-country survey instruments. Issues related to designing, adapting and pre-testing the survey instrument(s) are discussed further in Module 4.

**Development and Adaptation of Data Entry Program(s).** Data from the field will need to be processed using a data entry program (DEP), and the data quality expert may advise on or support the development of such a DEP. A useful DEP will integrate significant data quality measures such as out-of-range and consistency checks in order to minimize errors introduced at the point of data entry. Issues related to designing the data entry program(s) are discussed further in Module 4.

**Development and Execution of Training Program and Materials**. In order to ensure the quality of data, it is very important that supervisors, field teams and data entry personnel receive sufficient and well-executed training. Principal Investigators and Evaluation Coordinators are not typically very experienced in administering this type of training; therefore, we recommend that the data quality expert participate in and supervise the training of the data collection and entry teams. Issues related to training are discussed further in Module 5.

**Direct Supervision of Data Collection, Management and Entry**. In general, impact evaluations are only as good as the data collected. For this reason, it is crucial to ensure that the data quality measures are respected during data collection, collation (incl. transport of data from the field) and entry. We recommend that a data quality expert directly supervise data collection, management and entry once field work commences. Issues related to data collection, management and entry are further discussed in Modules 5 and 6.

The estimated time commitment for the first round of data collection depends on the work that can be managed by the survey firm. Table 7 outlines the estimated time commitment for a data quality expert.

### Table 7: Estimated Time Commitment for a Data Quality Expert

| Activities | Working Days |
|---|---|
| Review Survey firm Technical and Financial Proposal | 5 |
| Design, adapt and pre-test survey instruments | 20-30 |
| Develop and Adapt Data Entry Program(s) | 5-20 |
| Develop and Execute the Training Program and Materials | 5-20 |
| Supervise Data Collection, Management and Entry | 5-15 |
| Total | **40-90** |

## External Supervisor

While the data quality expert can provide technical expertise during the preparation and at the early stage of implementation of data collection, (s)he is usually an international consultant that will not stay in-country for the whole duration of data collection and entry. However, data quality is not only determined during preparatory stages, but highly depends on the implementation of data collection. Therefore, an external supervisor can be hired locally to assume a more perennial data quality assurance role, especially when the capacity of the survey firm is limited. Under the supervision of the data quality expert (supervision that can be exerted directly in country or later on remotely), the external supervisor is in charge of randomly controlling the carrying out of data collection and entry. The external supervisor allows for a quick response to data quality issues during field work, and maintains high data quality standards over time. The external supervisor reports both to the data quality expert and the Principal Investigator, and can rapidly advise the survey firm on corrective measures when the audit reveals threats to data quality.

The terms of reference of the External Supervisor can be adapted from those of the data quality expert. The estimated time commitment for the first round of data collection is:

**Table 8: Estimated Time Commitment for a Data Quality External Supervisor**

| Activities | Working Days |
|---|---|
| Random Controls of Data Collection, Management and Entry (on site) | 40 |
| Reporting on Data Quality and Implementing Corrective Measures | 15 |
| Total | **55** |

## Qualitative Research Expert

Quantitative analysis can answer whether RBF worked in a particular context for particular outcomes, but in many cases it does not answer why RBF worked or didn’t. In those circumstances, qualitative data can provide more detail on the specific context, insider perspectives, insight into processes and offer new explanations for certain results. Ideally, qualitative analysis should be incorporated into the IE at various points of the project cycle and planned for as part of the impact evaluation. It is outside the scope of this Toolkit to discuss qualitative research methods in depth; however, a qualitative research protocol and instruments will be developed for the second version of the Toolkit.

As qualitative research provides a holistic view of an intervention in the context of society, culture and/or a specific group of people, it typically requires collecting rich data from an “insider’s“ perspective, which requires substantial in-country presence. In many cases, it may be more cost effective to identify a local or regional consultant for this type of work. Depending on the scope of the proposed qualitative work, the estimated time commitment for the first round of data collection is:

Table 9: Estimated Time Commitment for a Qualitative Research Expert

| Activities | Working Days |
|---|---|
| Mission to assess context and identify research questions and methodology | 10 |
| Develop qualitative research protocol and tools | 10 |
| Recruit and train in-country qualitative team | 15 |
| Pilot test survey instruments and methodology | 10 |
| Manage data collection | 20 |
| Transcription, analysis and dissemination | 30 |
| Total | **95** |

## Cost Analysis Expert

Quantitative analysis can inform us whether and to what degree an RBF intervention worked, but in order to decide whether an intervention is worth expanding one also needs to consider its cost. Together with the impact evaluation results, cost analysis allows us to compute the cost-effectiveness, affordability and sustainability of RBF interventions. Currently, there is little information available on the costs and long-term financial requirements of both demand- and supply-side RBF interventions; therefore, we recommend that these data be collected and analyzed in the context of the IE data collection activities. The results of cost analysis can be used to assist policy-makers and program implementers to:

- Compare the costs and outputs of an RBF intervention(s) to business as usual, or other health investments
- Determine whether an RBF intervention(s) is (are) economically worthwhile investments;
- Assess if an RBF intervention(s) is (are) economically and financially feasible to scale-up;
- Evaluate the cost, affordability and possible means of sustaining RBF schemes; and
- Identify areas where possible efficiencies could be gained.

It is outside the scope of this Toolkit to discuss cost analysis and cost-effectiveness analysis in depth; however, a cost-effectiveness analysis protocol and instruments are being developed for the second version of the Toolkit.

The estimated time commitment for the first round of data collection in-country for a cost-analysis component is:

Table 10: Estimated Time Commitment for a Cost Analysis Expert

| Activities | Working Days |
|---|---|
| Mission/Remote support to assess context and identify cost-analysis methodology | 5 |
| Develop/adapt cost-analysis tools | 5 |
| Recruit and train in-country costing team | 10 |
| Pilot test survey instruments and methodology | 5 |
| Monitor data collection | 5 |
| Analysis and dissemination | 15 |
| Total | **45** |

## Involving Local Researchers in the Impact Evaluation

There are many challenges with implementing and managing a large-scale impact evaluation study, as discussed throughout this Toolkit. Depending on where the Principal Investigator and Evaluation Coordinator are based, one way to improve the success of a project's IE is to partner with local researchers. Local researchers may be able to:

1. *Build local ownership and presence of the study.* Even though local representatives in the Ministry of Health (MOH) and project design teams may support the study, they are often far removed from the actual implementation and management of the study. Bearing in mind that a typical impact evaluation lasts 3-5 years, it is crucial that local authorities and partners remain committed to the evaluation design and timeline, and local representatives of the IE team can facilitate this.
2. *Ensure direct and timely supervision for quality assurance*. A wide array of activities throughout the IE project cycle will require in-country engagement and/or direct supervision. By partnering with local researchers, the IE team can ensure consistent engagement with the MOH and other partners, as well as with the survey firm.
3. *Ensure cultural sensitivity and relevance*. The HNP hub has developed an array of resources for the country IE teams, including questionnaires, training materials and protocols. However, all of the these tools must be adapted to the local country context in order to ensure sensitivity to specific cultural characteristics, as well as ensure that the overall IE design and methodology is relevant to the country context.
4. *Build local capacity for impact evaluation*. Large-scale impact evaluations present an exciting opportunity to build local capacity on impact evaluation methodology, survey management and data quality control, reducing the reliance on international researchers. The skills acquired by local researchers in RBF impact evaluations transfer to other types of evaluation and research and to evidence-based policy making.

2012/02/05

## Country Spotlight: Assessing local capacity and needs to build the IE team
## Rwanda Community Performance-Based Financing Project

The planning for the Rwanda Community Performance-Based Financing (PBF) Project began with three joint missions by the project design and impact evaluation teams led by the project task team leader. The coordination between teams allowed for collaborative and consistent dialogue between the World Bank, Ministry of Health and development partners on the policy objectives of the Community PBF project and the related research priorities of its impact evaluation.

The project task team leader built the following team: (i) one principal investigator based in Washington, DC, USA to provide high-level technical support on the design of the evaluation; (ii) one coordinator based in Washington, DC, USA to provide technical support on the design, as well as provide intensive day-to-day support of the management of the evaluation team's time and deliverables and considerable time in-country supporting preparation and implementation; (iii) two researchers based in Kigali, Rwanda to provide technical support on the evaluation design, particularly related to questionnaire development, field sampling strategy and data quality assurance; and (iv) one data collection firm based in Kigali, Rwanda to manage data collection at the community health worker cooperative, community health worker and household levels. Two additional data quality assurance experts were contracted for two specific missions to provide technical support on development of the data entry program, field work management, transporting and entering data: one mission to pilot test the questionnaires and field work management strategy, and one mission following initiation of data collection to advise on on-going processes.

***Throughout the baseline preparation and implementation phases, the involvement of the two local researchers was crucial for addressing several challenges facing the quality of data collection****: (i) Significant support and supervision of survey firm during adaptation and translation of questionnaires, (ii) substantial guidance to the data collection firm on field sampling and field work management of large scale household survey, (iii) maintaining of quality standards through supervision, random spot checks and communication with field workers during field work.*
***Over the course of the preparation of the project, the local researchers were able to maintain dialogue with the Ministry of Health and development partners****, allowing coordinating activities between project implementation and baseline data collection.* ***They also represented the Principal Investigator of the study locally****.*
***The two data management experts conducted an extensive capacity building*** *for the data manager of the data collection firms and impacted the overall data management culture of the organization.*

Full story available: see Country Spotlights section of the Toolkit.

# Module 3

# Designing the Impact Evaluation

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

# Module 3. Designing the Impact Evaluation

**Main Recommendations and Available Tools for this Module**

| Recommendations | Critical | Important | Nice to have |
|---|---|---|---|
| • A prospective impact evaluation should be designed prior to or simultaneously with the intervention. | ✓ | | |
| • Teams should develop a results framework for the RBF project to identify the main pathway(s) by which the RBF program's activities will affect key outputs and outcomes. | ✓ | | |
| • The recommended identification strategy for the RBF Impact Evaluations is randomized assignment to intervention(s) and comparison groups. | | ✓ | |
| • Teams should assess present and future threats to the internal validity of the evaluation (e.g. contamination, lack of power) and monitor them over time. | ✓ | | |
| • Power calculations are an important part of the design of an impact evaluation. Without sufficient power, the impact evaluation may not be able to answer key policy questions. The sample size must allow for sufficient power. | ✓ | | |
| • The sample must be representative of the population that will ultimately benefit from the program. | ✓ | | |
| • When country counterparts buy into the concept of the impact evaluation and understand the importance of respecting the arms of the study, it will be easier to successfully keep treatment and comparison groups intact until the follow-up survey. | ✓ | | |
| • The choice of indicators for the study is critical – each indicator should be measurable with the chosen data collection instruments. | ✓ | | |
| • Teams can refer to *Impact Evaluation in Practice* (Gertler et al. 2011) for in depth discussion on appropriate identification strategies for impact evaluation. | | ✓ | |
| • When deciding on the unit of randomization, teams are balancing the power of the impact evaluation and the risk of contamination across randomization units. | | ✓ | |

**Tools**

- 3.01 RBF Indicators
- 3.02 WHO Output and Outcome Indicators
- 3.03 IE Design Paper Template
- 3.04 IE Budget Template
- 3.05 Ex-ante Power Calculation Example
- 3.06 Binary Power Calculations
- 3.07 Power Calculation References

## Module Contents

As an introduction, it is important to highlight how crucial the timing of the impact evaluation design is. Impact evaluations should be designed before or while the intervention is being designed. Although this means an extra burden for Task Team Leaders during project preparation, it ensures the design of the impact evaluation and the selection of treatment and comparison groups match the design and planned rollout of the program. In addition, it ensures political leverage for preserving the validity of the impact evaluation still exists: when political decisions have been made and publicized, negotiating changes for the sake of the impact evaluation will no longer be possible. To achieve a simultaneous design of the impact evaluation and the intervention, it is key that the IE and project teams collaborate.

In this module, we give an overview of the elements that should be included in the **Impact Evaluation Design Paper**[10]. This paper outlines the building blocks for the evaluation, including the results framework, research questions, identification strategy, data, staffing and budget. The first version of the paper can serve as the Concept Note for the purpose of peer review and approval by World Bank management (cf. Infra). After the concept note stage, the design paper should be regularly updated to reflect the status of the evaluation, any changes to the methodology, or challenges to implementation that affect the impact evaluation.

While in the rest of this module we consider prospective randomized impact evaluations, we acknowledge that other types of evaluations are possible and sometimes less costly, and can rely on existing data already available. More discussion on this is available in the *Impact Evaluation in Practice* handbook (Gertler et al. 2011). Related to this point, we would like to emphasize that the design and feasibility of the impact evaluation is very dependent on the design, timing and coverage of the intervention, the available budget and technical capacity, and the evolution of the policy dialogue at the preparatory stage. Teams should keep in mind that research questions and the design of the IE are interdependent, and the evolution of both is possible (and likely to occur) at the preparatory stage as a result of policy dialogue and further IE feasibility assessments. In this module, we hope to help IE teams design a rigorous randomized prospective impact evaluation. However, we recognize that teams need to be pragmatic, and adjust their design so that the IE remains feasible, rigorous and informative while fitting into the design and evolution of the intervention.

Each of the sections in this module corresponds to a section in the Impact Evaluation Design paper, and we give our recommendations about what we believe each section should contain.

[10] For more information on the technical design elements (such as the identification strategy), readers are referred to *Impact Evaluation in Practice* (Gertler et al. 2011).

**Outline of the IE Design Paper**

- Background and Rationale
- Results Framework
- Research Questions and Policy Relevance
- Output and Outcome Indicators of Interest
- Identification Strategy
- Sample
- Data
- Timeline
- IE Team
- Dissemination Plan
- Budget

## Background and Rationale

This section should answer the following questions:

- What are the main barriers/challenges to reaching the health related MDGs in the country?
- What evidence is available to suggest RBF may be used in this country context to accelerate progress to health-related MDGs?

The background should also include a brief discussion of any current evidence as it relates to the country RBF pilot program:

- Context: Within the country or similar country contexts (Sub-Saharan Africa, Latin America, etc.)
- Design: Beneficiaries, indicators, incentive levels, etc.

## Results Framework for RBF for Health

This section should outline the framework of inputs, activities, outputs and intermediate outcomes that will lead to the program's desired outcomes.

## Research Questions and Policy Relevance

Ideally, the purpose of impact evaluation is to generate evidence on how RBF programs can be used for accelerating progress towards the MDGs, not only within the country where the evaluation is taking place, but also globally. This section should answer two questions about the overall framework of RBF:

- What do we want to learn about RBF in the context where the intervention is taking place?
- How does this contribute to filling the global evidence gap on RBF?

An RBF project may have many components, such as changes in incentives, increased financing, increased supervision and monitoring, and the team will choose to evaluate only a sub-set of these components. If this is the case, this section should also include a brief summary of the RBF scheme and components, the components to be evaluated and the reasons for selecting these components.

## Country Spotlights: Defining the Research Questions of the Impact Evaluation based on Policy Priorities

### Kyrgyzstan Hospital-level Results-Based Financing Program

During consultations with the Kyrgyz Ministry of Health in 2010, it was agreed that measuring the impact of increased financing alone was not a primary policy question for the Ministry. For this reason, the impact evaluation aims to test the effectiveness and cost-effectiveness of RBF as well as one of its constituent components – the enhanced supervision of quality of care.

The primary research questions dictating the design of the impact evaluation are:

- Does the PBF package (including enhanced supervision) at the rayon hospital level improve quality of care?
- Does enhanced supervision *alone* improve quality of care at the rayon hospital level?
- What is the relative cost-effectiveness of the PBF package (including enhanced supervision) vis-à-vis enhanced supervision alone vis-à-vis business-as-usual in terms of quantifiable quality of care indicators?

### Zimbabwe Results-Based Financing Program

Zimbabwe confronts severe limitations and challenges in managing human resources for health in terms of training, financing, monitoring, and retention. The Zimbabwe IE team will explore the relationship between RBF, skill upgrading and capacity building in health facilities.

- What is the causal effect of the simultaneous introduction of results based financing with suspension of user fees on priority population health utilization and outcome measures in RBF districts?
- What is the effect of skill upgrading and capacity building of primary care nurses on priority health outcomes, utilization of services, and quality of care among the populations served, as well as the effect on health worker motivation in rural health facilities?
- What is the combined effect of capacity building of primary care nurses, RBF, and suspension of user fees on the aforementioned outcomes in rural health facilities?

### Cameroon Results-Based Financing Program

The focus is on the effect of RBF and its peripheral enhanced supervision, monitoring and evaluation on the quality of care and health outcomes:

- Does the PBF program increase the coverage of MCH services?
- Does the PBF program increase the quality of MCH services delivered?
- Is it the enhanced monitoring & evaluation and supervision or the link between payments and results that leads to improvements observed in quality or coverage?
- What is the contribution of enhanced supervision and monitoring to improving MCH service coverage and quality in the absence of increased autonomy or additional financial resources?
- Does the PBF program lower informal or formal charges for health services?
- Does the PBF program increase the quantity of funds available at the operational (i.e., facility) level?
- Does the PBF program improve physical and social accessibility of health services? Accessibility of health services will be examined in terms of the convenience of facility opening hours, availability of services through outreach, client perceptions of convenience of accessing health services and client perceptions of health providers' attitudes towards clients?
- Does the PBF program lower staff absenteeism?
- Does the PBF program increase demand generation activities by health facilities?

## Output and Outcome Indicators of interest

Research questions logically materialize into outputs or outcomes of interest. These indicators of interest, which are distinct from the RBF payment indicators, aim at measuring the impact of the RBF intervention with regard to the research questions chosen. They need to be clearly defined at the design stage, for the following reasons:

- Defining indicators allows the team to ensure the research questions of interest are actually measurable.
- It gives a sense early on to the team on what instruments will best allow the measurement of those indicators.

We encourage the teams to rely on international definitions of those indicators when such norms are defined, while considering additional national definitions for the sake of country relevance. Below is a non exhaustive list of references that can be used to define indicators, keeping in mind some of those tools do not aim at evaluating the impact of RBF, but may cover a broader spectrum of health service delivery and utilization issues.

- World Health Organization (WHO) **Indicators Compendium** on Health, which cover maternal and child health, published yearly.
- WHO Protocol on Integrated Management of Childhood Illness: for more information, visit http://www.who.int/maternal_child_adolescent/topics/child/imci/en/index.html.
- Other WHO protocols and guidelines.
- Guidelines and tools focusing on service delivery, such as the WHO Service Availability and Readiness Assessment (SARA), the WHO Service Availability Mapping (SAM) and the Measure DHS Service Provision Assessment (SPA), the Maternal and Child Health Integrated Program (MCHIP) MamaNatalie and NeoNatalie tools.
- National protocols for relevant indicators in-country.

A list of proposed **RBF indicators**, how to calculate them and using what instruments is provided in this Toolkit.

## Identification Strategy

In order to have a successful impact evaluation, each team will need to develop an evaluation strategy that allows for the identification of the causal impact of the intervention. For this to be possible, the strategy will need to include treatment and comparison groups, as well as collection of baseline and post intervention data on treatment and comparison groups.

In the “Identification Strategy” section of the impact evaluation design paper, the following questions should be addressed:

### Which elements of the RBF program will be evaluated?

We recommend that each element to be evaluated should constitute its own so-called “arm” of the study. Additional arms of the study can be used to evaluate the interaction of different components. Thus, the design will need to specify *how many treatment (or sub-treatment) arms will be included in the impact evaluation.* Even when programs are designed with several “complementary” interventions, it may make sense to try and disentangle the separate effect of the different interventions, in order to avoid continuing to implement potentially costly but ineffectual project components.

> **Example 1:** A Performance-Based Financing (PBF) program may involve performance-based payments and health worker training. One could potentially test the impact of different components of the program separately (one treatment arm for performance-based payments, one arm for health worker training). If one also wants to measure the impact of having both performance-based payments AND health worker training, then one would need a third treatment arm (performance-based payments + health worker training) in order to compare the effect of the package to the effect of the individual components.
>
> **Example 2:** When introducing a demand-side incentive program, the Government may be interested in understanding whether the results will differ when the incentive is in cash versus in-kind. One could test the impact of introducing a cash incentive versus an in-kind incentive of the same value with two treatment arms.

### Country Spotlight: Arms of the study
### Rwanda Community Performance based financing program

The Rwanda PBF program includes two types of interventions:

- **Model 1**: Conditional in-kind incentives for women are in-kind payments directly to women conditional on:
  - **Antenatal Care**: Pregnant women accompanied/ referred to health center before or during 4th month of pregnancy: $5 USD value
  - **Institutional Delivery**: Pregnant women who deliver in health facility: $7 USD value
  - **Postnatal Care**: Mother-child pairs receiving postnatal care at health facility within 10 days of birth/discharge: $4 USD value
- **Model 2**: CHW incentives are cash payments to CHW cooperatives conditional upon demonstrated performance on specific health indicators. The indicators are:
  - **Nutrition Monitoring**: % of children (6-59 months) monitored (condition on correct referral to health center for malnourished children),
  - **Antenatal care**: % of women accompanied/referred to the health center for antenatal care before or during 4th month of pregnancy,
  - **Institutional Deliveries**: % of women delivering at the facilities,
  - **Family Planning**: % of new family planning users referred by CHWs cooperatives to the health center,
  - **Family Planning**: % regulars users of modern contraceptives at the health center

For the Rwanda community PBF impact evaluation, the Government wanted to generate evidence on the impact of the proposed demand-side and CHW models separately, as well as the additive impact of combining the two.

## How will the team estimate the counterfactual?

An important aim for determining the effect of the intervention in any impact evaluation is to estimate the counterfactual, this is, what would have happened to the treated group in absence of the treatment/intervention. Given that one cannot *observe* the treated group without the treatment (because, by definition, it is being treated!), one would need to find a comparison group that will allow one to *estimate* what would have happened to the treated group in absence of the treatment. The exact strategy for selecting the comparison group will depend on the operational rules of the intervention. Within the context of the operational rules, the comparison group must be selected to obtain an accurate estimate of the counterfactual: i.e. what would have happened to treatments in the absence of the program. The comparison group should satisfy two requirements: first, the observed and unobserved characteristics of the treatment and comparison groups should be identical on average; second, treatment and comparison groups should have the potential to react in the same way to the treatment and be subjected to the same external shocks over time. When those conditions are satisfied, any differences in the average outcome measurements of treatment and comparison groups following the implementation of the program can be attributed to the intervention.

*Where possible*, assignment of units to the treatment and comparison groups should be done in a randomized way, this is, using a lottery-type mechanism. Additional identification strategies include randomized promotion, regression discontinuity design, and difference-in-differences. Strategies that are not considered proper estimates of the counterfactual include before-and- after comparisons and comparison of enrolled and non-enrolled units. Please refer to *Impact Evaluation in Practice* (Gertler et al. 2011) for an in-depth discussion on appropriate identification strategies for impact evaluation. While randomizing the assignment of a program is not a panacea, in most cases it is the most straightforward and cheapest in terms of data requirements.

## Choosing the unit of randomization

The unit of randomization is the unit that is used for assignment to treatment or comparison groups. Once the unit of randomization is chosen, all entities within the unit of randomization will have the same assignment as the unit of randomization to which they belong. For example, if a country is subdivided in provinces, districts and wards, and the team decides to allocate treatment at the district level, then all wards, health facilities and health workers will have the treatment or comparison status depending on the status of the district to which they belong.

In traditional clinical trials such as medicine trials, the unit of randomization is often the same as the unit of analysis:  the patient gets individually assigned to the treatment or comparison group (hence the patient is the unit of randomization); in addition, the analysis is done at the patient level (the analysis compares patients in the treatment group to patients in the comparison group). By contrast, in evaluations of RBF programs the unit of analysis will often be nested within the unit of randomization. For example, we may decide to assign treatment and comparison status at the level of the health facility, but we may decide to analyze the data at the level of the health worker. In this case, all health workers working in a facility will "inherit" the status of their respective facility. Patients and households may be other units of analysis.

2012/04/03

**Country Spotlight: Unit of Randomization**

**Rwanda Community Performance based financing program**

The unit of randomization for this evaluation was the health sector, which is a geographical area below the district level. All health facilities, community health worker cooperatives, and households in area health sector "inherited" the treatment status of the health sector to which they belonged. Each of the sectors was assigned to one of four study arms: (i) demand-side incentives only, (ii) community health worker incentives only, (iii) demand and community health worker incentives, or (iv) comparison group.

The choice of units of analysis and of randomization will affect the statistical power and validity of the evaluation in an important way. Say, for example, that four districts will be participating in the program, and that those four districts are chosen randomly from a group of 8 districts. Even though the choice of "treatment" districts is randomized, it is very unlikely that the characteristics of the treatment and comparison districts will be balanced, simply because there are only a small number of districts over which to average out the differences. Interestingly, the lack of balance between treatment and comparison group will most likely apply not only for district level indicators, but also for health facility level indicators, even if each district has many health facilities.

When the evaluation can only include a limited number of districts or administrative areas, usually statistical power will be too low to obtain a solid impact evaluation. In those cases, it is worth checking whether randomized assignment can be done at a lower level, for example at the level of the health facility. This will make it easier to balance the characteristics of treatment and comparison groups. However, there are trade-offs in this choice: choosing a lower level of randomization, say the health facility, may be operationally difficult and create spillovers between treated and non-treated unit. For example, if health facilities within one district get randomly assigned to the treatment group or to the comparison group, district level administrators may find it difficult to treat facilities within their districts distinctly. In addition, randomization at such a low level may be politically difficult: for example, health providers may talk with each other and complain about differential treatment within the same district. This communication between health providers may invalidate the assumption that the treatment and comparison units are independent. It may also create contamination between treatment and comparison groups.

2012/02/22

**Country Spotlight: Addressing and preventing compliance issues**
**Zambia Health Results-Based Financing intervention**

Compliance by program implementation with evaluation design is a central theme to all evaluations. Challenges have been noted during the malaria program evaluation [in Zambia]. Yet, this could not prevent compliance challenges experienced during the implementation of the RBF program. The IE team discovered before baseline that one district in the comparison group had been contaminated. To correct this, the team included one additional district in each of the 3 remaining treatment groups. This led to an increase in the IE budget. While large programs do not take place in a laboratory environment in which we can control for everything, there are some factors that often lead to contamination and would be relatively easy to improve on, including coordination, communication, and reporting.

Full story available: see Country Spotlights section of the Toolkit.

Finally, randomizing at the facility level creates another challenge: hospitals are very unique health facilities that cannot easily be compared to lower-level facilities. In general, hospitals are excluded from the evaluation for this reason.

Ultimately, the decision on the level of randomization should reflect the nature of the program being evaluated and depends on the country's IE team. This decision needs to be well justified and explained in the Design Paper.

## Sample

The design paper should answer the following questions:

### What are the inclusion criteria for the sampling frame?

The sampling frame is the set of units from which the sample will be selected. It is important, especially for a pilot RBF program, that the sampling frame be representative of the population that will eventually benefit from the intervention and for whom the desired impact of the intervention is to be determined. This is to ensure that the results will apply to the population of interest, and is known as external validity.

Practically, the sampling frame is a physical or electronic list of units that could potentially be sampled. While it is not possible to assemble that list at the design stage, the design should include the inclusion criteria for the sampling frame. Inclusion criteria are critical because, for equal sample sizes, they can lead to wide variations in the power of the study. At the design stage, the IE team will need to define inclusion criteria for sampling frames at multiple levels.

- For example, if the treatment and comparison units are assigned at the level of the district, one will need to start with the list of districts.
- The next step will be to determine the second level of sampling: e.g. will it be health facilities or households? One can first sample households and select health facilities used by or in the enumeration area of surveyed households. Conversely, one can start by surveying health facilities first and sample households that are within the catchment area of the facilities. Each method has pros and cons and challenges (see table below). The survey team should assess this tradeoff.

<table>
<tr><th>Health facilities sampled first</th><th>Households sampled first</th></tr>
<tr><td>• Easier to list health facilities within the sampling frame. A listing may be available from the MOH or National Statistics Institute.<br>• Remote households may be excluded from catchment area of facilities, with implications on the representativeness of the data.</td><td>• More time consuming and costly to list households within the sampling frame.<br>• May exclude certain health facilities if the matching from households to health facilities is based on actual health seeking behavior and certain health facilities are not visited by sampled households.<br>• May include facilities outside of the randomization unit if households visit them and matching from households to facilities is based on actual health seeking behavior.</td></tr>
<tr><td colspan="2">Define criteria to link households and health facilities<br>• What is the catchment area? And does the perception of the catchment area from the facility match the perception from households?<br>• Actual versus theoretical catchment area: do we consider actual versus theoretical health seeking behavior?<br>• How many households for each facility, or how many facilities for each household?</td></tr>
</table>

2012/02/05

### Country Spotlight: Where to Start in the Sampling frame?
### Benin Results-Based Financing Project

[In Benin] determining the sampling frame raised a few questions and discussions within the team. While the team benefited from political endorsement from the Ministry of Health, they did not obtain the authorization from the National Institute of Statistics to use their census data to build the sampling frame. The team resorted to building the sampling frame themselves, and identified eligible households via preparatory field work. This led to a significant increase in costs, which had not been originally included in the terms of reference of the survey firm and not planned in the original IE budget. It also delayed baseline survey field work.

The team also faced a 2-month long discussion on the connection between households and health facilities: should the team start by surveying households and asking them about the health facility they most frequently used, or should health facilities be surveyed first, and a catchment area defined around each health facility to locate households “attached” to each facility? Since all health facilities were included in the IE survey, no sampling of the facilities was necessary. The team decided to go for the second solution, and defined health facility catchment areas of 5 kilometers in the South and 15 kilometers in the North of the country. For each facility, ten households were surveyed, based on the initial mapping of sampling units conducted that gave the number of households in each sampling unit.

Full story available: see Country Spotlights section of the Toolkit.

- Once the sample of treatment and comparison units have been selected, the team can then decide on the inclusion criteria for health centers, health workers, households, mothers, children, etc. in a waterfall fashion.

  **Example:**

  *Selection criteria for sampling frame of districts:* all districts in the country. Sample: 200 districts, randomly chosen from the sampling frame of districts.

  *Selection criteria for the sampling frame of health facilities:* "All of the health facilities, public or private, located in the selected sample of districts."

  *Selection criteria for health workers:* "The health workers working in a sampled health facility and responsible for maternal and child health services."

  *Selection criteria for households:* Say we believe that RBF will improve the quality of prenatal care, and that we believe this will increase birth weight. In this case, birth weight is our outcome of interest. In our surveys, we will therefore need to obtain birth weight for a large enough sample of births that we believe will be impacted by RBF. Say RBF starts in January 2011 and the endline data collection is planned for August 2012, 20 months later. For a pregnancy to be exposed to RBF during its full duration, it must have started in January 2011 or later. At the endline survey stage, this corresponds to pregnancies that ended 10 months ago or less. Therefore, we need to maximize the number of observations on pregnancies that ended 10 months ago or less. In order to allow a proper comparison between treatment and comparison groups both at baseline and at endline, the criterion for inclusion of households for both surveys will be: "Households with at least one pregnancy that ended 10 months ago or less, living in the catchment areas of sampled health facilities." In practice, it may be difficult to use this criteria because pregnancies that have ended are impossible to observe. An alternative criterion would be "Households with at least one child 0-10 months old". However, one should be aware that this does not include households were there was a pregnancy that ended 10 months ago or less, but where the baby did not survive. If the program affects the child survival rate, relying on the presence of a child that is still alive may introduce bias in the sample.

**Country Spotlight: Building the sampling frame**
**Cameroon Health Sector Support Investment Project**

The study is a blocked-by-region cluster-randomized trial (CRT), having a pre-post with comparison design. Individual health facilities in each region will be randomized to one of the 4 study groups. The study groups are: (T1) full PBF package; (C1) Same per capita financial resources as PBF but not linked to performance; Same supervision and monitoring and managerial autonomy as T1; (C2) No additional resources but same supervision and monitoring as PBF arms and T 1 and C1; and (C3) Status quo. In total, approximately 270 primary care facilities are included in the impact evaluation.

Individuals living in the intervention zone of a facility assigned to a given study group (e.g. C3) may visit a health facility assigned to a different group (e.g. T1). To minimize contamination between study arms, it was decided that facility-level catchment areas were to be used as a sampling frame to randomly select households for the household survey component of the IE. Yet in Cameroon, the most decentralized official geographic health unit is the "Health Area", which can include one, many or even no primary care facilities. As a result, official facility-level catchment areas do not exist. Worse still, household estimates for enumeration zones from the National Census Bureau were unreliable and outdated.

In response to this challenge, all 14 pilot districts were visited (…) to collect information related to health facilities existing in each Health Area, and the villages/town sectors associated with each facility. Data collected also included the population size of each location, and its distance to the associated facility. The sampling frame generated from this work was then used by the Baseline Survey Firm for first-stage sampling of locations within health facility catchment areas. Once in the field, survey teams from then conducted a household listing in each sampled location, where from 16 households who met the study's inclusion criteria (at least one pregnancy or birth in the 2 years preceding the survey) were randomly selected.

Full story available: see Country Spotlights section of the Toolkit.

## What is the required sample size?

The sample consists of those units that the survey firm will try to interview/survey. The sample is pulled from the sampling frame during sampling. At the design stage, one does not need a list of the units that will be been sampled, but one does need to know the intended number of units to be included in the sample, aka the sample size.

The IE team should estimate the sample size needed to be able to detect a minimum acceptable impact on the indicators of interest. This estimation is based on so-called statistical power calculations. Statistical power is the probability of detecting an effect given that it exists. Being a probability, it will always be bounded by 0 and 1 and we want studies designed with as much statistical power as possible (so a statistical power close to 1, usually 80%-90%). Studies with low statistical power will very likely fail to statistically detect a treatment effect even if there is one.

The amount of statistical power of an evaluation study depends on several quantities, including the sample size available at each level of analysis. For studies in which the unit of analysis is the same unit of randomization there is a pretty straightforward relationship between the total number of units and the amount of statistical power: a larger sample results in higher statistical power. However, for studies in which the unit of analysis (e.g. the households) is nested, or clustered, within the unit of randomization (e.g. the health facilities), which will often be the case for evaluation studies of RBF health programs, the relationship between sample size and statistical power is a bit more convoluted. The general guideline is that the number of units of randomization and the number of units of clustering will often be more important than the number of units of analysis within each randomization or clustering unit respectively.

- For example, a study with 100 health facilities and 10 patients for each health facility will have more statistical power than a study with only 10 health facilities and 100 patients for each health facility, despite both studies having a total of 1,000 households (and assuming that there is indeed some nesting of the unit of analysis within the unit of randomization). Thus this nesting or so-called clustering of households within health facilities needs to be accounted for in the power calculations.
- Similarly, if an impact evaluation is randomized at say the district level, and health facilities are clustered within each district, increasing the number of health facilities in the sample will not increase power as much as increasing the number of districts. In countries with a limited number of districts, randomizing at the health facility level may be preferable for that reason, even though it poses other challenges (e.g. contamination).
- In the case of studies with multi-stage clustering (e.g. an impact evaluation randomized at the district level, with health facilities clustered within each district, and households clustered within the catchment area of each facility), increasing units of randomization (e.g. district) or increasing clustering units (e.g. health facility) will increase power more than increasing analysis units (e.g. households).

In summary, power will increase faster in the number of clusters than in the number of units within clusters. It is important to note that the proportion of total outcome variability lying between the clusters, known as the intraclass or intracluster correlation (ICC) in the case of continuous outcomes, is inversely related to the amount of statistical power: a larger ICC means less statistical power. It is crucial to have an estimate of the ICC to conduct power calculations of randomized clustered impact evaluations.

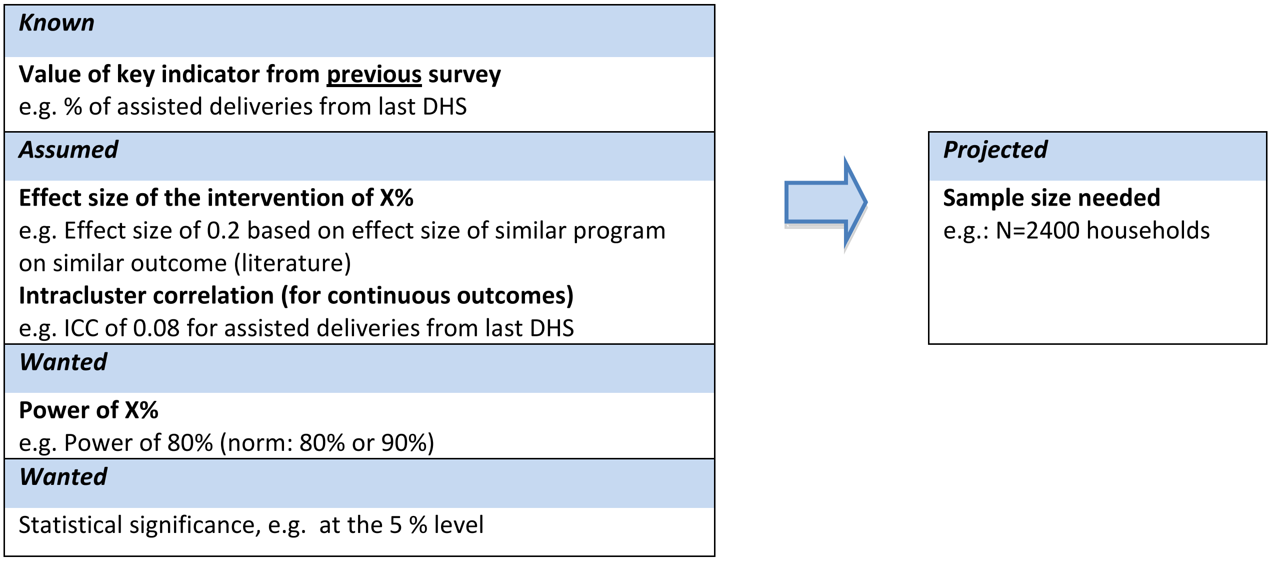
The amount of statistical power of an evaluation study will also depend on the expected size of the intervention effect on the outcome variables of interest. Naturally, smaller effects are harder to detect and thus studies that aim at detecting smaller effects will have—all other things held constant—less statistical power than studies that aim at detecting larger treatment effects. While the size of the intervention effect is often beyond the control of the evaluation team, it must be estimated based on prior knowledge on the nature of the intervention and of the context in which the intervention is being implemented.

Other parameters and study design options that play a role in the power calculations include the baseline values (e.g. for continuous outcomes the mean and standard deviation) of the chosen outcome indicators, the number of arms in the study, whether the design is balanced or unbalanced (e.g. same number of units in each treatment arm), whether the sampled units will be chosen fully at random or following a stratified design (e.g. 1000 randomly chosen patients versus 10 patients in each of 100 health facilities), whether a blocked design was used (e.g. whether the units of randomization were classified by poverty ranking or any other ancillary information before randomization to create more comparable study arms) and whether covariates are included in power calculation.

Based on previous studies, we recommend that the team use a minimum power of 0.9 with a significance level of 0.05. The technical paper should also include the definitions of indicators and the data sources used to generate the power calculations, the sample size estimations at every level of observation (health facility, health worker, household, woman, etc.), and a detailed description of the adopted study design.

There are two different types of power calculations, usually referred to as “ex ante” and “ex post”. Ex ante power calculations take into account the values of key variables from available surveys, the size of the expected effect of the intervention, the desired power of the study (80 or 90% in general) and the desired statistical significance level (0.05 in general) to calculate the required sample size for the baseline.

**Figure 5: Ex-ante Power Calculations**



***Known***

**Value of key indicator from <u>previous</u> survey**
e.g. % of assisted deliveries from last DHS

***Assumed***

**Effect size of the intervention of X%**
e.g. Effect size of 0.2 based on effect size of similar program on similar outcome (literature)
**Intracluster correlation (for continuous outcomes)**
e.g. ICC of 0.08 for assisted deliveries from last DHS

***Wanted***

**Power of X%**
e.g. Power of 80% (norm: 80% or 90%)

***Wanted***

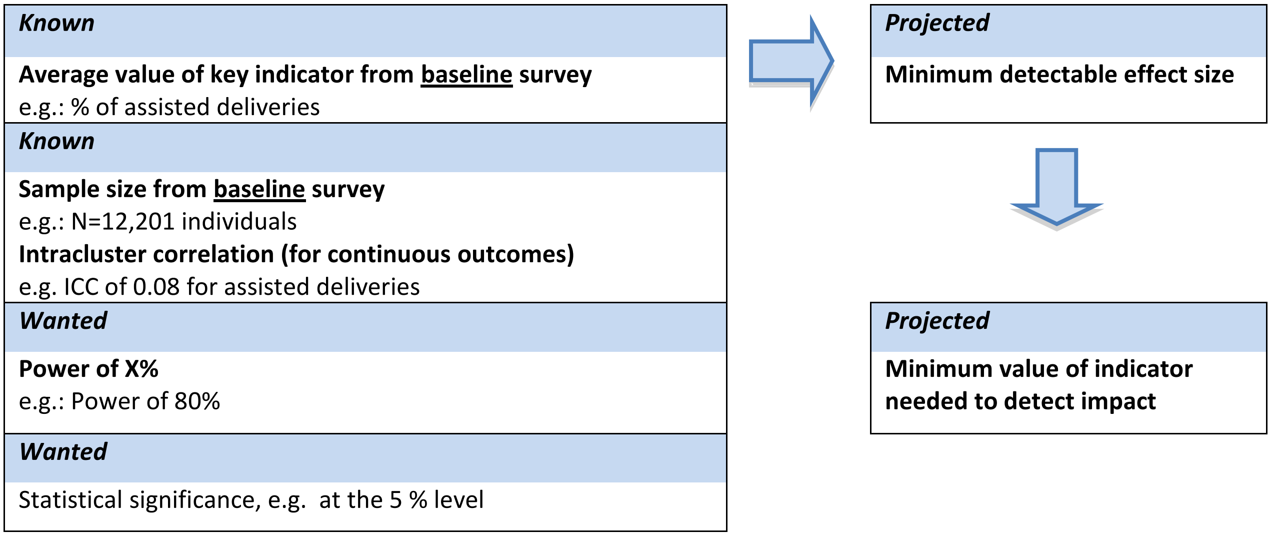
Statistical significance, e.g. at the 5 % level

***Projected***

**Sample size needed**
e.g.: N=2400 households

Another approach is to use “ex post” power calculations. Despite the name, “ex-post” power calculations do not need be undertaken after the design stage of the study. For some studies it is possible that from the outset there are some known limitations in terms of the sample sizes that can be included in the study. In these studies it may be informative to approach the power calculations in a way that allows determining the smallest intervention effect the study will be able to statistically detect given the available sample size. For studies with serious sample size limitations this approach will help determine if the study is worth undertaking. It is crucial to avoid conducting studies that from the outset are known to have low statistical power.

Figure 6: Ex-post Power Calculations



***Known***

**Average value of key indicator from <u>baseline</u> survey**
e.g.: % of assisted deliveries

***Known***

**Sample size from <u>baseline</u> survey**
e.g.: N=12,201 individuals
**Intracluster correlation (for continuous outcomes)**
e.g. ICC of 0.08 for assisted deliveries

***Wanted***

**Power of X%**
e.g.: Power of 80%

***Wanted***

Statistical significance, e.g. at the 5 % level

***Projected***

**Minimum detectable effect size**

***Projected***

**Minimum value of indicator needed to detect impact**

2012/04/03

**Country Spotlight: Arms of the study (cont'd)**

**Rwanda Community Performance based financing program**

Rwanda is divided into four provinces, 30 districts and over 400 sectors. Keeping in mind the intervention would be implemented at the sector level, power calculations demonstrated that the impact evaluation required 50 sectors for each of the four study arms, resulting in a total of 200 sectors. The following inclusion criteria served to define the sampling frame for sectors: (i) minimizing geographic disbursement by limiting the number of districts to 19 (from universe of 30), (ii) excluding 30 sectors which were purposively selected to receive the treatment at the request of the MoH, (iii) sectors had to have at least one health center. A total of 223 sectors in 19 districts satisfied those criteria. Since only 200 sectors were needed for the study, the IE team used statistical software (STATA) to randomly select 50 sectors to be assigned to each for the four study arms. The remaining 23 sectors were assigned to "reserves".

Note that most indicators of interest regarding maternal and child health are non-continuous outcomes bounded between zero and one (e.g. percentage of deliveries assisted by skilled provider). Power calculations are usually not conducted the same way for continuous or non-continuous outcomes.

An example of **Ex-ante Power Calculations** and a more detailed explanation of **Binary Power Calculations** for an RBF impact evaluation are included as tools in this Toolkit.

We recommend that the IE teams include a basic diagram that illustrates the arms of the study and the proposed sample size at each level of the intervention. For example:

**Country Spotlight: Arms of the study (cont'd)**

**Rwanda Community Performance based financing program**

| | |
|---|---|
| T1<br>In-kind incentives only<br>*50 sectors*<br>*600 households*<br>*1200 community health workers* | T2<br>CHW incentives only<br>*50 sectors*<br>*600 households*<br>*1200 community health workers* |
| T3<br>In-kind incentives + CHW incentives<br>*50 sectors*<br>*600 households*<br>*1200 community health workers* | C<br>Comparison group<br>*50 sectors*<br>*600 households*<br>*1200 community health workers* |

## Data Requirements

This section will answer the following questions:

- What is the type of data that will be collected and analyzed? (quantitative and/or qualitative)
- What are the sources of data?
- At what level(s) will the data be collected (facility, household, etc)?
- Who will be interviewed? I.e. who will be the respondents?
- What kind of data will be collected?
- What survey instrument(s) will be used?
- How frequently will the data be collected?

While there are endless possibilities of surveys at various levels of the intervention, we will describe five potential surveys that can be fielded during the baseline and endline data collection for RBF. These are:

***Health Facility Survey*.** The healthcare facility survey measures (i) the main characteristics of the facility, typically including staffing, infrastructure, service availability, structural and process quality, etc., (ii) health worker characteristics such as training, salary and time use, knowledge and practice, satisfaction and motivation, and (iii) the quality of care delivered through patient exit interviews. Data should be collected on all health facilities in the evaluation sample, and may require 1.5-2 days per facility.

***Household Listing*.** Often the IE team may not have access to a recent census to construct the household sample. In those cases, the survey firm will be required to conduct a listing of all households in the selected Primary Sampling Units (PSU) through a rapid screening and listing survey. Enumerators should collect basic information from each household in the PSU, taking no more than a few minutes per household to assess household eligibility, basic demographic and re-contact information.

***Household Survey.*** A complete household survey is usually implemented on a random sample of households in the selected enumeration areas of the evaluation sample. It should take on average 90-120 minutes per household to implement. Information should be collected through interviews as well as direct observation. The household survey is described in more detail later in this module.

***Biometric Indicator(s).*** Household biological indicators of health and welfare may include the following indicators:

- Height and weight for a specific sub-sample of household members: Pregnant women, children < 5 years old.
- Hemoglobin measurement (a way to test for anemia) for a specific sub-sample of household members: Pregnant women, children < 5 years old.
- Malaria tests for a specific sub-sample of household members: Pregnant women, children < 5 years old.

***Community Questionnaires.*** A community survey typically covers community leaders in the sample clusters, and takes approximately 60-90 minutes to administer. The survey is normally collected through interviews with 1-2 community leaders or key respondents in each community, and includes information on community characteristics, services, infrastructure, access to markets, prices and community-level shocks.

We suggest that teams use the format in Table 11 to summarize data requirements for the impact evaluation:

### Table 11: Format for Data Requirements

| Data | Type | Source | Level | Respondents | Description of Data | Survey Instrument | Frequency |
|---|---|---|---|---|---|---|---|
| **Example 1** | Quantitative | HMIS | Facility | N/A | Monthly vaccination coverage of children <5 years | N/A | Quarterly |
| **Example 2** | Quantitative | Primary | Facility | Health care workers | Staff work load, compensation, motivation and KAP | Health Facility Questionnaire | Twice (baseline and follow up) |
| **Example 2** | Qualitative | Primary | Facility | Pregnant women | Focus group of pregnant women to understand health seeking behavior for prenatal, delivery and postnatal services | Focus Group Questionnaire | Twice (baseline and follow up) |

## Timeline

In order to ensure that the operational design of the intervention and the evaluation design are consistent, the IE team should be engaged early in the program planning process. The design paper should address the following questions:

- When will baseline data be collected (if primary) and/or extracted (if secondary)?
- When will follow up data be collected (if primary) and/or extracted (if secondary)?
- What is the commitment of the Government and other key counterparts to adhere to this timeline to ensure (i) that the baseline data collection is completed before any activities take place on the ground; and (ii) that there is sufficient time between the start of the intervention and the endline survey so that the treatment group will be subjected to the treatment long enough in order for their outcomes to be impacted?

### Figure 7: Sample Timeline for an Impact Evaluation Design Paper

| Phase | Year 1 | | | | Year 2 | | | | Year 3 | | | | Year 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 |
| Program design | | ▨ | ▨ | ▨ | | | | | | | | | | | | |
| Impact evaluation design | | ▨ | ▨ | ▨ | | | | | | | | | | | | |
| Evaluation preparation | | | | ▨ | ▨ | | | | | | | | | | | |
| Baseline data collection | | | | | | ▨ | ▨ | | | | | | | | | |
| Initiation of RBF intervention | | | | | | | | ▨ | | | | | | | | |
| Exposure to RBF treatment | | | | | | | | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | | | |
| Baseline data documentation and storage | | | | | | | | ▨ | | | | | | | | |
| Baseline analysis and report | | | | | | | | ▨ | ▨ | | | | | | | |
| Evaluation preparation | | | | | | | | | | | | ▨ | ▨ | | | |
| Endline data collection | | | | | | | | | | | | | | ▨ | ▨ | |
| Endline data documentation and storage | | | | | | | | | | | | | | | | ▨ |
| Impact analysis and report | | | | | | | | | | | | | | | | ▨ |

While a general timeline is adequate for the purpose of the IE design paper and concept note, the IE team should develop a more detailed Gantt chart which summarizes all key activities under the IE and their corresponding timeframe. The IE Gantt chart is discussed in Module 4.

## IE Team

The design paper should also detail the composition of the IE team. (Cf. Module 2) In summary, the following information would need to be included:

- Who will lead the impact evaluation (Principal Investigator)?
- Who will provide technical support to the Principal Investigator? (Co-Principal Investigator, Evaluation Coordinator, Data Quality Expert, Research Assistant, etc.)
- Who or what organization may be involved as local research counterparts for study design and management?
- Who will lead complementary research activities (e.g. cost analysis, qualitative research)?

## Dissemination Plan

The IE team should also have a plan on how results from the baseline analysis, as well as initial impact analysis results, will be disseminated.

At a minimum, the IE team should plan for the following dissemination activities:

- Baseline Descriptive Report.
- Baseline Brown Bag Lunch (BBL) or other presentation.
- Impact Analysis Report.
- Impact Analysis Workshop.
- Impact Analysis BBL.

## Budget

The budget for an evaluation is an important part of its design, and we recommend that it include the expected expenses for the following components of the impact evaluation:

- *Impact Evaluation Team*. The budget should include all staff and consultant time for managing the impact evaluation, including design, implementation and analysis.
- *Data collection***.** The team should identify all primary and secondary data collection requirements and provide a budget for completing it (minimum baseline and follow up data), including qualitative and/or cost-analysis data collection requirements when applicable.
- *Travel***.** The budget should include all necessary travel costs for required project supervision, including air, hotel and subsistence.
- *Technical Assistance***.** The budget should include any additional consultant time and travel for technical assistance (such as survey instrument development, data quality control, data entry program development)
- *Dissemination Plan.* The budget should include any costs associated with travel or logistics for at least one field-based presentation at baseline and one at follow-up, as well as any costs associated with producing written materials.

- *Miscellaneous***.** The budget should include any additional costs related to the impact evaluation, such as payments for Institutional Review of the research protocol.

The Toolkit includes an **IE Budget Template** to help in drafting the IE budget.

2012/02/22

**Country Spotlight: Addressing financial challenges during data collection with pragmatism**
**Zambia Health Results-Based Financing intervention**

Given Zambia's sparsely populated geography in a number of districts, poor road networks in general and especially during the rainy season, the high cost of gas, and price volatility due to exchange rate fluctuations, transportation has been an important cost driver within the IE budget. While the knowledge spillover from the malaria to the HRBF survey has contributed to containing costs, such as through reduced costs for product development and training, the variable costs for transportation have not improved as much. This is not fully within the team's control but efforts such as end-of-day data quality checks have helped reducing travel costs. Information regarding case load can help field entry planning to reduce transport costs [for additional visits when too few patients were interviewed in health facilities].

Full story available: see Country Spotlights section of the Toolkit.

2012/02/05

**Country Spotlight: Adjusting timeline and budget on time**
**Benin Results-Based Financing Project**

Benin is starting to implement a Results-Based Financing (RBF) intervention in 34 districts. The intervention consists in combining different measures: increasing management autonomy, and/or lump sum staff bonuses, and/or RBF bonuses. The combination of those measures defines five study arms: (i) Increased autonomy + RBF bonuses, (ii) Increased autonomy + lump sum bonuses, (iii) RBF bonuses only, (iv) Lump sum bonuses only and (v) Status quo (comparison group).

During the preparatory stage, the status quo comparison group was added last minute to the IE design, mainly because the team was concerned that there would be small differences between the four initial study arms. This implied sudden increases in the budget and the time needed to complete the baseline survey. The team revised their timeline and requested additional budget in order to conduct the experiment and the survey as needed.

In addition, while the team was preparing to implement the intervention and the IE, other multilateral donors became interested in financing RBF interventions in the country. The integrity of the comparison group got threatened in the early stages of the IE, and the team had to negotiate a delaying in other donors' activities in order to keep the comparison group long enough.

Full story available: see Country Spotlights section of the Toolkit.

# Module 4

# Preparing the Data Collection

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

# Module 4. Preparing the Data Collection

| Main Recommendations and Available Tools for this Module | | | |
|---|---|---|---|
| **Recommendations** | **Critical** | **Important** | **Nice to have** |
| • The research protocol should contain all relevant information related to the protection of human subjects, including specific sampling criteria, informed consent and data confidentiality protocols. | ✓ | | |
| • The impact evaluation must be approved by an Institutional Board: the Principal Investigator should plan for contracting this board to conduct the ethical review and approve the research <u>prior</u> to the beginning of field activities. | ✓ | | |
| • A Project/impact evaluation Gantt Chart can help teams coordinate activities and timelines from the project and from the impact evaluation. | | ✓ | |
| • The impact evaluation team should agree with Government counterparts what will be the policy of accessing the data from the impact evaluation. A written Memorandum of Understanding can help prevent misunderstandings. | | ✓ | |
| • The decision between CAFE and centralized data entry has major implications for the selection of the survey firm and should be decided in advance of survey firm procurement. | | ✓ | |
| • Hiring a survey firm is a time intensive process, which typically requires 3-6 months and should be initiated in the early stages of project planning. | ✓ | | |
| • Depending on the situation and expertise in country, it may be preferable to hire one survey firm that would conduct both health facility and household surveys, or for two separate firms. As a general rule, we recommend that teams use a competitive selection process. | | ✓ | |
| • The survey management team should include a Project Manager, a Field Manager and a Data Manager during the full duration of the preparation and implementation of the data collection. | ✓ | | |
| • Negotiations with the survey firm require a clear understanding of budget and time constraints, which have implications for field team composition and survey duration. | ✓ | | |
| • The survey firm should be supported from the early stages of survey preparation by a data quality expert, especially in local survey firms with limited capacity. | | ✓ | |
| • The structure and quality of the survey instruments are crucial for data quality and comparability of results across countries. We recommend that project teams use the RBF Facility and Household questionnaires as a basis. The Principal Investigator of the evaluation should determine which modules are appropriate and which are not, and ensure key outcomes of interest can be calculated from the questionnaires. Teams should feel free to make the adjustments that they deem necessary. | | ✓ | |

**Main Recommendations and Available Tools for this Module**

| Recommendations | Critical | Important | Nice to have |
|---|---|---|---|
| • The toolkit questionnaires are meant to be comprehensive – teams may want to limit the number of modules to limit the cost and time requirement for administering the questionnaires. | | ✓ | |
| • Community surveys can allow measuring infrastructures and existing support networks within the community. They can also be used as a complement to household surveys, especially when household surveys need to be drastically shortened. | | | ✓ |

**Tools**

- 4.01 Impact Evaluation Gantt Chart
- 4.02 Data Access Memorandum of Understanding
- 4.03 Research Protocol Example
- 4.04 Informed Consent Templates
- 4.05 Health Facility Survey Firm TOR
- 4.06 Household Survey Firm TOR
- 4.07 Data Collection Budget Template
- 4.08 Consumables and Equipment for Biomarker Data
- 4.09 Health Facility Questionnaires
- 4.10 Household Questionnaires
- 4.11 Community Questionnaires
- 4.12 Costing Questionnaires
- 4.13 Data Entry Program
- 4.14 Anemia Referral Guidelines
- 4.15 Anemia Referral Form
- 4.16 How to Translate Questionnaires
- 4.17 Institutional Review Board TOR
- 4.18 Certificate of Accurate Translation

## Module Contents

In this module, we give an overview of the steps involved in preparing for the baseline data collection. We cover all of the steps should be taken after the concept note is approved until the survey firm comes on board. These steps include: (i) planning the evaluation and its baseline survey; (ii) defining the research protocol and obtaining ethical clearance; (iii) defining the data entry strategy; (iv) hiring a survey firm for the baseline survey; (v) understanding and developing the questionnaires.

## Planning the Activities using a Gantt Chart

For a baseline to accurately represent the "before treatment" situation, it is crucial that all data collection be complete before the RBF intervention starts, especially in the chosen treatment groups. We recommend that one start with the desired start date for the intervention and work backwards to plan all of the activities that need to take place in order to successfully complete the baseline. This can be done using the **IE Gantt Chart** template. On the basis of experience with other surveys, here are some tips on how long key processes will take:

**Tip 1: Allocate sufficient time for procurement of survey firms** - Due to the size of the budget, the procurement of survey firm(s) will almost always be competitively selected, and such a process can take from **4-6 months** to complete.

**Tip 2: Allocate sufficient time for procurement of equipment and consumables** - The materials for anthropometric and biomarker testing (such as anemia or malaria) are not available in many countries. While these materials can be purchased and managed by the survey firm(s), this process can take **2-3 months** to complete.

**Tip 3: Allocate sufficient time for obtaining ethical clearance** - Some institutional review boards (IRB) have set schedules for reviewing in-country ethical clearance proposals. In most cases, the research package - including research protocol, questionnaires, and informed consent templates - must be submitted a number of weeks before the study can be reviewed by the IRB; the time often depends on the specific IRB. Teams should plan **2-3 months** to obtain ethical clearance.

## Defining Ownership of Data

Defining ownership of data can be a complicated matter, because large surveys typically involve many interested parties with a stake in the data:

- Governments: because the data because they were collected on their territory.
- Investigators because they invested time and intellectual effort into data collection.
- Funders/sponsors: may each have their own implied assumptions about who owns the data.
- Survey firms: may insist on reserving the rights to the data in their contracts.

All of these examples show that it is very important to have a clear understanding of the topic of access to the data and ownership of the data. This can be achieved through a **Memorandum of Understanding (MOU)** between the Government and the Bank or other financier before the start of the data collection. The memorandum of understanding should cover the following topics:

- Who will own the data once they are collected?
- What are the agreements on access to the data? Within what time span will the data be unavailable, available for licensed use, or available publicly? What will be the conditions for licensed use and usage during the "unavailable" time? More details on the different types of access to the data are given in Module 6 on data storage.
- Agreements on storage and preservation of the data.
- Who will manage the data?

Investigators and survey firms should never "own" data collected during an impact evaluation. However, it is possible to make certain provisions for investigators in the access to the data agreement so that they will have sufficient incentives to invest their time and effort into the data collection, while also ensuring that the data will be available for further use within a reasonable time span.

A further discussion on how to make data available after data collection, and on data access policies, can be found in Module 6.

## Protecting Human Subjects

While impact evaluations are linked to project operations and Government interventions, it is also a research activity that involves "human subjects". The households, doctors, nurses and administrators that respond to questionnaires are subject to harm if the information they provide is made publicly available without sufficient safeguards. As such, impact evaluations need to be conducted within good practice of "human subjects protection".

**Examples of Violations to Protection of Human Subjects:** (i) During the administration of a survey, a woman may be endangered if her husband overhears her divulging confidential information regarding her family planning practices. (ii) The safety of households may be jeopardized if individuals are able to identify specific families' income or asset holdings from data posted on the internet. (iii) The study did not disclose information on the risks of biometric tests. (iv) A participant in the survey asked to withdraw from the study half way through the survey but was instructed to finish the interview by the enumerator.

### Basic Principles of Human Subjects Protection

It is the responsibility of the principal investigator and other investigators to safeguard the rights and welfare of human subjects involved in research in accordance with the appropriate national code of ethics or legislation. In the absence of national ethical guidelines, the investigator should be guided by the Helsinki declaration adopted by the Twenty-Ninth World Medical Assembly in Tokyo (October 1975) and Article 7 of the International Covenant of Civil and Political Rights, adopted by the United Nations General Assembly on 16 December 1966.

**WHO criteria:** The basic criteria recommended by WHO for assessing the research projects involving human subjects include:

- the rights and welfare of the subjects involved in the research should be adequately protected;
- freely given, informed consent should be obtained;
- the balance between risk and potential benefits involved should be assessed and deemed acceptable by a panel of experts independent of the institution(s); and
- any special national requirements should be met.

**In the USA**, the following three principles form the foundation of guidelines for the ethical conduct of human subjects research[11]:

- Respect for persons: How will the researchers obtain informed consent from their research subjects?
- Beneficence: How will the researchers ensure that the research (i) does not harm and (2) maximizes potential benefits and minimizes potential harms
- Justice: How will the researchers ensure that the benefits and burdens of research are fairly and equitably shared?

These principles are based on the 1979 Belmont report on Ethical Principles and Guidelines for the Protection of Human Subjects of Research, available at http://www.hhs.gov/ohrp/policy/belmont.html.

## Human Subjects Training

The United States National Institutes of Health (NIH) recommends that all Principal Investigators, co-Principal Investigators and research coordinators be trained in the protection of human subjects and that they take yearly refresher courses. The online NIH training is very informative and only takes about one hour to complete. It is available at http://phrp.nihtraining.com/users/login.php and www.ohsr.od.nih.gov.

[11] These principles are laid out in the Belmont Report, which was drafted by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in 1974.

Please note that HRITF funds for baselines and follow up surveys can only be transferred to the teams if the proposed Principal Investigator, co-Principal Investigator and research coordinator can provide evidence of recent human subjects training. The NIH online course includes a test, and upon completion a certificate number will be generated, which can be used for this purpose.

Another source for human subjects training is the CITI program. CITI offers international IRB courses in several languages, though the program has a fee (75 USD per person). Please see www.citiprogram.com.

## The Research Protocol

The **Research Protocol** details the purpose of the study, evaluation objective, methods and procedures, and lays out how the researchers will ensure that human subjects are protected. As such, it is one of the most important written documents in an evaluation's documentation.

The research protocol is used by different stakeholders in the evaluation:

- **The IE and project teams** use it as a record of methods.
- **The survey firm** uses it as a guide for field methods.
- **Analysts** rely on it to understand how the data were generated.
- **Ethical Review Boards /Institutional Review Boards** (IRBs) rely on it to determine whether or not they can clear the proposed research.

### What is an Ethical Review Board?

An **institutional review board** (**IRB**), also known as an **independent ethics committee** (**IEC**) or **ethical review board** (**ERB**), is a committee that has been formally designated to approve, monitor, and review biomedical and behavioral research involving humans.

The purpose of an IRB review is to assure, both in advance and by periodic review, that appropriate steps are taken to protect the rights and welfare of humans participating as subjects in a research study. IRBs attempt to ensure protection of subjects by reviewing research protocols and related materials. IRB protocol review assesses the ethics of the research and its methods, promotes fully informed and voluntary participation by prospective subjects capable of making such choices (or, if that is not possible, informed permission given by a suitable proxy), and seeks to maximize the safety of subjects. (Wikipedia)

Many universities have their own institutional review board, and many countries also have a national ethical review board.

The structure of a research protocol can vary depending on the requirements of the Ethical Review Board / Institutional Review Board, but it normally includes the following elements:

1. **Purpose of the study:** This section outlines the proposed intervention, the current evidence gap as to the impact of the proposed intervention, and how the proposed impact evaluation will address this gap. This section will be similar to the introduction and motivation section in the IE design paper.

2. **Evaluation Objectives, Policy Questions and Methodology**: This section should detail the IE objectives and the policy questions that will be addressed in the evaluation, and the selected methodology for isolating the causal impact of the proposed intervention on intended outcomes.

3. **Analysis Plan**: This section should detail the econometric methods that will be used to conduct the impact analysis.

4. **Subject Selection**: This section is an extension to the Sample section of the IE design paper and overlaps with the Field Sampling Plan (see Module 5). It should outline the following:

   - ***Number of subjects:*** This is the number of subjects that are required at the facility, provider, household and individual levels, according to the study requirements.
   - ***Inclusion criteria:*** The inclusion criteria are those criteria that are used to decide whether a unit is or is not to be included in the sampling frame. Inclusion criteria are critical because, for equal sample sizes, they can lead to wide variations in the power of the study.
   - ***Definition of the Sample According to Age, Gender and Racial Origin:*** The IE team is required to fully detail the inclusion criteria and justify any exclusion from the study.
   - ***Protection of Vulnerable Subjects:*** The IE team should identify which part of the sample is considered vulnerable and what the team will do to minimize risks to those subjects. Most importantly, the IE team should be concerned about minimizing the risk of coercion of subjects who are poor, illiterate, children, or otherwise defined as vulnerable, to take part in the study.
   - ***Privacy and Autonomy:*** This section should identify the measures the survey team will take to protect respondent privacy, confidentiality and autonomy.

5. **Methods and Procedures:** This section should detail the data that will be collected (including the required biomarker data such as anemia, malaria, and anthropometric measurements), the field procedures for selecting the households to be interviewed, and the required respondents. It should also detail how the survey firm will collect and manage these data, including the procedures it will use for storing the data and ensuring confidentiality in the field, during data entry and upon completion of the study.

6. **Risk/Benefit Assessment**: The IE team should assess any potential risks (indirect or direct) and benefits (indirect or direct) for study participants. Risks may include loss of confidentiality, harm from anemia, malaria, or anthropometric testing. In most RBF impact evaluations, the risks to study participants should be minimal. Benefits may include a small compensation for participant time (based on local custom) and an assessment of child health status from immediate results of anemia and malaria tests.

7. **Subject Identification and Recruitment**: This section should detail how the field team will assess potential participants' capacity to comprehend the study procedures and survey instruments. If a field team finds that a potential participant has limited capacity or comprehension of the study, then they should replace that household with the next available eligible household, based on the inclusion criteria and sampling strategy.

8. **Documentation**: The research protocol should include any relevant documents or tools as annexes, including but not limited to: (i) questionnaires, (ii) informed consent forms, (iii) child health cards, and (iv) description of respondent benefits or compensation package if applicable.

The WHO published guidelines on how to write a research protocol for research involving human participation at http://www.who.int/rpc/research_ethics/guide_rp/en/index.html.

## Informed Consent

Informed consent is one of the cornerstones of human subjects' rights in any study or intervention. It requires that respondents have a clear understanding of the purpose, procedures, risks and benefits of the study. By default, ***informed consent*** by an adult[12] respondent requires a written document (form) that includes a section on the methods used to protect respondent confidentiality, a section on the respondent's right to refuse or cease participation in the study at any point in time, an explanation of potential risks and benefits, contact information for the event the respondent wishes to contact the survey team or investigator, and space for the respondent to record their formal written consent to participate with a signature. The field team keeps one copy of the informed consent form for documentation purposes, and leaves one copy with the respondent after the interview. In contrast to able adults, *minors*[13] cannot consent to participate in a survey; they may ***assent*** to participate ***after*** written ***permission*** by their parent or guardian. Assent and permission are verified in the same formal written way as consent. Table 12 summarizes the expected respondents for the RBF impact evaluation data collection activities, and required verification.

[12] As defined by country law, usually individuals over 18 years of age

[13] Or otherwise defined by country law, usually individuals under 18 years of age

Table 12: Survey Instruments, Respondents and Verification Methods

| Survey Instrument | Respondent | What verification is needed? |
|---|---|---|
| Health Facility Assessment | Health Facility Manager and/or Administrator | Informed consent |
| Health Worker | Individual Health Worker(s) for service of interest | Informed consent |
| Patient Exit Interview | Patients for service of interest | Informed consent<br>Permission  from adult guardian + assent from child if child is being examined |
| Main Household Modules | Head of household and/or spouse. Includes proxy responses for all household members for education, labor market activity and health utilization. | Informed consent |
| Female Health Modules | All women in household 15-49 years old | Informed consent |
| Child Health Modules | Mother and/or caregiver for children 0-5 years old | Permission  from adult guardian + assent from child if child is being examined |

The IE team may opt to request two waivers from the requirement to obtain formal written consent/assent from respondents.  These waivers would be requested from the Ethical Review Board.

**Waiving formal written consent among eligible, potential adult respondents**: For the main study, IE teams may request a waiver of formal written consent in the form of a signature, and request to replace it with documented verbal consents. If possible, the enumerator should document any verbal consent with an accompanying signature from the supervisor (as a witness). IE teams can use the following arguments to justify documented verbal consents. First, most of the activities in the study present no greater risks of harm than those encountered in daily life, and secondly, the IE team will ensure the field staff is trained extensively in the proper presentation of the verbal consent and study introduction. In such a case, each family will be provided with a copy of the consent form so they can contact the study staff should they have any future questions about the study (illiterate families will be able to take the card to literate members in their community for assistance). Thirdly, the scientific validity of the study could be compromised by non-response if non-literate participants refuse to participate on grounds that they do not wish to sign their name on a document that they cannot read and if non-literate individuals end up consenting to participate at a lower rate than literate individuals. This is important because literacy is a key factor in many of the causal processes that this study plans to evaluate, such as health care utilization, quality of care received.

**Waiving written assent for children**: The IE team may request a waiver of children’s written assent for the child participants under the age of 5 years in cases where the survey asks questions about the health of children under age 5 years but it would be unreasonable to assume that the infants and children included in the survey will understand the risks and benefits of participating. For those sections, permission from the parent on behalf of the child should be sufficient. For survey modules that require direct contact with children (e.g. anthropometric measurement, hemoglobin measurements, and malaria testing), the Ethical Review Board may wave written assent from children. However, if children refuse to be measured or tested by field staff, field staff should respect their wishes and use the "refusal" code when filling in the questionnaire forms. Please note that parental permission for participation of a minor in the survey cannot be waived.

**Informed Consent Templates** are available for teams.

## Protecting Respondent Confidentiality

All information provided during the course of the interview is strictly confidential, and although results of the study may be published for scientific purposes, it should be written in such a way that identification of an individual or household is not possible. To ensure confidentiality, each subject of the survey should be assigned a unique identification number (ID) and all names and identifiers should be deleted from the database that is used for research purposes. The assignment of IDs is discussed in Module 5.

## Obtaining Ethical Clearance

**Clearance in Country**: The Principal Investigator is responsible for identifying all the institutions which require country clearance or approval of the study, particularly when the study requires household or individual level data collection. Typically, the team will be required to obtain ethical clearance from the respective country's Ethical Review Board or Institutional Review Board. The following International Compilation of Human Subject Protections is a useful resource for identifying and contacting international and national boards: http://www.hhs.gov/ohrp/international/intlcompilation/hspcompilation-v20101130.pdf

The IE team should also work with its respective MoH counterpart to ensure compliance with all country-level requirements for conducting research on human subjects. In many countries, the IRB will require both the Principal Investigators and the survey firm to present the study and its compliance with research guidelines.

**Clearance in the United States**: While the World Bank does not have its own IRB there are two additional possibilities for ensuring the IE protocol adheres to international standards. First, Principal Investigators based in academic institutions (such as Johns Hopkins University, University of California Berkeley) are required to go through their IRB for clearance to participate in the study. Secondly, the IE team can contract an independent US-based IRB in order to provide third-party review of the research protocol. A **template of the terms of reference for the IRB** is available for teams.

The ethical clearance process involves submission of the research protocol, questionnaires, informed consent and other study materials to the IRB, review by the IRB, and revision of the protocol and materials based on any recommendations from the committee. The process normally takes 2-3 months, though it can be heavily dependent on how often the IRB meets. The IE team should plan to have all required clearances prior to piloting the survey with the field team.

## Defining the Data Entry Strategy

Defining a data entry strategy is a crucial step in determining which survey firm to hire. Two common data entry strategies are Computer Assisted Field Edits (CAFE) and centralized data entry. Under CAFE, each field team includes a data entry operator with a laptop who is responsible for digitizing the data in the field. In centralized data entry, the paper questionnaires are transported to a central data entry location where a team of data entry operators digitizes the data from all field teams.

For HRITF funded evaluations, CAFE is the recommended data entry strategy, because evidence from implementation of the LSMS suggests that it optimizes the overall quality of the data collected. However, this data entry strategy has implications for the budget and timing of the survey. Therefore, the IE team should decide on whether they want to use the CAFE option *before* hiring the survey firm.

Implementing CAFE requires hardware, software, an effective organizational structure and realistic planning of survey field work. In particular, it requires (i) a committed and dedicated core staff team, and (ii) implementation of the team approach to field work. The following general considerations are important when comparing CAFE with centralized data entry:

**What resources are necessary for CAFE?** To implement CAFE effectively it is necessary to have one data entry operator (DEO) armed with a laptop on each field team. The survey firm will likely need to procure the laptops, which will affect both the budget and timeline. The survey firm may also consider not using a data entry operator, and training the enumerators to enter the data themselves. Team supervisors will also be responsible for entering the data, analyzing the error reports and deciding on corrective revisits. Training the team supervisors on CAFE responsibilities is difficult at first, and therefore central supervision of CAFE teams needs to be far more active, technical, capable and committed than usual central supervision of field teams. Central supervisors must know better than anybody else how to deal with the daily reports generated by CAFE.

**How much time is needed for CAFE training**? Evidence suggests that a survey firm should budget 3-5 days for training data entry operators, which will be subject to the local capacity. The CAFE training should focus on preparing the data entry operators and team supervisors to enter the data, analyze the error reports and make corrective revisits as necessary.

**What are the differences in data quality between CAFE and centralized data entry?** The essential benefit of CAFE is that outliers and inconsistencies are dealt with in the field, by direct confrontation with the household's reality, rather than through office guesswork, as is sometimes required in centralized data entry. Typically, the long and frustrating process of “data cleaning” from a central level becomes unavoidable, and threatens the policy-making relevance of the data. This is because “data cleaning” becomes a process of ensuring data is internally consistent, but does not necessarily represent the reality in the field. CAFE not only allows for immediate identification of inconsistencies in the field, but also provides immediate feedback on the performance of the field staff, allowing early detection of inadequate behaviors.

**How much time is saved by using CAFE vs. centralized data entry?** With centralized data entry the survey firm will typically need to spend at least 2-3 months entering and cleaning the datasets after the data has been collected. With CAFE, the delivery of finalized databases is immediate following the completion of data collection -- all the data is available at the same moment the last cluster is surveyed and its file delivered. However, CAFE may lead to a slight extension of the time required in the field, especially if many re-visits are necessary. Yet, the benefits of superior data quality outweigh the costs of longer field work.

## Hiring a Survey Firm

The IE team will require a survey firm to conduct all the major data collection activities, including: (i) health facility survey, (ii) household listing, (iii) household survey, (iv) biometric measurement, and (v) community survey. Depending on the local capacity, the IE team may find that they need to hire different survey firms to complete different activities. For example, the survey firm most qualified to lead a smaller health facility survey may not have the capacity or experience to lead a large-scale household survey. For this reason, the IE team should consider competitively bidding these two core activities separately in order to identify the most competent firm for each survey. However, there are also advantages of bidding for one unique firm to complete all the surveys. Table 13 highlights some of the pros and cons of both approaches when conducting both household and health facility surveys.

### Table 13: Pros and Cons of Choosing One vs. Several Survey Firms

| Different firms for different surveys | One firm for all surveys |
|---|---|
| + Specific expertise and sharper skills in conducting the given survey (e.g. experience in large scale surveys for household survey; experience with medical context for health facility survey)<br>+ Less training required as survey firm already specialized in either large scale survey or survey in medical context<br>- More difficult coordination of both surveys and transaction costs<br>- Higher procurement costs<br>- May imply more supervision required | + Better coordination of activities and synergies between household and health facility surveys<br>+ Training and capacity building done by the IE team for the first type of survey serving the quality of the second type of the survey<br>- However training and capacity more intensive at the beginning<br>- High field staffing required to conduct surveys simultaneously<br>+ Lower transportation costs, and transaction costs in general<br>+ Matching of households and health facilities easier at the sampling stage |

2012/02/05

**Country Spotlight: Potential and Challenges in hiring a unique survey firm**
**Benin Results-Based Financing Project**

The team launched a competitive bidding for survey firms to conduct both household and health facility surveys. The rationale behind having one firm for both surveys was to save on transaction and transportation costs. The winning firm was a local firm, chosen on the grounds of lower costs, but also with the concern of building local capacity. The firm benefited from the presence of skilled staff but it lacked experience in large scale surveys. The World Bank hub team and their partner quality assurance firm Sistemas supported and trained the members of the survey firm during the pilot phase to limit the risks of poor data quality. (…) The team emphasized the supervision of household data collection with the presence of external controllers. As a result, the household data collected was of good quality.

However, the health facility survey was not conducted with such supervision and control. This got combined with very technical content: enumerators had to administer fifteen health facility instruments. The measurement of absenteeism also implied two visits, one announced and one unannounced, which added to the complexity – and the cost – of the survey. Finally, the firm did not have a lot of expertise in health facility surveys in the first place. For those reasons, the team is concerned that the quality of the health facility data may not be as good as that of the household data, and is considering recollecting part of the data at follow-up if necessary.

Full story available: see Country Spotlights section of the Toolkit.

2012/02/22

**Country Spotlight: Making Choices when Hiring a Survey Firm**
**Zambia Health Results-Based Financing intervention**

[In Zambia], a central question during the survey firm procurement stage was the availability of firms with experience in managing large-scale data collection and a good track record of delivering high-quality data (…). Following a competitive tender through the Bank's procurement system, the team selected a local survey firm with a good track record. However, because implementing multi-site facility and household surveys simultaneously requires a large field force and substantive excess capacity, the team chose to split the survey implementation into a facility survey, (…) and a household survey. The (…) Zambia team engaged in capacity building of the local survey firm.

While taking a local capacity building approach is expected to lead to positive externalities over time and across programs, it is associated with some risks. To minimize the potential risks, there were built-in quality and fiduciary controls. For example, the Terms of Reference for the firm that implemented the household survey included conditionality on milestones for fund release and data quality for engagement in any follow-up survey. These incentives have been critical during managing the evaluation.

Since the facility and household surveys are implemented by two different entities, this provided an opportunity for cross-survey-firm support during the implementation of the HRBF surveys (…).

Full story available: see Country Spotlights section of the Toolkit.

As mentioned earlier this module, the competitive selection process for a survey firm(s) is lengthy: it may take 2-3 months to prepare and evaluate proposals, and another 2-3 months to negotiate and sign a contract.

The following is a summary of key points that we believe are important when hiring a survey firm:

- Selecting the survey firm:
  - Will you hold a competitive bidding for the data collection? If so, make sure your timeline includes time for the call for proposals, review of proposals, and contracting the firm.
  - Based on the study requirements, will you competitively bid different data collection activities separately (i.e. health facility, household, qualitative)?
- CAFE vs. centralized data entry:
  - Will you use field data entry? If yes, then make sure this is included in the TOR as it affects the required qualifications, equipment, training, and the budget.
- Household (or health facility) listing:
  - Does such a listing already exist? If not, then make sure this is included in the TOR as it highly affects the budget and timing.
- Ethical Review Board:
  - What are the local requirements for conducting research and what will be required of the survey firm?

    - Is there a fee for the ethics committee? Who will pay for it? If it is the survey firm, then this should be included in their TOR.
    - Who is required to present the study to the committee (IE team and/or hired survey firm)? If it is the survey firm, then this should be included in their TOR.
- Supervision:
    - What are the reporting and supervision mechanisms planned for the survey firm to report on issues?
- Timeframe:
    - What are the time constraints for data collection?
    - How much time is required per field team to complete data collection in a specific cluster?

## Survey Firm Staffing

We recommend that the health facility and household level data collection should be viewed as two distinct activities, regardless of whether one or more survey firms are involved. For each component, the survey firm(s) should be prepared to build a team that is composed of survey managers, enumerators and data entry operators.

***Survey firm management***. It is important to note that multiple managers will be necessary to supervise different aspects of the work, but that all of them will need to participate in all phases of the survey, on a full-time basis. The key managerial positions are:

- Project Manager: plans, supervises and manages the entire survey with the assistance of the field and data managers. The Project Manager must be based in-country for the entire duration of the survey.
- Field Manager: plans, supervises and manages the field work.
- Data Manager: plans, supervises and manages data entry, error checking, processing and consolidation of data.

In many cases, survey firms will require substantial technical support from the IE team and the Data Quality Expert (Cf. Module 2). It is important to ensure that from the beginning the survey firm is aware that the Data Quality Expert may need to be involved in all aspects of data collection or entry if it is deemed necessary by the IE team.

Regular reporting systems between the survey firm and the IE team, and among survey teams, will also need to be set up.

2012/02/22

**Country Spotlight: Supervising Survey Firms and Field Work**
**Zambia Health Results-Based Financing intervention**

For example [in Zambia], the TOR with the survey firm stipulated not only survey implementation mid-time and end-line reporting but also reports on challenges when they arise. The need for such just-in-time reporting became clear during the survey work related to the malaria program evaluation when field teams experienced hostility from communities because of perceptions and beliefs related to the blood testing within the biomedical component. In addition to the reporting requirements for the firm defined in the TOR, an IE coordinator receives regular updates from the field, about every week, or more if needed.

[Among survey teams,] the HRBF surveys questionnaires are checked on a daily basis, and interviewers debrief every evening to reduce discrepancies/missing data, and thus return visits, which also means lower travel costs for the survey.

Full story available: see Country Spotlights section of the Toolkit.

***Field teams.*** The recommended structure for each field team includes:

*1 field team supervisor*: The field team supervisor should be responsible for assisting enumerators in solving any problems encountered during field work. Consequently, enumerators should channel all questions, comments, observations, and complaints through the field team supervisor. This will allow for all issues to be handled in a systematic way, allow for timely responses, and help the centrally-based field manager to respond to field supervisors, instead of the each team member independently. The field team supervisor should also be responsible for assessing the work of the enumerators on their team: he/she will need to randomly observe the interviews with household members and revisit some households selected at random. As questionnaires are completed, the field supervisor should review each questionnaire for errors, and ensure that enumerators return to households for whom there were errors or incomplete data. We believe that field team supervisors play a significant role in ensuring high quality data collection, and their performance should be assessed by project managers based on the quality of the data.

*2-4 enumerators*: Depending on the country context and nature of the data, the survey firm may need to have a mix of male and female enumerators in each team. The correct composition of the field team should be determined based on data collection requirements, and the gender sensitivity to specific questions in the questionnaires.

*1 anthropometrist/biomarker collector:* The field teams may also include an individual who is qualified and trained in collecting anthropometric and/or biological data collection. In some contexts and countries, this individual may be required to have a nursing or other medical background.

*1 data entry operator*: **If the CAFE system for data entry is used, a** data entry operator will generally be required. We recommend that this individual should be responsible for entering data from questionnaires as they come out of the field, and performing immediate checks for inconsistent or implausible data. The field team supervisor should review these checks, and determine if a second visit to the household or facility is necessary.

Budget and time constraints will be important factors in determining the number of field teams. With fewer field teams, survey managers will have an easier task supervising them and ensuring that data quality measures are properly implemented; on the other hand, there needs to be a sufficient number of field teams in order to complete the survey within the allocated amount of time. As previously mentioned, it is vital that baseline data collection be completed before the intervention begins in the treatment groups and that endline data collection be completed before the intervention is rolled in the comparison group.

### Deliverable and Payment Schedule.

In order to minimize the cash flow risks during the implementation of surveys, the survey firm's deliverables should be clearly linked to required activities, and follow a schedule which allows for release of funds at critical points during preparation and implementation of evaluation activities, while minimizing the risk that the firm will abandon the work without completing it. Table 14 presents an example of a deliverable and payment schedule that should be specified in the **terms of reference of the survey firm(s)**:

### Table 14: Survey Firm Deliverable and Payment Schedule

| Deliverable | Date | Payment |
|---|---|---|
| Signature of Contract | Month 0 | 10% |
| Deliverable 1:<br>1.1 IE Gantt Chart with all proposed activities, deliverables and timeframe for each<br>1.2 Adapted research protocol and informed consent forms<br>1.3 Evidence of ethical clearance, insurance and permits needed to implement the survey. | Month 2 | 20% |
| Deliverable 2.1:<br>Adapted questionnaires in English<br>Initial translation of questionnaires into local language | Month 2 | |
| Deliverable 2.2:<br>Pre-testing report including timing of modules, comments from enumerators and supervisors and necessary changes to the questionnaire<br>Final local language questionnaire<br>Final corresponding English questionnaires. | Month 3 | 20% |
| Deliverable 3: Written Sampling Plan approved by the Evaluation Team. | Month 3 | |
| Deliverable 4: Written Biomarker Data Collection Protocol approved by the IE Team. | Month 3 | |
| Deliverable 5: Written Field Work Plan approved by the IE Team. | Month 3 | |
| Deliverable 6:<br>6.1 Written data entry protocol for data entry agents detailing program<br>6.2 Final data entry program adapted for the local questionnaires<br>6.3 Dataset dictionary with all variables labeled and defined | Month 3 | |
| Deliverable 7: Roster of recruited personnel with their corresponding qualifications. | Month 4 | 20% |
| Deliverable 8: Procurement and Training<br>8.1 Procured materials (anthropometrics, GPS, biomarker data collection)<br>8.2 Training materials and field manuals<br>8.3 Report with the results of the interviewers' evaluations | Month 4 | |
| Deliverable 9: Final Pilot Report and Data successfully transferred to the Evaluation Team. | Month 5 | |
| Deliverable 10: Project Manager's Final Written Baseline Data Collection Report | Month 9 | 30% |
| Deliverable 11: Final Databases and Final Data Delivery Report | Month 10 | |
| Deliverable 12: Timely delivery of Project Manager's bi-weekly Progress Reports | 10 months | |

## Budget

A number of factors will influence the **survey firm(s) budget**:

- Local wage and benefit levels
- Reliance on international staff (wages and travel)
- Sample size and geographic distribution
- Anticipated duration of interviews (health facility, household and/or community)
- Composition of field teams
- Number of field teams
- Duration of field work

- Required procurement of materials (anthropometrics, biomarker tests)

In order for a survey firm to provide an accurate estimate of the survey budget in its financial proposal, the terms of reference (TOR) used in the Request for Proposals (RFP) should contain all relevant information to identify the particular requirements of the data collection and inform the factors detailed above. In some cases, time is a major factor as the data must be collected prior to implementation or scale-up of the RBF intervention(s). In this case, the TOR should specify the time constraints so that the survey firm will be able to estimate the composition and number of field teams required to complete data collection in the required timeframe. In addition, the TOR should specify all materials required for the data collection and be explicit about which entity will be responsible for procuring them. In many countries, the recommended anthropometric and biomarker testing materials must be shipped from outside the country, and this typically results in a greater time and costs to the survey firm. Sometimes though, these materials can be obtained from donors or national programs (e.g. national anti-malaria program) for free. Regardless of their origin and price, the procurement costs and timing must be included in the TOR.

Experience demonstrates that the majority of data collection budgets are underestimated. We advise the following:

- TOR should be as detailed as possible and provide clear description of all required data collection activities
- Provide a Survey Firm Budget Template to avoid omission of key budget items
- A data collection expert should review the technical and financial proposals to ensure no major gaps exist.
- Allow for a 10% miscellaneous budget item to avoid budget overrun.
- Carefully check the proposed methodology, timing and budget for the household listing survey, if required. Unfortunately, survey firms usually under-budget this type of activity.

## Understanding and Adapting the Survey Instruments

The **RBF Survey Instruments** and corresponding **Data Entry Program** were developed to capture an extensive amount of data at the health facility, health worker, household and individual levels. This section provides an overview of the basic format of these questionnaires, recommendations for adaptation and translation, as well as detailed descriptions of the survey instruments in terms of sample, respondents, timing and content.

*Please note: Every impact evaluation in the RBF network is unique in its setting, research questions and context. Therefore, it is not possible to put together a unique or uniform set of data collection instruments. The RBF instruments are provided as resources to the teams – by using a common resource, we hope that teams will be able to add to the collective knowledge about how to measure RBF impact and benefit from each other's experience.*

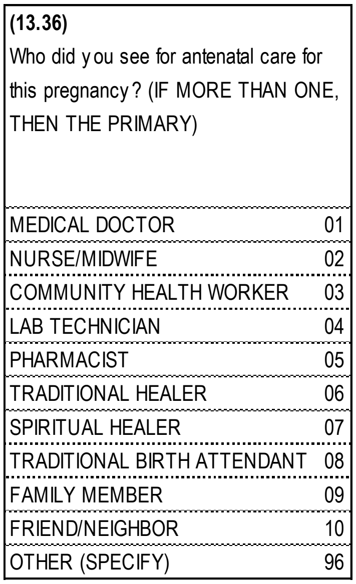
## Universal Formatting Guidelines

The format of the questionnaires is an important determinant of the quality of the data; therefore, we recommend that the questionnaires follow universal guidelines for format and design. We strongly advise that any edits to the questionnaires should follow these guidelines, in order to ensure that they do not lead to inconsistencies between modules, country evaluations, or teams.

**CAPITALIZED vs. lower case Font.** The following instructions should be included in the Enumerator Manual and made available to the survey firm and enumerators.

- Lower case Font: The lower case font is intended to be read aloud, and is used for the formulation of interview questions and the listing of all response options, when the question requires the enumerator to read the possible responses aloud.
- CAPITALIZED font is intended to not be read aloud by the enumerator, and is used for instructions to the enumerator, when a question requires instructions to the enumerator, and for the listing of response options, when the responses are not intended to be read out loud, and only recorded if mentioned by the respondent.

**Response Codes. We recommend that** all response codes be written in two-digit format. The following is an example from a previous questionnaire



**(13.36)**
Who did you see for antenatal care for this pregnancy? (IF MORE THAN ONE, THEN THE PRIMARY)

| | |
|---|---|
| MEDICAL DOCTOR | 01 |
| NURSE/MIDWIFE | 02 |
| COMMUNITY HEALTH WORKER | 03 |
| LAB TECHNICIAN | 04 |
| PHARMACIST | 05 |
| TRADITIONAL HEALER | 06 |
| SPIRITUAL HEALER | 07 |
| TRADITIONAL BIRTH ATTENDANT | 08 |
| FAMILY MEMBER | 09 |
| FRIEND/NEIGHBOR | 10 |
| OTHER (SPECIFY) | 96 |

While we recommend that all questions are coded with a two digit code, one exception to that rule is the Yes/No response, where Yes and No are usually coded as 1 and 2.

**“Don’t know” and “Refusal” responses.** Most questions in the questionnaire do not have an option for a “don’t know” answer. This is because previous experience with household surveys has shown offering this option makes an interviewer less likely to persist in obtaining an answer from the respondent. Of course, there may be cases where the respondent genuinely does not know the answer to the question; in this case the interviewer should spell out “Don’t know” in the paper questionnaire. Similarly, when respondents refuse to answer a particular question, we recommend that interviewers write “Refusal” in the paper questionnaire. Data entry operators will code these special responses into pre-defined codes at the time of data entry. The RBF data entry program uses -7 (minus 7) for "Don't know" and -8 (minus 8) for "Refusal".

We recommend that the IE team identify any questions that have an unusually high “Don’t know” or “Refusal” responses during the pilot or pretest, and either (i) revise the wording of the question or (ii) remove the question. High rates of "Don't know" or "Refusal" responses give an indirect measure of the quality of the interviews and, ultimately, of the quality of the survey firm’s data collection.

**“Other” response.** We strongly recommend that the code for “Other” response be “96” in order to maintain consistency across all questions, modules, and previous surveys. In addition, all "other” responses should be recorded by interviewers exactly as they are declared by respondents, using the respondent’s phrasing and diction. The data entry modules provide a specific place for recording the text of "other responses", and the recorded responses should be reviewed by quality assurance staff in case the recorded response is a pre-coded response, in which case it should be converted into its respective code.

### Adaptation of RBF Survey Instruments

2012/02/22

**Country Spotlight: Adapting and testing questionnaires**
**Zambia Health Results-Based Financing intervention**

(…) As every project is unique the survey-related products required content customization. The HRBF team relied on instruments developed for a malaria project in Zambia for the household survey, and on instruments developed by the Zambia IE team with support from the Bank’s evaluation team at the hub for the health facilities. Field testing, which lasted for about six months, allowed identifying significant adjustments to be made. In particular, the questionnaire, initially administered in four hours, was reduced to 1.5 hours by removing redundant or difficult-to-administer sections from the socio-economic and health books, including biometric and physical activity questions. Despite the significant time spent on adaptation, the costs were lower compared to developing a new product.

Full story available: see Country Spotlights section of the Toolkit.

Obvious country-specific content appears in red in the RBF survey instruments. IE teams have the option of adapting and adjusting the RBF Survey Instruments to the local context in further depth, but should be cognizant that this may lead to inconsistencies inside their questionnaire and with data collected in other countries.

**Including relevant questions to respond to the research questions.** When adapting the questionnaires, one essential element to keep in mind is that ultimately, the instruments aim at measuring the impact of the intervention on the outputs and outcomes of interest. Therefore, it is essential to make sure that:

- The most appropriate instruments are used to measure indicators of interest.
- Within each instrument, the questions and answers that will allow calculating the indicators of interest are indeed all included, administered to the appropriate respondents and regard the appropriate timeframe (see RBF Indicators list in Module 3).

In the rest of this section, we provide recommendations for three types of changes that country teams may need to make to the questionnaire: (i) adding country-specific questions, (ii) adding and/or dropping responses to existing questions and (iii) dropping questions that are not relevant. These changes should be coded in specific ways to indicate that they are local additions to the questionnaire.

**Adding country-specific questions.** When adapting survey instruments to the country-specific context, questions can be coded with the country-specific 2-letter nomenclature, e.g.:

AF – Afghanistan
BJ - Benin
ZR – Democratic Republic of Congo
GH- Ghana
KG- Kyrgyzstan
RW – Rwanda
ZM – Zambia
ZW- Zimbabwe

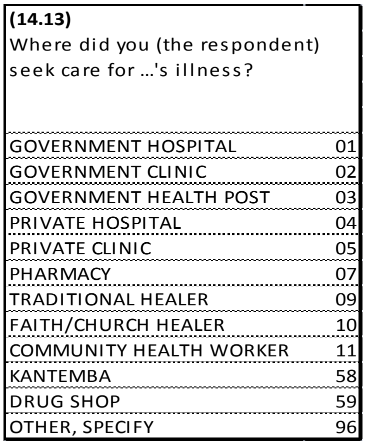
For a full list of countries, please refer to http://cds.worldbank.org/Pages/CntryGroup.aspx.

For example, if the team in Rwanda needs to add three questions after question 10 in the questionnaire, it can code the questions as RW10A, RW10B and RW10C. This way, the subsequent question numbering will not be affected and the data entry program and STATA do files can still be used.

The questionnaires include automatic skip patterns that let the interviewer skip questions that are not relevant based on previously obtained answers. When country teams add questions to the questionnaire, we suggest that they pay very close attention to the skip patterns throughout the section where the new questions are added to ensure that skip patterns are not disrupted by the additions and that the questions will be applied to all intended respondents. In general, it may be simpler for teams to add new content in their own section(s) at the end of the module/questionnaire, to avoid the potential problems in altering existing sections.

**Adding country-specific response codes**. When adjustments are made to responses to questions, new response options should be given codes clearly that are clearly different from the standard existing responses. Codes for responses that are present in the global version should remain unchanged, even if some responses are dropped. This may mean that the codes for responses are not sequential, but codes do not need to have any sequence. Typically, new codes that are used for country-specific responses have been defined using codes over 50. Codes that reflect different naming conventions in different countries can be over-written. For example, if a medical doctor is normally coded 01 but medical doctors are called medical officers, then the team would use the code 01 for medical officers. By using similar coding guidelines, teams can help ensure that data from different countries will be easily comparable.

For example: Say we want to add two possible responses that are specific to Zambia in question 14.13 "Where did you seek care for [YOUR CHILD]'s illness?" To do so, we would add the Zambia specific responses with codes 58 and 59": Kantemba – 58 and Drug Shop – 59.



| **(14.13)**<br>Where did you (the respondent) seek care for ...'s illness? | |
|---|---|
| GOVERNMENT HOSPITAL | 01 |
| GOVERNMENT CLINIC | 02 |
| GOVERNMENT HEALTH POST | 03 |
| PRIVATE HOSPITAL | 04 |
| PRIVATE CLINIC | 05 |
| PHARMACY | 07 |
| TRADITIONAL HEALER | 09 |
| FAITH/CHURCH HEALER | 10 |
| COMMUNITY HEALTH WORKER | 11 |
| KANTEMBA | 58 |
| DRUG SHOP | 59 |
| OTHER, SPECIFY | 96 |

**Dropping questions that are not relevant.** A number of questions in the standard questionnaires are marked as optional. In addition, other questions may not be relevant to the country context, or not be aligned with the country's research questions. We recommend that the Principal Investigator determine whether the questions are relevant or appropriate, and if not, drop them from the questionnaire. When a question is dropped we recommend that the numbering of the subsequent question number be maintained as in the original questionnaire when possible, so that they remain consistent across countries.

2012/04/03

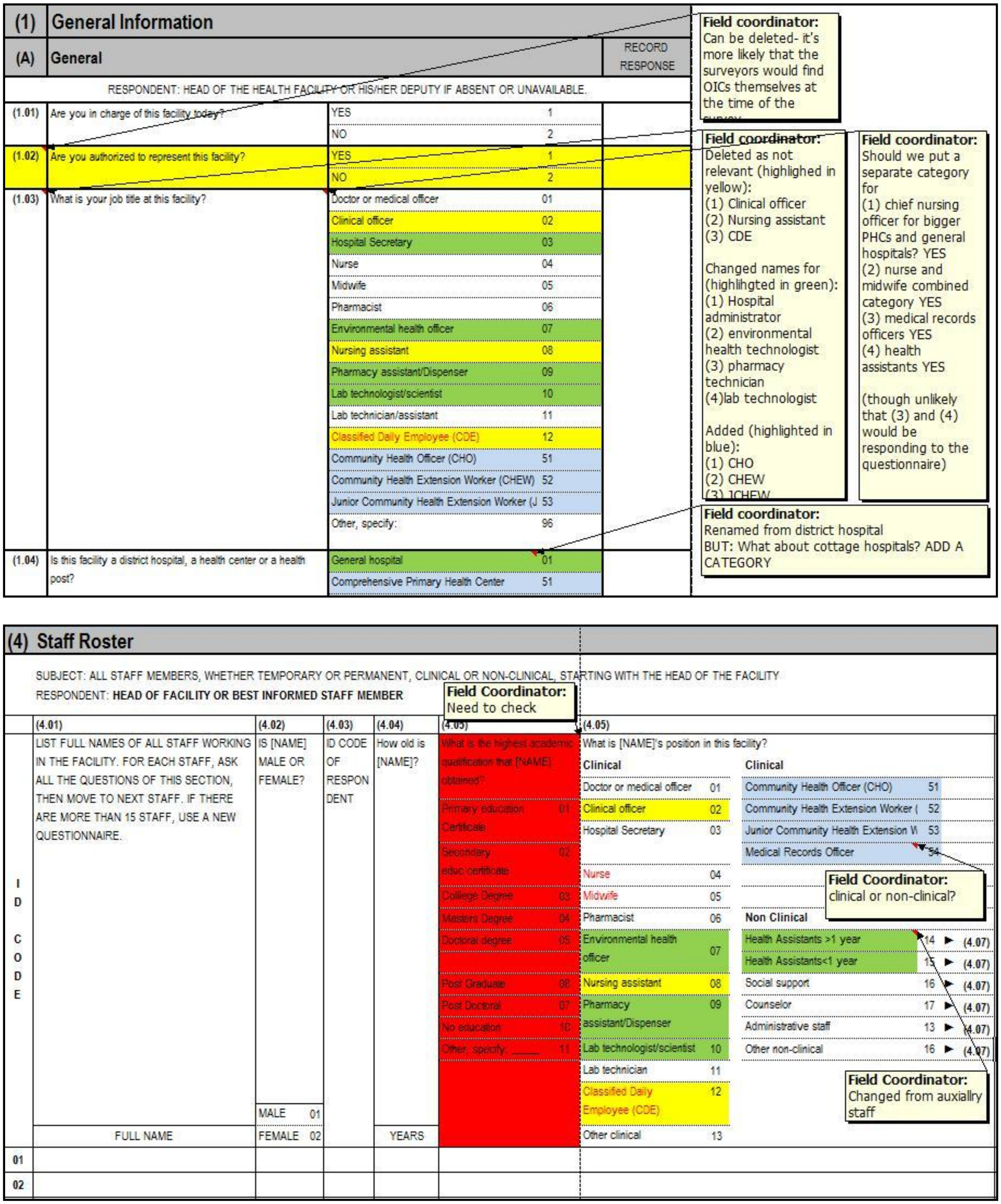
## Country Spotlight: Adapting survey questionnaires
## Nigeria State Health Program Investment Credit

In preparation for their baseline survey, the Nigeria team used the HRITF questionnaires and adapted them to their local context and research questions. This requires collaboration and back and forth between team members, especially between Principal Investigator, IE coordinator and research assistant. Therefore, it is useful to flag changes. The Nigeria team used the following to flag changes to the questionnaires during the adaptation process:

- Cells highlighted in yellow are to be deleted
- Cells highlighted in green were modified from the original HRITF questionnaire
- Cells highlighted in blue were added to the original HRITF questionnaire
- Cells highlighted in red need further checking
- The team used comments on those changes for further clarification and to point out remaining issues.

Below are screenshots of the questionnaires in progress to illustrate the adaptation process.

## Figure 8: Screenshots of Toolkit Questionnaires during Country-specific adaptation



| | | | | |
|---|---|---|---|---|
| (1) | General Information | | | |
| (A) | General | | | RECORD RESPONSE |
| | RESPONDENT: HEAD OF THE HEALTH FACILITY OR HIS/HER DEPUTY IF ABSENT OR UNAVAILABLE. | | | |
| (1.01) | Are you in charge of this facility today? | YES | 1 | |
| | | NO | 2 | |
| (1.02) | Are you authorized to represent this facility? | YES | 1 | |
| | | NO | 2 | |
| (1.03) | What is your job title at this facility? | Doctor or medical officer | 01 | |
| | | Clinical officer | 02 | |
| | | Hospital Secretary | 03 | |
| | | Nurse | 04 | |
| | | Midwife | 05 | |
| | | Pharmacist | 06 | |
| | | Environmental health officer | 07 | |
| | | Nursing assistant | 08 | |
| | | Pharmacy assistant/Dispenser | 09 | |
| | | Lab technologist/scientist | 10 | |
| | | Lab technician/assistant | 11 | |
| | | Classified Daily Employee (CDE) | 12 | |
| | | Community Health Officer (CHO) | 51 | |
| | | Community Health Extension Worker (CHEW) | 52 | |
| | | Junior Community Health Extension Worker (J | 53 | |
| | | Other, specify: | 96 | |
| (1.04) | Is this facility a district hospital, a health center or a health post? | General hospital | 01 | |
| | | Comprehensive Primary Health Center | 51 | |

**Field coordinator:**
Can be deleted- it's more likely that the surveyors would find OICs themselves at the time of the survey

**Field coordinator:**
Deleted as not relevant (highlighed in yellow):
(1) Clinical officer
(2) Nursing assistant
(3) CDE

Changed names for (highlihgted in green):
(1) Hospital administrator
(2) environmental health technologist
(3) pharmacy technician
(4)lab technologist

Added (highlighted in blue):
(1) CHO
(2) CHEW
(3) JCHEW

**Field coordinator:**
Should we put a separate category for
(1) chief nursing officer for bigger PHCs and general hospitals? YES
(2) nurse and midwife combined category YES
(3) medical records officers YES
(4) health assistants YES

(though unlikely that (3) and (4) would be responding to the questionnaire)

**Field coordinator:**
Renamed from district hospital
BUT: What about cottage hospitals? ADD A CATEGORY

### (4) Staff Roster

SUBJECT: ALL STAFF MEMBERS, WHETHER TEMPORARY OR PERMANENT, CLINICAL OR NON-CLINICAL, STARTING WITH THE HEAD OF THE FACILITY
RESPONDENT: **HEAD OF FACILITY OR BEST INFORMED STAFF MEMBER**

**Field Coordinator:**
Need to check

| | (4.01) | (4.02) | (4.03) | (4.04) | (4.05) | | (4.05) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| I D CODE | LIST FULL NAMES OF ALL STAFF WORKING IN THE FACILITY. FOR EACH STAFF, ASK ALL THE QUESTIONS OF THIS SECTION, THEN MOVE TO NEXT STAFF. IF THERE ARE MORE THAN 15 STAFF, USE A NEW QUESTIONNAIRE. | IS [NAME] MALE OR FEMALE? | ID CODE OF RESPONDENT | How old is [NAME]? | What is the highest academic qualification that [NAME] obtained? | | What is [NAME]'s position in this facility? | | | | |
| | | | | | | | Clinical | | Clinical | | |
| | | | | | | | Doctor or medical officer | 01 | Community Health Officer (CHO) | 51 | |
| | | | | | Primary education Certificate | 01 | Clinical officer | 02 | Community Health Extension Worker ( | 52 | |
| | | | | | Secondary educ certificate | 02 | Hospital Secretary | 03 | Junior Community Health Extension W | 53 | |
| | | | | | | | | | Medical Records Officer | 54 | |
| | | | | | | | Nurse | 04 | | | |
| | | | | | College Degree | 03 | Midwife | 05 | | | |
| | | | | | Masters Degree | 04 | Pharmacist | 06 | Non Clinical | | |
| | | | | | Doctoral degree | 05 | Environmental health officer | 07 | Health Assistants >1 year | 14 | ► (4.07) |
| | | | | | | | | | Health Assistants<1 year | 15 | ► (4.07) |
| | | | | | Post Graduate | 06 | Nursing assistant | 08 | Social support | 16 | ► (4.07) |
| | | | | | Post Doctoral | 07 | Pharmacy assistant/Dispenser | 09 | Counselor | 17 | ► (4.07) |
| | | | | | No education | 10 | | | Administrative staff | 13 | ► (4.07) |
| | | | | | Other, specify: _____ | 11 | Lab technologist/scientist | 10 | Other non-clinical | 16 | ► (4.07) |
| | | | | | | | Lab technician | 11 | | | |
| | | MALE 01 | | | | | Classified Daily Employee (CDE) | 12 | | | |
| | FULL NAME | FEMALE 02 | | YEARS | | | Other clinical | 13 | | | |
| 01 | | | | | | | | | | | |
| 02 | | | | | | | | | | | |

**Field Coordinator:**
clinical or non-clinical?

**Field Coordinator:**
Changed from auxiallry staff

### Translation and Back Translation

Once the English instrument(s) is adapted to the local context, the survey firm should translate the instrument(s) into any necessary local language(s). To verify that the translation was done accurately, each local language version can be back-translated by an independent translation team, who was not involved in the initial translation. This activity should be included in the survey firm's TOR and budget, as well as the general timeline. We estimate that translation and back-translation each take 2-3 weeks.

As **Translator Excel macro** is provided in this toolkit to facilitate the translation of formatted questionnaires.

## Survey Instruments

### Health Facility Questionnaires

The health facility questionnaires were designed to provide primary data on service delivery, facility structures, process quality, human resources and infrastructure. A health facility survey involves visiting and collecting data for all health facilities identified in the sampling plan. Special methods, such as record review, observing client-provider interaction and using standardized patients can add considerable value to the facility assessment. Additionally data collected from record reviews and staffing inventories can be used to validate routine administrative statistics on the volume of services delivered and on the availability and geographical distribution of human resources, such as the data available in the HMIS system. However, they can also increase the costs and complexity of the data collection.

#### *Facility Assessment*

**Sample**. The facility assessment should be applied to all health facilities in the sample as defined by the sampling plan.

**Respondent**. The desired respondent for the facility assessment is the health facility manager or administrator. Although some sections may require follow up with the heads of accounting, pharmacy, and laboratory, the enumerators should initiate the facility assessment with the health facility manager. The health facility manager can then determine whether other focal points are required to complete specific sections.

**Timing.** We estimate that the health facility assessment will take approximately **1 day**. The interview should be carefully timed around service delivery hours to minimize disruption to patient care, as well as around the availability of the manager. It is highly recommended that field teams call the health facility in advance to ensure the manager will be present on the day of the interview.

**Content**. There are fifteen sections in the health facility survey instrument:

1. *General Information and Universal Precautions*: This section gives an overview of the facility infrastructure, service hours, referral services and financials. In the context of RBF, this data is important for understanding how the facilities' management responds to the incentive structure (i.e. expanding patient rooms, extending service hours to increase utilization, reallocate budget to improve facility quality, etc.)
2. *Administration and Management*: This section collects data on the facility management in terms of its relationship with the community, developing a business plan, supervision, internal and external assessment, and budget planning. This section also gives a sense of the facility's autonomy and authority to procure drugs and make management decisions.
3. *Human Resources*: This section gives an overview of the facility's staff, including recent hiring, turnover and vacant positions as well as staff training and collaboration with community health workers. In addition, the section includes a staff roster with key staff characteristics.
4. *Roster:* This section collects basic data on all staff members of the facility, such as level of education, qualifications, and workload.
5. *Laboratory Services*: The section collects data on the laboratory services available at the facility.
6. *Services*: This section collects data on the vaccination, prenatal, delivery, post-natal, tuberculosis and malaria services offered and the respective protocols followed at the health facility.
7. *General HMIS*: This section collects basic HMIS data on the composition of the catchment area population and general composition of patients.
8. *Health Services Utilization*: This section collects data on the utilization of health services based on HMIS data.
9. *User fees*: The sections collects data on patient fees, the transparency on fee rates, exceptions, authority on who sets fees, as well as how the income from fees is used within the facility. One of the potential impacts of RBF is a reduction of out-of-pocket expenditures by patients in order to increase utilization rates. This data will complement population-level data to determine the impact on user fees.
10. *Leadership*: This section collects information on the type of leadership exerted by the health facility manager.
11. *Authority*: This section collects information on the degree of autonomy of decision at the health facility level.
12. *Direct Observation*: In this section, enumerators will fill out what they observe regarding the general state of the facility and postings of user fees and national protocols.
13. *Equipment*: This section collects data on the facility's available and functional equipment, particularly the equipment required for providing key maternal and child health services. This data is used for constructing the structural quality of the facility.
14. *Drug Supply*: The section collects data on the stock of key drugs in the last 30 days for general uses, malaria, family planning, tuberculosis, obstetric, and vaccination services. This data will also be used for constructing the facility's structural quality.
15. *Catchment Area*: This section collects additional nominative data on the villages included in the catchment area.

### *Health Worker Interview*

**Sample**. The sample will depend on the objectives of the IE, and in particular the type of service that is of interest (eg. Prenatal care, child care, adult care, etc). For example, if an objective of the evaluation is to measure the impact on the quality of prenatal care, then the survey will need to include at least one health workers who provide this specific service.

**Respondent**. If there is only one provider for the service(s) of interest present on the day of the interview, then that provider should be selected for the interview. However, if there is more than one provider for the service(s) of interest present on the day of the interview, then the field team will need to randomly select one of the providers in the field of interest. We recommend that field teams identify whether or not certain services are offered on desired interview day(s) in order to ensure that there is at least one provider in the field of interest will be present on the chosen day.

**Timing.** We estimate that the health worker interview will last 45-60 minutes. The interview should be timed so as to minimize disruption to patient care.

**Content**. There are thirteen sections in the health worker interview:

1. *General Information*: This section collects basic demographic information and data on the health worker's position, experience and responsibilities.
2. *Training*: This section collects data on the health worker's training in key health service areas in the last year or more, as well as training needs.
3. *Hours worked*: This section details hours and days worked, as well as reasons for absence.
4. *Salary*: This section details the monthly salary, payment regularity, as well as potential employment and salary options outside of health care.
5. *Other Compensation*: This section details the value of other forms of compensation, including travel, housing, remote location and bonus payments.
6. *Supervision*: This section collects information on the supervision activities at the health facility, the health worker's feedback from the supervisor and the supervisor's contributions to the work environment.
7. *Secondary Job*: This section collects information on any secondary work the health worker is involved in, as well as the supplemental income generated by this work.
8. *Well Being*: This section collects information on the general well-being of the health worker.
9. *Satisfaction (optional):* This section collects data on the health worker's satisfaction related to various elements of her working conditions.
10. *Personal drive (optional)*: This section collects data on the health worker's motivation related to various elements of her working conditions.
11. *Innovation (optional)*: This section collects information on the ability of the worker to respond to changes in the facility and/or the community in the vicinity of the facility.
12. *Vignettes*: Vignettes measures knowledge through case scenarios. Vignettes describe a situation where a patient comes in with specific symptoms or conditions. The provider is asked what (s)he

would do to take care of the patient. The focus is on assessing the medical knowledge of the health worker.

13. *Vignette* for prenatal care protocol.

For more information on how to measure quality of care through vignettes, please refer to the section below on Measuring Quality of Care.

*Please note: The expected duration of the health worker interview will likely run over 45 minutes if all thirteen sections are included in their entirety. Teams should feel free to adjust the content of the data collection instruments and swap modules in/out so that their final questionnaire reflects the proposed research questions.*

### *Patient Exit Interview*

**Sample**. As with the health worker interview, the service(s) of interest will determine which types of patients need to be interviewed at exit (prenatal care, child care, other). Typically, the enumerators will interview between 8-12 patients per service(s) of interest.

**Respondent**. If there are only a few patients on the day of the interview (12 or less), then the enumerators can interview all the patients. However, if there is a large number of patients for the service(s) of interest (15+), then the field team should randomly select 8-12 patients interest. It is highly recommended that field teams identify whether or not certain services are offered on desired interview day(s) in order to ensure that there are enough patients for the service of interest will be present on the chosen day.

**Timing.** The patient exit interview should last approximately **20-30 minutes**.

**Content**. There are eight sections in each of the two patient exit interview templates:

1. *Exit identification*: This section collects data on the health facility (for cross-checking purposes), the education level and marital status of the patient (or the patient's caretaker).
2. *Treatment and counseling*: This section collects data that is crucial to the analysis of the quality of care. This data reflects the provider effort during key maternal and child health consultations and provides a checklist of all the questions asked, examinations and lab tests conducted, medications and counseling provided during the consultation.
3. *Time and expense*: This section collects additional data related to the quality of care, with a focus on waiting time and time with the provider. In addition, this section collects information on any fees paid during the consultation.
4. *Satisfaction*: This section collects data on the patient's satisfaction related to various elements of her consultation.
5. *Security and trust*: This section collects data on the patient's feelings regarding security and trust of the various elements of her consultation and of the area around the health facility.

6. *Household 1 and 2*: This section collects data on the patient's socioeconomic status, including land ownership, household structure and asset holdings.
7. *Community health worker*: This section collects some information on the presence of community health workers and the services they provide in the community. This section may only apply to certain contexts.
8. *Traditional Birth Attendant (optional):* This section collects information on the presence of traditional birth attendants and the services they provide in the community. This section may only apply to certain contexts.

*Please note: The expected duration of the patient interview will likely run over 30 minutes if all sections are included in their entirety. Typically, data collected on patient satisfaction is on average very high, does not have much variability and de facto eliminates potential patients that do not even come to the facility on the grounds of dissatisfaction with services provided. It is therefore an optional section. In addition, depending on the research questions the team may want to consider reducing the data collected in sections 5 (security and trust) and 6 (household 1 and 2).*

### *Measuring Quality of Care*

A number of methods have been used to measure provider knowledge and delivery of health care (Franco, Daly et al. 1997; Hermida, Nicholas et al. 1999; Peabody, Luck et al. 2000; Bessinger and Bertrand 2001; Franco, Franco et al. 2002; Leonard and Masatu 2005; Leonard and Masatu 2006; Das and Hammer 2007).

Table 15 provides a summary of the quality of care methods, by comparing their validity, reliability, feasibility and relative costs.

### Table 15 Summary of Methods to Assess Provider Knowledge and Delivery of Care[14]

| Method | Strengths | Weaknesses | Reliability | Feasibility | Relative Costs |
|---|---|---|---|---|---|
| **Provider knowledge** | | | | | |
| Provider interviews – knowledge, attitudes | Measures practical knowledge; measures provider perceptions | Hawthorne effect – Altered behavior or performance of the health worker resulting from awareness of being part of an experimental study; Does not measure actual practice | Good for provider perceptions and knowledge (but many poor questions are used) | Can be done in large sample sizes, wide variety of cases, including rare events | Low |
| Provider vignettes | Measures practical knowledge, decision-making | Hawthorne effect – Altered behavior or performance of the health worker resulting from awareness of being part of an experimental study; Does not measure actual practice | Variability based on interviewer; can have low reliability (though may be good in certain circumstances) | Depends on high quality interviewer, uses lower sample size and limited variation in cases, some rare events | Higher interviewer qualifications and training required |
| **Health service delivery** | | | | | |
| Observation of care | Measures actual quality delivered | Hawthorne effect – Altered behavior or performance of the health worker resulting from awareness of being part of an experimental study; Measures optimum care delivered | Good -- Limited inter-rater problems, but not well documented | Requires high case load &/or common conditions; more intrusive than other methods | Modest training (e.g. 1 week) of literate or student health workers; supervision costs |

[14] Gupta and Peters (2010). Please see the tool "**3.01a RBF Indicators**"

| Method | Strengths | Weaknesses | Reliability | Feasibility | Relative Costs |
|---|---|---|---|---|---|
| Exit Interview – History taking and physical exam | Lower Hawthorne effect than observation if provider can be blinded to study | Does not measure actual sequence of physical exam | Poor patient/caretaker recall: low correlations with actual history documented | Can do large numbers, but better for common conditions | Modest training (e.g. 1 week) of literate or student health workers; supervision costs |
| Exit interview – Counseling, Perceptions, Patient characteristics | Measures what caretaker actually understands;<br>Measures immediate perceptions ;<br>Can measure key characteristics (e.g. wealth) which can be compared with population parameters | Perceptions at point of care may not reflect those measured later | Good – limited inter-rater problems; some questions have shown good correlation with actual history | Can do large numbers if facilities have high volumes, better for common conditions | Modest training (e.g. 1 week) of literate or student health workers; supervision costs |
| Simulated clients (mystery patients) | Lowest Hawthorne effect if able to keep interviewer blinded | Poor verisimilitude for children and women in delivery<br>Not possible for pregnant women and sick children (ethical issues) | Variability based on actor | Low sample size; least variation in conditions to test | Higher interviewer qualifications and training can increase costs |
| Patient record review | If good record keeping, can reflect care intended to be provided; Better for diagnosis and treatment (if systematically recorded) | Poor records are the norm; Limited information on actual tasks performed reflects what providers say patient care was rather than actual care delivered; | Poor, especially for non-standardized record keeping | Non-intrusive, quick, but records rarely of adequate quality for use other than volume of service | Cheapest |

The impact evaluation toolkit contains three types of vignettes:

1. Case scenario vignettes (prepared by Shivam Gupta and David Peters, Johns Hopkins University): In these vignettes, the enumerator reads the case of a patient with particular symptoms, and asks the provider for all of the actions and prescriptions that the provider would take to provide this patient with the most appropriate treatment. In most of the scenarios, apart from the initial reading of the case, the enumerator does not provide any further information to the provider. However, in one of the vignettes, the case has two parts. These vignettes can be found on sheet 12 of the Health Worker questionnaire.
2. Vignette on protocol for prenatal care: This vignette tests provider knowledge of the protocol for a first prenatal visit for a pregnant woman. The case scenario does not contain any information on additional symptoms other than visible pregnancy. This vignette can be found on sheet 13 of the Health Worker questionnaire.
3. World Bank-ISERDD vignettes (developed by Jishnu Das and Jeffrey S. Hammer, in collaboration with the Institute for Socio-Economic Research on Democracy and Development (ISERDD) (Das and Hammer, 2004; Das and Hammer, 2007): In these vignettes, the enumerator reads an initial case of a patient with particular symptoms. The enumerator then asks the provider to treat him/her as if he/she were the patient. The provider can ask the enumerator/“patient” any question to be able to come to a diagnosis, and can also propose exams. The enumerator/“patient” provides the answer to the questions of the provider, and gives the results of the proposed exams to the provider. The provider is asked to come to a diagnosis and treatment proposal. There may be follow-up questions from the enumerator as to the treatment that the provider would give to the patient. Please note that these vignettes are not included in the Health Worker questionnaire; however, they are a separate tool in the **health facility questionnaire** folder.

In addition to vignettes, the Toolkit includes a direct observation module that was also developed by Jishnu Das and Jeffrey S. Hammer, in collaboration with the Institute for Socio-Economic Research on Democracy and Development (ISERDD) (Das and Hammer, 2004; Das and Hammer, 2007): This very short module allows one to record basic information on the activities of the health provider during the visit, whether history questions were asked, if the provider conducted any physical exam, ordered certain tests and prescribed medication. This module was developed in the Indian context, where most visits are very fast. Teams should consider tailoring this module to their country context.

## Household Questionnaires

Household surveys allow the IE team to measure outcomes at a population-level. One of the key risks of an RBF project is the incentive for providers to over-report output levels. Although strong verification systems are required for a functional RBF, by using independent, primary data from the population-level, the impact evaluation is able to provide evidence of whether the impacts of the RBF program are trickling down to the target population of the program.

### *Household-level*

**Subjects**. Field teams should complete one household-level questionnaire for each household selected for the study as defined by the sampling plan.

**Respondent**. The main respondent for the household-level questionnaire is the best informed person in the household. This can be the head of household and/or spouse. However, the main respondent may ask for support from other household members on specific questions.

**Timing**. We estimated that the household interview will take **60-90 minutes** depending on how many modules are included. Since the respondent may have other responsibilities, the field team may plan two visits to the household and return at a later time in the day, or the next day, to complete the interview when the respondent has time.

**Content**. The household survey was developed in consultation with several RBF IE teams and includes the following sections:

1. *Roster*: This section collects the basic demographic data, including age, birth date, marital status and parental education levels, for all household members.
2. *Education*: This section collects data on current and completed levels of education, current attendance and time allocation for all household members 5 years and older.
3. *Labor*: This section collects data on current primary and secondary employment activities (either income generating or other), as well as other sources of compensation such as insurance, unemployment or retirement benefits for all household members 12 years and older.
4. *Housing*: This section collects data on all characteristics of the house, such as floor, roof and wall material, water and sanitation, fuel sources and rent.
5. *Assets*: This section collects data on the asset holdings and value of assets of the household, including land, equipment and animals.
6. *Income:* This section collects data on all possible sources of income, including investments, rentals, scholarships, remittances and inheritance.
7. *Consumption*: This section collects data on all consumption categories, including food, consumables and durables over a weekly, monthly and annual basis.
8. *Mortality*: This section collects data on any deaths of household members in the last 12 months, as well as the cause of death.
9. *Health Status and Utilization*: The section collects data on morbidity of all household members. This data is complemented by a more in-depth section in the maternal and child questionnaires.

19. *Contact*: This section collects data on how to contact the household for any follow-up information or surveys.

*Please note: The expected duration of the household-level interview will likely run over 90 minutes if all nine sections are included in their entirety. If time is a significant constraint we suggest that the team first consider reducing the data collected in sections 5 (assets), 6 (income), 7 (consumption) and/or 8 (mortality). However, please keep in mind that the information on assets / transfers / consumption provides valuable indicators on poverty, particularly important for conducting equity analysis.*

### *Female and Child Health Modules*

**Subjects**. All female household members 15-49 years old residing in households selected for the study should be interviewed for the maternal health modules, and there should be a minimum of one maternal health interview per household. All children 0-5 years residing in households selected for the study.

**Respondent**. The respondent(s) for the maternal health modules is the female household member(s) 15-49 years old. The team should not accept proxy response, i.e. one household member responding on behalf of another. The respondent for children 0-5 years is each child's primary caregiver.

**Timing**. We estimate that the maternal health interview will take **60 minutes** per female household member. Since  the respondent may have other responsibilities, the field team may plan two visits to the household and return at later time in the day, or the next day, to complete the interview when the female household member(s) is available. We estimate that the child health interview will take **30 minutes** per child. Since the respondent may have other responsibilities, the field team may plan two visits to the household and return at later time in the day, or the next day, to complete the interview when the caregiver(s) is available.

**Content**. The maternal health instrument has the following sections:

10. *Activities of Daily Living*: This section includes data collection on activities of daily living (ADL), and measures a person's physical ability to do common daily life activities.
11. *Mental Health*: This section collects data to assess the woman's mental health, and any treatment for recent depression and/or anxiety.
12. *Pregnancy History*: This section collects summary data on the pregnancies within the woman's lifetime, including live births, miscarriages and stillborns, as well as a summary of all the woman's living and non-living children.
13. *Antenatal and Postnatal Care*: This is an extensive section which collects data on the woman's prenatal, delivery and post-natal care for pregnancies in the last 2 years (this time frame can be adjusted based on the country impact evaluation objectives). For most of the RBF projects, prenatal, delivery and post-natal care utilization are core indicators for the success of the project. This section collects data on service utilization and quality of care (as measured by provider's adherence to national protocol).
14. *Reproductive Health*: This section collects data on the woman's desire for more children, history of contraceptive use, as well as current use.

15. *Vaccination*: This section collects data on the child's vaccination history, both at the facility and during community health campaigns.
16. *Anthropometrics*: This section collects data to measure the child's nutritional status by collecting the child's height and weight. This data is used to compute the child's Z-score. In some contexts, this data may also be collected for women.
17. *Other Biomarkers*: This section collects data on additional biomarker tests, such as malaria and anemia, which may be conducted during the household survey. In some contexts, this data may also be collected for women.
18. *Community health workers*: This section collects data on the woman's satisfaction with community health worker services. This may be adapted or removed depending on the country context. In some countries, the introduction of RBF has increased collaboration between the health facility and community health workers in order to induce demand for key services.

*Please note: The expected duration of the maternal interview will likely run over 60 minutes if all nine sections are included in their entirety. If time is a significant constraint the team may first consider reducing the data collected in sections 10 (Activities of Daily Living), 11 (mental health) and 18 (satisfaction). Depending on the objectives of the evaluation, the team may want to reduce sections 14 (Reproductive Health). Depending on time and budget constraints, the IE team may decide to collect biomarker data for a sub-sample of children and/or women, or eliminate it altogether (sections 16 and 17). Eliminating or reducing section 12 (pregnancy history) would be tricky because the information is being used as the basis for determining whom to interview in section 13 (antenatal and postnatal care). Section 13 is critical for the calculation of maternal care indicators.*

## Community Questionnaires

**Subjects**. Teams may collect one community-level module collected for each village selected for the study.

**Respondents**. The respondent(s) for the community modules should be the community leader(s) or local administrator most familiar with the community-level characteristics.

**Timing**. The community interview should take **45 minutes**.

**Content**. The community instrument has ten sections:

1. *Direct observation*: This section requires the enumerator to collect data based on direct observation of the community, including information on community sanitation, cleanliness, housing and topography.

2. *Composition*: This section collects data on the composition of the interview panel in case multiple community leaders or representatives are required to complete the community interview.
3. *Demography*: This section collects data on the number of homes and people in the community, as well as the main religion, language, and ethnicity.
4. *Basic Services*: This section collects data on the community's access to basic services, including health facilities, schools, roads, markets, water and sanitation.
5. *Social Capital*: This section collects data on community-level organizations and memberships, as well as women's rights to land and inheritance.
6. *Economic Activities*: This section collects data on the main economic activities of the community members.
7. *External Shocks*: This section collects data on any major external shocks that have impacted the community in the last 10 years, including floods, earthquakes, droughts, and disease.
8. *Programs*: This section collects data on the recent social programs, including health, education, water and sanitation and access to credit, that have been introduced in the community in the last 3 years.
9. *Prices*: This section collects data on the prices of food, health services and education.
10. *Costs*: This section collects data on the cost of the RBF project within the community.

*Please note: While the community level data allows for teams to control for additional observable characteristics in the analysis, there should be a balance between the treatment and comparison areas on these observables because of randomization. The module is included as a reference but may not be needed.*

# Module 5

# Implementing the Data Collection

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

# Module 5. Implementing the Data Collection

| Main Recommendations and Available Tools for this Module | | | |
|---|---|---|---|
| **Recommendations** | **Critical** | **Important** | **Nice to have** |
| • The impact evaluation team and survey firm should define the protocol for uniquely identifying observations in the data bases, as well as linking across databases. | ✓ | | |
| • The impact evaluation team should define the protocol for identifying the treatment and comparison areas within the databases. | ✓ | | |
| • The quality and duration of the training of field teams are key to the success of data collection. | ✓ | | |
| • While survey firms are in charge of data collection, the impact evaluation team should work with the survey firm to ensure appropriate and timely reporting on field work. | ✓ | | |
| • The research protocol and survey manuals should contain all the information needed by the survey firms to ensure data collection is conducted ethically and according to plans. | ✓ | | |
| • The safety and confidentiality of the data collected should be safeguarded carefully during data collection and entry. Field teams should report any logistical or security challenge. | ✓ | | |
| • The impact evaluation team should closely monitor the quality of data collection and data entry, and may want to hire a data quality expert to help in this process. | | ✓ | |
| • Local survey firms may have limited capacity in data entry programming, entry and management. The Toolkit contains data entry forms for CS-Pro software that correspond to the household and health facility questionnaires in the toolkit. | | ✓ | |
| • It is preferable to enter the data concurrently with field work, rather than after its completion. | | | ✓ |

**Tools**

- 5.01 Interview Duration Tracking Sheet
- 5.02 Enumerator Evaluation Form
- 5.03 Survey Progress Report I (Word)
- 5.04 Survey Progress Report II (Excel)
- 5.05a Household survey Field Manual
- 5.05b Household survey Training Program
- 5.05c Household survey Training PPTs
- 5.06a Health Facility Survey Field Manual
- 5.06b Health Facility Survey Training PPTs
- 5.07 Survey Training, CAR & Cameroon Examples
- 5.08 Health Facility Supervisor Checklist

- 5.09 Health Facility Arrival Checklist
- 5.10 Health Facility Supervisor Tracking Form
- 5.11a Daily Listing of Under 5 Exit Interviews
- 5.11b Daily Listing of ANC Exit Interviews
- 5.12 Cash Management Sheet

## Module Contents

In this module, we give an overview of the steps involved in implementing the data collection. We cover all of the steps that need to be taken after the survey firm comes on board, which include: (i) determining the sampling frame and the sample; (ii) defining unique identifiers; (iii) defining the treatment and comparison identifiers; (iv) pre-testing the questionnaires; (v) planning and managing data entry; (vi) planning field work; (vii) recruiting and training field teams; (viii) pilot test; (ix) managing field work; (x) reporting.

## Sampling Frame and Sample

In the design of the impact evaluation, the IE team should have defined (Cf. Module 3):

(i) The unit of randomization: the unit level that was used for assignment to the treatment and comparison groups;

(ii) The inclusion criteria: those criteria that are used to decide whether a unit is to be included in the sampling frame. Inclusion criteria are critical because, for equal sample sizes, they can lead to wide variations in the power of the study. The inclusion criteria define the population of interest.

(iii) The number of units in each arm of the study and overall sample size at each level (health facilities, health workers, households, women, or children).

During the preparation of the data collection, the IE team will need to put together *first the sampling frame, and then the sample*. As a reminder, the sampling frame is the list of units from which we will select the sample. The sampling frame should be representative of the population of interest. Practically, the sampling frame is a physical or electronic comprehensive list of units that could potentially be sampled. The sample is a sub-list that is drawn from the sampling frame, and it lists those units that need to be interviewed by the survey firm.

In most evaluations of RBF, the randomization will occur at a geographic level, e.g. districts, sectors, departments, etc. The sampling frames for all lower units (e.g. health facilities, health workers, etc.) will be limited to only those geographic units that belong to an arm of the evaluation, either treatment or comparison.

### Health facilities

If the Ministry of Health can provide a comprehensive list of health facilities, then the IE team should be able to extract the sampling frame from this list using the established inclusion criteria, e.g. “All facilities, public and private, located in the districts that belong to one of the evaluation arms”. From the sampling frame, the IE team and the survey firm will then need to extract the sample. In many countries, it will be necessary to survey *all* health facilities included in the sampling frame so as to reach sufficient power for the evaluation. In some cases, there may be more health facilities in each of the treatment and comparison groups than the number required to achieve sufficient power. In that case, the IE team may decide to select a random sample.

2012/04/03

**Country Spotlight: Sampling frame and sample**

**Rwanda Community Performance based financing program**

In the Rwanda Community PBF case, the *sampling frame* for facilities included all health facilities located in the 200 sectors identified for the evaluation. To maximize power, the *sample* of health facilities included all health facilities in the sampling frame.

## Health workers

The survey firm will first need to put together the sampling frame of health workers. For example, if the inclusion criterion was "health workers delivering prenatal care services on the day of the health facility interview", then the survey firm should list all of those health workers. Then the firm will require instructions as to how to draw the sample: e.g. "randomly draw two health workers delivering prenatal care services on the day of the health facility interview".

2012/04/03

**Country Spotlight: Selection criteria for the sampling frame**

**Rwanda Community Performance based financing program**

In the Rwanda Community PBF program, there are a total of four Community Health Workers (CHWs) assigned to every village: (i) a Maternal and Child Health focal point, (ii) a social affairs focal point, (iii) an HIV/AIDS focal point and (iv) an administrative focal point. The inclusion criteria for the sampling frame for CHWs working in selected villages located in the 200 sectors was "The community health worker responsible for Maternal and Child Health Services, and another randomly selected CHW regardless of gender or position".

## Households

Most evaluations will have an inclusion criterion for households of the following type "Households with children under X years of age living in the catchment area of a sampled health facility". While the criterion is simple enough, assembling the sampling frame at the household level can be quite time consuming because one would need to list all households in the catchment area; in most cases, there are many of households, and there is usually no readily available list.

**Country Spotlight: Selection criteria for the sampling frame**
**Rwanda Community Performance based financing program**

In the Rwanda Community PBF Case, the inclusion criteria for the sampling frame for households located in selected villages located in the 200 sectors was “the household with the most recent birth (at least one child 0-4 months)”.

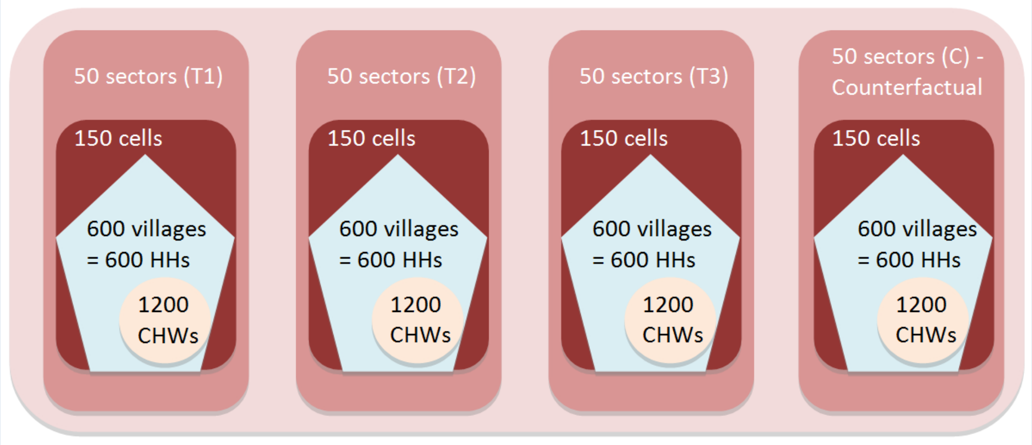
As an alternative to listing every single household in the catchment area of a facility, one can resort to two-stage random sampling, which can be done as follows:

*(i) Determine the Primary Sampling Units (PSU) that constitute the catchment area of the health facility.* Preferably, this is the smallest geographic unit such as a village or census block.

*(ii) Randomly select PSUs for each health facility catchment area.* If the health facility catchment area can be defined from data that are available at the *central level*, then the Primary Sampling Units (PSU) in each catchment area can be listed. However, in many countries, it will not be possible to define the health facility catchment area from centrally available data, and the catchment area information will need to be obtained from the staff in each of the *health facilities separately*. In this case, the household sampling may have to be done after the health facility surveys: the survey firm can be responsible for listing all PSUs in the catchment area of each facility and providing this information to the IE team, who can then draw a random sample of PSUs from this list.

### Country Spotlight: Random selection of Primary Sampling Units
### Rwanda Community Performance based financing program

In the Rwanda Community PBF Case, power calculations demonstrated that 12 households were required for each of the 50 sectors in each of the four study arms => 50 sectors*12 households = 600 households per study arm. The selection of PSUs was defined at a central level, as the IE team obtained the list of all cells and villages in each sector from the Ministry of Health Community Health Desk. Using these data, three cells were selected for each of the 200 sectors in order to minimize geographic disbursement within sectors. Second, four villages were selected for each of the cells. This resulted in a random sample of 12 villages per sector.



*(iii) List all households in sampled PSUs, the so-called "household listing".* If the country has recent census data that can be linked to the health facility PSUs, then the IE team will be able to extract the list of households in each sampled PSU from the census, and this would constitute the sampling frame for households. In most cases though, such information will not be available; therefore, the survey firm will need to list all household(s) that meet the household inclusion criteria and live in the sampled PSU. The complete household listing should collect basic information on the household and its members, including age and sex, in order to ensure a sample frame to draw an eligible household following the inclusion criteria. This household listing may be done prior to, or concurrently with the household data collection, but in either case it is crucial to have adequate supervision of the process to ensure that the household listing accurately represents the population of interest. We also recommended that the team responsible for the household survey does not also conduct the household listing. When the same team conducts both the household listing and the household survey, there can be an implicit incentive to exclude households that are hard to reach within the PSU from the household listing, and this can result in an under-representation of hard-to-reach households in the survey.

**Country Spotlight: Household Listing**

**Rwanda Community Performance based financing program**

In the Rwanda Community PBF Case, the IE team did not have access to a recent census. Initially, the team planned to conduct a household listing in each of the villages selected. However, due to financial and time constraints, the IE and field teams could not conduct the listing. Rather than conduct a full listing of every village to identify the most recent births (0-4 months), the Field Supervisor met with the village leader and CHWs in each village to identify households with births in the last 4 months (on average, there are 2 births per month per village). This sampling strategy was vulnerable to risks of moral hazard, and required rigorous field supervision by the Project Manager and IE team.

*(iv) Sample households from the household listing.* Once the survey firm has completed the listing of households in each sampled PSU, the IE team will be able to randomly draw a sample of households to be included in the household survey.

## Defining Unique Identification Codes

During data collection, the survey firm will collect many pieces of information at the health facility, community, household and individual levels, and field teams will be visiting numerous locations, often several times. Furthermore, many more people may process the questionnaires, data, and biological samples collected, and a few years after the completion of the baseline data collection, another (or the same!) survey firm will want to go back to the same locations to collect another round of data. In addition, analysts want to be able to link different sources of data to each other. For example, information on a household is only useful insofar as we know in which catchment area the household is located. Therefore, the data will need to be organized in such a way that anyone working on them will be able to properly track the origin of each response, but without violated respondents' confidentiality.

Proper organization of survey (and other) data rests on the consistent use of identification codes (ID codes) throughout the data collection and entry processes. This should ensure that all information can be traced and linked, no matter where it came from. Establishing this protocol of ID codes is normally the responsibility of the principal investigator, and the Principal Investigator will typically need to work with the survey firm to develop a protocol for these codes. Without a clear ID code protocol, the principal investigator will not be able to finalize the questionnaires and the survey firm will not be able to plan field work. During baseline field work, the survey firm will need to use the agreed ID code protocol to label all paper questionnaires, biological samples, etc. The same ID codes would then have to be used during the follow up survey(s).

There are two types of ID codes, each with a distinct use.

- Geographical ID codes identify the exact location of each unit that is included in the database, but should not be made publicly available because they do not maintain respondent anonymity.
- Field ID codes uniquely identify each unit for which data is being collected, whether it is a person, household, village or health facility, but do not give information on the exact physical location of the unit. They can be used in databases that shared and used for analysis.

Both geographical and field ID codes are included in the original paper questionnaires.

Good ID codes, whether they are geographical or field ID codes, have the following characteristics:

- They uniquely identify a unit of observation: i.e. all information pertaining to a particular individual will bear the same ID code, whether it's the person's response to a questionnaire or her blood sample.

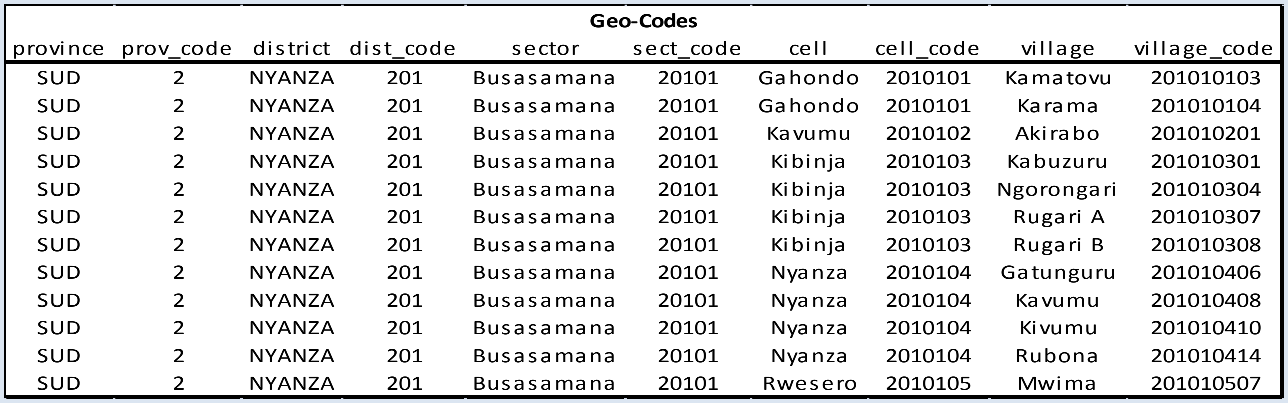
- They are numeric. ID codes should not include any letters or special characters. They should only contain numbers.

### Geographical ID codes for raw databases

Raw databases normally contain a geographical ID code that identifies each location. Every geographic unit (region, province, district, village…) should have its own geographic ID code, and the geographic ID code should be constructed in a hierarchical way, starting from the highest geographical unit, so code for the smallest location unit has a unique ID. This code may be based on existing health information or census data in order to maximize compatibility between the impact evaluation databases and existing data sources.

### Country Spotlight: Defining Geo-Codes
### Rwanda Community Performance based financing program

Rwanda is geographically organized by provinces, districts, sectors, cells and villages. For the purposes of the CPBF evaluation, the randomization unit was the sector, with a total of 200 sectors in the sample. Surveys include data at the village level, as well as at the household and health worker level. The smallest geographic unit is the village, and therefore the database contains the geographical ID codes for province, district, sector, cell and village.



| | | | | Geo-Codes | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| province | prov_code | district | dist_code | sector | sect_code | cell | cell_code | village | village_code |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Gahondo | 2010101 | Kamatovu | 201010103 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Gahondo | 2010101 | Karama | 201010104 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Kavumu | 2010102 | Akirabo | 201010201 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Kibinja | 2010103 | Kabuzuru | 201010301 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Kibinja | 2010103 | Ngorongari | 201010304 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Kibinja | 2010103 | Rugari A | 201010307 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Kibinja | 2010103 | Rugari B | 201010308 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Nyanza | 2010104 | Gatunguru | 201010406 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Nyanza | 2010104 | Kavumu | 201010408 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Nyanza | 2010104 | Kivumu | 201010410 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Nyanza | 2010104 | Rubona | 201010414 |
| SUD | 2 | NYANZA | 201 | Busasamana | 20101 | Rwesero | 2010105 | Mwima | 201010507 |

While the principal investigators need to know the geographic location of the units sampled for the survey (district, health facility, etc) in order to generate panel data, ***these ID codes can easily compromise respondent confidentiality if shared***. For example, say that a database contains information on households, and that for each household, we are able to identify which village the household lives in. Say the household has 5 children and 17 cows and this information is available in the database. If the village only has few households with 5 children and 17 cows, then any user of the database could easily indentify which household responded to the survey. This would violate the household's confidentiality.

Therefore, the principal investigators should not share any databases that include geographical identifiers, and only the raw databases (which are not to be shared outside of the principal investigators) should contain the geographical ID codes.

## Field ID Codes for Data Collection and Data Sharing

Field ID codes are unique, simplified identification codes for each unit of observation, but they do not contain geographical information per se. Apart from protecting the confidentiality of respondents, field ID codes also facilitate data collection and data coding for the survey firm.

Assume for example that an IE team wants to survey between 400 health centers and corresponding catchment areas. This survey would have 400 "survey areas", which would each cover one health center and the corresponding respondents. We will need to use a three-digit numeric code, since a two-digit numeric code would cover at most 99 survey areas. For each survey area, there would be exactly one health facility, so health facilities can be numbered in the same way as survey areas. For each health facility/survey area, there typically would be several additional levels of data collection, for example health workers working at that facility, patients exiting the facility, and households living in the catchment area of the facility, which all need to be linked to the health facility code. These other units to be interviewed can typically be coded with an additional 2 digits. For example, in survey area number 232:

- Health facility: there would be one health facility with unique ID code 232.
- Health workers: If 3 health workers are selected for the health worker interview in facility number 232, the health worker questionnaires would be coded 232-01, 232-02 and 232-03.
- Antenatal care (ANC) patients: If a total of 10 ANC patients are interviewed when exiting facility 232, then the questionnaires would be coded 232-01 to 232-10.
- Under 5 patients: If a total of 8 guardians of under 5 patients are interviewed when exiting facility 232, then the questionnaires would be coded 232-01 to 232-08.
- Households: If a total of 12 households are interviewed in the catchment area of facility 232, then the household questionnaires in this catchment area will be coded 232-01 to 232-12.

With this simplified coding, the survey firm will be able to easily track the completion of surveys by area: for each survey area, the field team will track completion of an expected "package" of surveys and measurement. In the above example, a package 1 health facility assessment (Form F1), 3 health worker questionnaires (Form F2), 10 ANC exit questionnaires (form F3), 8 under 5 exit questionnaires (Form F4) and 12 household questionnaires (Form HH).

2012/04/03

### Country Spotlight: Defining Field IDs

### Rwanda Community Performance based financing program

The Rwanda CPBF evaluation covered 200 sectors, which corresponded to the survey areas. In each sector, the IE sampled 1 health facility, 12 villages per health facility and 2 community health workers per village. This resulted in a total of 1 health facility, 12 household and 24 community health worker interviews per sector. Three field ID codes were generated:

- field_id1: A three-digit ID that uniquely identifies the sectors and health facilities
- field_id2: A three-digit ID that uniquely identifies the villages and households selected for each health facility
- field_id3: Identifies the community health workers selected for each health facility

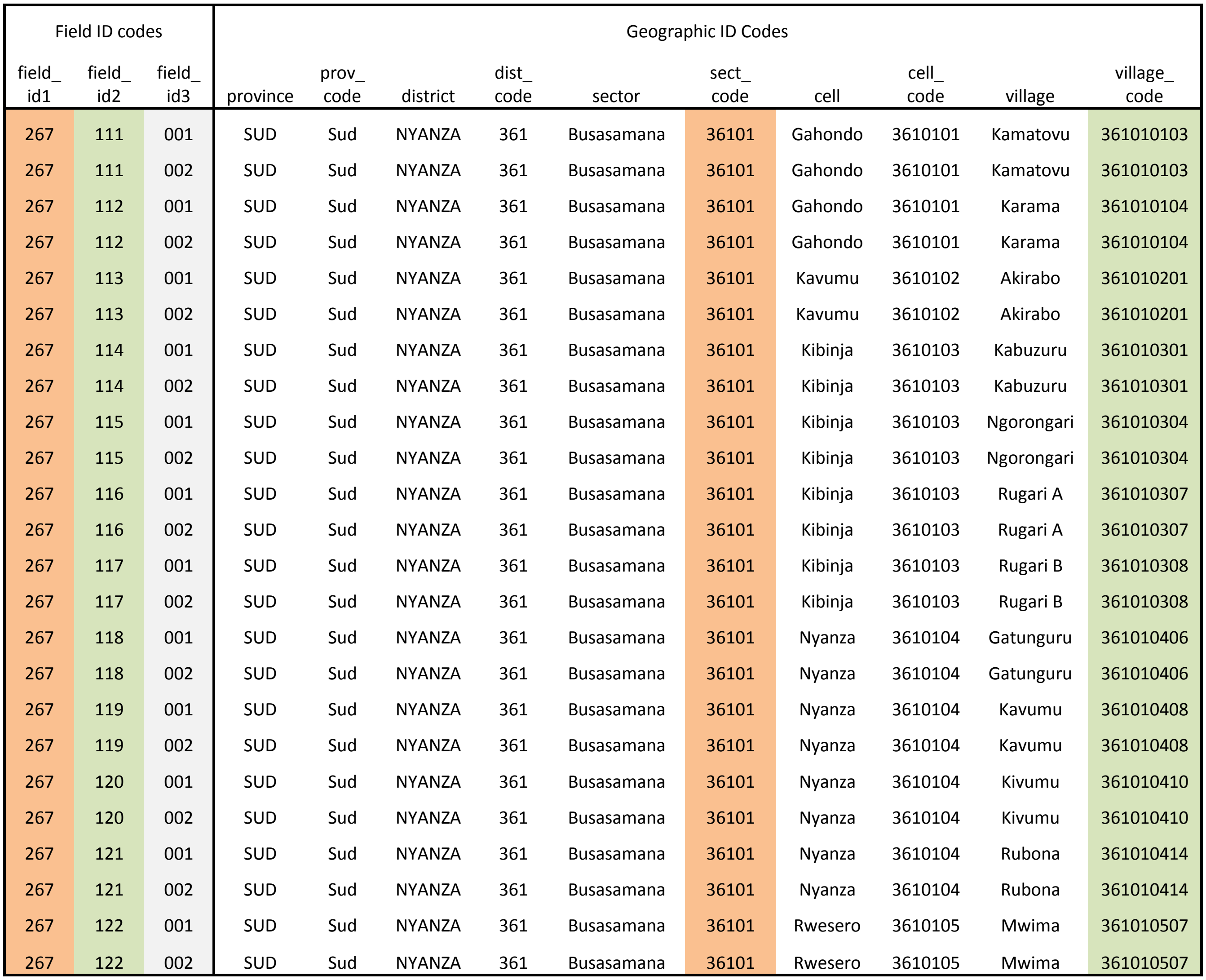
### The ID control file: linking geographical and field ID codes

The ID control file is a file that lists the field ID codes and the corresponding geographic ID codes, typically in an excel format. This file is normally prepared by the survey firm in collaboration with the principal investigator, before the start of the baseline survey.

The ID control file plays a crucial role at the time of the follow up survey, when the investigators may want to go back to the field and interview in the same health facilities and catchment areas as in the baseline. If they only had field ID codes, they would be able to know which facility in the dataset corresponds to which facility in the country, in other words they would not be able to physically locate the health facilities that are in the baseline dataset. With the ID control file however, the investigators will be able to locate each baseline facility geographically, and they will be able to use the same field IDs in the follow up survey as in the baseline survey.

Note that the ID control file is a sensitive document, because with it, anyone could identify the physical location of respondents in the dataset, and this would violate respondent confidentiality. It is the responsibility of the principal investigator to ensure that this information is stored safely. In Module 6, we discuss the use of a data enclave to safeguard this type of sensitive data.

### Figure 9: Screenshot of an ID control file



| Field ID codes | | | Geographic ID Codes | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| field_ id1 | field_ id2 | field_ id3 | province | prov_ code | district | dist_ code | sector | sect_ code | cell | cell_ code | village | village_ code |
| 267 | 111 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Gahondo | 3610101 | Kamatovu | 361010103 |
| 267 | 111 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Gahondo | 3610101 | Kamatovu | 361010103 |
| 267 | 112 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Gahondo | 3610101 | Karama | 361010104 |
| 267 | 112 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Gahondo | 3610101 | Karama | 361010104 |
| 267 | 113 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kavumu | 3610102 | Akirabo | 361010201 |
| 267 | 113 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kavumu | 3610102 | Akirabo | 361010201 |
| 267 | 114 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Kabuzuru | 361010301 |
| 267 | 114 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Kabuzuru | 361010301 |
| 267 | 115 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Ngorongari | 361010304 |
| 267 | 115 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Ngorongari | 361010304 |
| 267 | 116 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Rugari A | 361010307 |
| 267 | 116 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Rugari A | 361010307 |
| 267 | 117 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Rugari B | 361010308 |
| 267 | 117 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Kibinja | 3610103 | Rugari B | 361010308 |
| 267 | 118 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Gatunguru | 361010406 |
| 267 | 118 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Gatunguru | 361010406 |
| 267 | 119 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Kavumu | 361010408 |
| 267 | 119 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Kavumu | 361010408 |
| 267 | 120 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Kivumu | 361010410 |
| 267 | 120 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Kivumu | 361010410 |
| 267 | 121 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Rubona | 361010414 |
| 267 | 121 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Nyanza | 3610104 | Rubona | 361010414 |
| 267 | 122 | 001 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Rwesero | 3610105 | Mwima | 361010507 |
| 267 | 122 | 002 | SUD | Sud | NYANZA | 361 | Busasamana | 36101 | Rwesero | 3610105 | Mwima | 361010507 |

Note: In the case of the Rwanda Community Performance-Based Financing impact evaluation, field_id1 code uniquely identifies each sector, but does not include any geographic information. field_id2 uniquely identifies each village, or each household since one household per village was interviewed, but does not include any geographic information. field_id3 uniquely identifies each Community Health Worker, since two Community Health Workers per village were interviewed, but does not include any geographic information.

Please note that field IDs and geo-codes were modified in this table to keep any identifying information confidential.

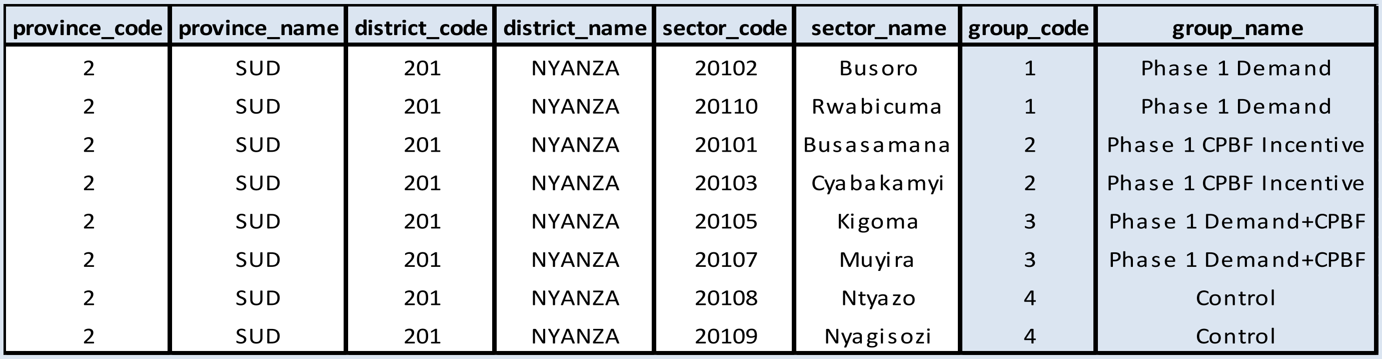
## Treatment and Comparison Identifiers

The IE team will need to ensure that there is a way to identify which geographic areas are assigned to which treatment and comparison groups, so there should be a list of geographical areas and their corresponding assignment to the treatment or comparison groups.

On the other hand, analysts will also need to be able to identify which areas in the dataset are assigned to the treatment and comparison groups, so there should be a list of field ID codes and their corresponding assignment to the treatment and comparison groups.

Note that the file that links the geographical information and the treatment/comparison information, and the file that links the field IDs and the treatment/comparison assignment, should be separate. If the geographical information, the field IDs and the treatment/comparison assignment information were all included in one single file, it would be possible to use it to physically locate respondents, and this would violate confidentiality.

Figure 10: File with Geographic and Treatment/Comparison Identifiers



| province_code | province_name | district_code | district_name | sector_code | sector_name | group_code | group_name |
|---|---|---|---|---|---|---|---|
| 2 | SUD | 201 | NYANZA | 20102 | Busoro | 1 | Phase 1 Demand |
| 2 | SUD | 201 | NYANZA | 20110 | Rwabicuma | 1 | Phase 1 Demand |
| 2 | SUD | 201 | NYANZA | 20101 | Busasamana | 2 | Phase 1 CPBF Incentive |
| 2 | SUD | 201 | NYANZA | 20103 | Cyabakamyi | 2 | Phase 1 CPBF Incentive |
| 2 | SUD | 201 | NYANZA | 20105 | Kigoma | 3 | Phase 1 Demand+CPBF |
| 2 | SUD | 201 | NYANZA | 20107 | Muyira | 3 | Phase 1 Demand+CPBF |
| 2 | SUD | 201 | NYANZA | 20108 | Ntyazo | 4 | Control |
| 2 | SUD | 201 | NYANZA | 20109 | Nyagisozi | 4 | Control |

Note that field IDs are not included in this file.

## Pre-Testing the Questionnaires

Pre-testing the questionnaires is typically done by the IE team in conjunction with a select number of managers and enumerators from the survey firm. The objectives of the questionnaire pre-test include:

- Ensuring that questionnaires are properly adjusted to the local context
- Reviewing the translation of the questionnaire in order to ensure it is adapted to the local context.
- Identifying any adjustments that need to be made to the questionnaire in order to minimize “don’t know” responses and refusals.
- Testing field work organization and management: eg. the division of labor between enumerator in the facility questionnaire, and the role of the team supervisor
- Testing the gathering of the biometric measures and adjusting procedures and equipment if necessary.
- Evaluating enumerators’ ability to administer the questionnaires.

- Estimating how long it takes to administer each questionnaire in the field. The IE team and the survey firm need to understand the average duration of each data collection component in order to properly plan field work. See **Duration of Interview Tracking Sheet.**
- Documenting the changes and adaptations suggested by the team.

## Managing Data Entry

In this section, we review a number of important considerations around data entry.

**Approach.** As previously mentioned, we recommend that data entry be done in parallel with field work using the CAFE approach, rather than waiting until field work is finished.

**Software**. The data entry software (preferably CSPro) should be programmed to include range checks where necessary and to identify key-punch errors, which give impossible values for that question (e.g., entering 7 on a question which asks a scale from 1 to 5, or entering 150 for age, etc.). The data management software should also be programmed to flag internal inconsistencies between questions. If any of these flags are triggered, an enumerator or the field team supervisor should first recheck the data, and if needed, revisit the household to clarify the inconsistent responses. For example, the software should flag any observations that trigger the following inconsistency checks:

- Age is greater than her mother's age minus twelve
- Years of education is greater than age minus four
- Total number of rooms for exclusive use is greater than total number of rooms
- Gender checks across modules – for example, pre-natal care should not be entered for a male.

**CSPro Data Entry Modules** are available for country teams to adapt and use. By following the questionnaire adaptation guidelines detailed above, the country teams should be able to maximize their use of the available data entry programs and minimize additional costs for their adaptation.

**Hardware**. The IE team should assess the hardware that the survey firm proposes to use for data collection. Data entry and management conducted on out-dated, old and unreliable computers increases the security risk to the data. It is worthwhile to conduct this assessment prior to contracting the survey firm(s) (See Hiring the Survey Firm).

However, computers and laptops used in data entry and data management do not need to be latest-model machines. CS-Pro can run on less expensive computers that have the following characteristics: at least 1 Gb of RAM memory, a processor with a modest speed of around 1.5 ghz, some 100 Gb of free disk space, screens with a contemporary resolution (at least 1266 x 768), a reasonable screen size (not less than 14" for field work laptops, recommended 17" to 19" for desktops used in an central entry room), at least 3 USB ports, and basic networking capability. At the same time, the firm should have another more powerful computer (either a laptop or a desktop), which can be used as concentrator.

**Anti-Virus Protection**. The IE team should also assess the survey firm's anti-virus protection software prior to initiating data collection. Without a high quality, functional anti-virus, there is an increased security threat to the data. We recommend that machines have reliable, lightweight antivirus such as quickly updatable NOD32 or Karpersky, rather than heavyweight packages such as Norton or McAfee that must be constantly updated.

**Networking Computers**. The concentrator computer needs to be able to access all field work laptops and/or desktops used for centralized data entry. We suggest using a private network based on wifi to avoid complicated networking protocols. In most of cases, a simple mapping schema of the C drives in each computer will give the concentrator access to those drives as remote units of the concentrator, i.e. allocating them drive-letters M / N / … / Z. This approach allows the concentrator computer to take control of up to 14 entry stations. We suggest that Ethernet-cabled networks should be avoided unless they are previously available as they are expensive and difficult to maintain.

**Labeled Boxes to Hold Paper Questionnaires**. The survey firm should establish a protocol for storing and managing the paper questionnaires as they come out of the field. One recommendation is to store the paper questionnaires in folders or boxes (depending on the size) and label them using the Field ID code. With this strategy, the team can easily monitor whether or not data collection from a health facility or health area is complete. The packs or boxes should be stored in easily accessible shelves, sorted in ascending order by the unique RBF IE code. The room that holds the paper questionnaires should be protected from unauthorized access at all times.

**Post-processing the data.** Once the data is entered into the raw data files, the survey firm team should conduct a final re-examination of the data to compare the data reported by the field supervisors and the data recorded in the data entry. At a minimum, we recommend the following checks:

- Check the raw data files to ensure that the number of questionnaires filled (as per the supervisors' reports) corresponds to the number of records in the data files. For example, if the supervisors reported 7 ANC exit interviews in facility number 232, the ANC exit interview data file should contain 7 records for this facility.
- Inspect the raw data files to ensure there are no corrupted files.
- Compare coding for each variable in the dataset to the coding in the questionnaire – are any responses out of range? As a ballpark figure, 2% of out of range values is acceptable, but 10% is a serious problem.
- Manually check all "Other" entries to ensure that the description does not fall under one of the pre-coded responses. This task can require visual inspection of several thousand cases, can easily take 6 to 8 days, and will often require re-examining the original paper questionnaires to verify the accuracy of recorded response.

**Labeling Databases.** We recommend the following protocol for labeling and organizing data files.

`country'_`x'_`nn'

*Country*: The prefix identifies the country where the data were collected in. For example, in Nigeria, the prefix is “NG”.

*x*: The x represents the survey instrument with the following coding, for example:

- HH Household
- F1 Health Facility
- F2 Health Workers

*nn*: The nn represents the sequence number of the file for each form.

- 00 is for variables that have only one response, such as variables on household assets (one response per household) or variables on facility characteristics (one response per health facility)
- 01 if for variables that have multiple occurrences within a questionnaire, such as variables in the household roster (one questionnaire has education and health information on multiple household members)

2012/04/03

**Country Spotlight: Consistently Naming Databases**
**Nigeria State Health Program Investment Credit**

- NG_HH_00 would contain the household-level data from the household questionnaire
- NG_HH_01 would contain the household roster data, i.e. those variables that report information about individual household members.

**Double-Entry**. When using a centralized entry facility, we recommend that the survey firm use double entry (or "verification entry") to validate and measure the accuracy of data entered by the various operators. This can be done over a flexible proportion of the workload: initially, the IE team and survey firm can agree on the number of data files that should be double-entered for each survey area. Typically this would be done by the best entry. If the comparison between primary and verification entry has more than approximately 10-15 entry mistakes per questionnaire, the survey firm should take corrective action, such as reducing the speed of the primary entry, increasing the percentage of questionnaires to be double-entered, and in severe cases replacing the data entry operator. Note that data entry using the CAFÉ system uses an entirely different approach and that blind double entry is usually not needed under that system.

## Risks to Internal Validity of the IE Design during Data Collection

When moving from IE design to data collection, it is crucial to pay attention to potential threats to internal validity of the evaluation design. For an impact evaluation to be internally valid, one needs to maintain an accurate estimate of the counterfactual (i.e. what would have happened in the absence of the RBF program).[15] Data collection activities may pose a threat to the internal validity of the evaluation design if the survey firm ends up treating the treatment and comparison groups differently. By treating the two groups differently, the firm then introduces a bias, which can invalidate the design! Here are a few examples:

- **Example 1**: Data Collection Team 1 is the more experienced team and demonstrates a higher skill level in the training than Team 2. Team 1 is assigned to collect data in the treatment group and Team 2 is assigned to collect data in the comparison group. After the baseline data is received, the IE team runs difference in means tests and finds that the average means for outcomes in the treatment group are significantly lower than those in the comparison group. After some research, the IE team finds that Team 1 was better able to catch over-reporting and incorrect reporting by respondents than Team 2. So the data collected in the treatment group is more accurate than in the comparison group, and therefore it is not possible to compare the means of the outcome variables between the two groups.
- **Example 2**: During field work planning, the survey firm decides to start collecting data first in the treatment group, and then in the comparison group. Data collection in the treatment group starts in January, while data collection in the comparison group starts in April. After data collection, difference in means tests show that there is a much higher prevalence of malaria and reduced utilization of key health services in the treatment group compared to the comparison group. After some research, the IE team finds that January-March is the rainy season, while April-June is the dry season, both of which have clear seasonal impacts on key health indicators. While the data were accurately collected in both groups separately, they are not comparable between the two groups.
- **Example 3**: During the baseline survey, the project team asks the survey firm to deliver an envelope documents to the facilities in the treatment group. Upon opening the envelope, the head of the facility realizes that the facility will be participating in an RBF program, whereby the personnel will receive performance-based incentive payments. She quickly informs her staff of what is coming up and directs them to collaborate with the enumerators so as to “look good” to the RBF administrators. As a result, health workers in the treatment facilities end up being more cooperative and patient with the enumerators than those in the comparison facilities, and data are more accurate in the treatment facilities. As a result, the data from treatment and comparison facilities are no longer comparable.

[15] Please see 0 and “Impact Evaluation in Practice” for a discussion of how randomized assignment of treatment and comparison state is helpful in accurately estimating the counterfactual.

These examples illustrate the following points to keep in mind during data collection: (i) the timeline for data collection should not favor either treatment or comparison facilities; (ii) there should be no difference between the two groups in terms of field team abilities, supervision, etc; (iii) the survey teams should be kept fully separate from the implementation and monitoring of the RBF intervention itself. Whenever possible, field teams should not be informed of the treatment status of the facility, and should be trained to not ask or enquire about this status.

## The Field Work Plan

The field work plan is a document that is prepared by the survey firm and details how the firm will manage its field teams and how it will implement the facility and household surveys. Here are a few elements to look out for:

### Management of Field Teams

The field work plan should detail team composition, roles and responsibilities; the schedule; expected output and logistics.

- **Composition**. The plan should clearly detail how many teams will be deployed to the field, and the composition of each team (supervisor, enumerators, data entry personnel, other).
- **Roles and Responsibilities**. As each survey has multiple components, we recommend that the plan details which team member is responsible for completing which module(s).
- **Schedule**. The plan should include a completed **Sample Control File** which details the specific dates the field team will be interviewing in each unit of observation (health facility, household). This will incorporate the expected output as well, as this defines how long a field team will be in each location.
- **Expected Output**. The plan should include each field team's expected daily output for each questionnaire type.
- **Logistics**. Field team supervisors are responsible for managing logistics in the field, including managing the consumables stock (paper questionnaires, pens, materials for biomarker data collection, etc). In addition, field team supervisors manage travel and transportation, such as hotel arrangements and ensuring that there is a functional vehicle for the daily data collection (gas, tires, etc). The plan should include a discussion on the daily and/or weekly requirements in order to manage logistics for each field team.
- There are also specific issues to consider when developing the field work plan for the facility and the household data collection activities.

### Facility-level Data Collection

The facility-level data collection field work plan should take into consideration the following aspects:

- **Patient hours**. Field teams will first need to determine the start and end of patient hours because, in most cases, interviews with providers will need to take place after patient hours; conversely, patient exit interviews can only be collected during patient hours.
- **Days services are offered**. The field teams will need to know the days of that the services of interest are offered by the sampled facilities. For example, if facilities only offer prenatal care on Mondays, Wednesdays and Fridays, then field teams cannot plan to collect prenatal provider and patient exit interviews on Tuesdays or Thursdays. This has major implications for field work planning.
- **Scheduling.** In some countries, schedules for health facilities may be uniform and therefore they can be incorporated in the overall field work plan. In most cases though, the days and hours that services are provided are decided at the facility level. Therefore, field teams will need to gather this information before they can estimate the required number of days to complete all facility-level interviews. As a general rule of thumb, IE teams can expect that field teams will need 1.5-2 days per facility.

### Household-level Data Collection

The household-level data collection field work plan should take into consideration the following aspects:

- **Respondent availability**. The respondents in household interviews are the head of household and main caregivers, and therefore will have limited availability to respond to a 2-3 hour interview. Ideally, the field team composition should allow teams to conduct parallel interviews with several respondents within one household, so as to minimize the amount of time spent in any one household.
- **Duration of interviews**. The household questionnaire is quite long and complex, and can be expected to require **2-3 hours** per household. If the full questionnaire were to be administered in one session, the quality of the answers could be affected by the respondent's fatigue. Therefore, it is recommended that the survey firm schedule in the field work plan more than one visit to each household. The first visit should be for initial data collection, while the second one should involve interviewing any additional household members not present during the first interview and solving any inconsistencies discovered by the data entry operator. As discussed above, this would be for both CAFE and centralized data entry.

## Recruiting and Training Field Teams

The survey firm is responsible for recruiting field staff. However, it is in the IE team's interest to ensure that all hired staff are sufficiently qualified and that they meet a set of standard requirements.

- **Number of field staff recruited**. The survey firm should recruit and train at least 15% more individuals than are required to form the survey field teams. It should be expected that some enumerators will drop out even when the field work plan included measures to minimize enumerator fatigue and burn-out, such as proper field work management and questionnaire design. In those cases, the survey firm needs a contingency plan to ensure there are no interruptions to data collection.

- **Qualifications of field staff**. The survey firm should recruit and train individuals according to their role on the field team in accordance with local regulations. For example, anthropometrists and individuals responsible for biomarker data collection may need special clearance or qualifications, such as a nursing or other medical background. The firm should identify these requirements before it starts recruiting staff.

## Training Program and Materials

The training of supervisors, enumerators and data entry operators is an essential step in ensuring the quality of survey data.

- **Program.** General training should be given to all supervisors, enumerators and data entry operators – this will help create a team environment and give team enough flexibility to substitute roles in case a team member is temporarily absent due to illness or another emergency. Field team supervisors should receive additional training following the general training.
- **Logistics.** In the best case scenario, the survey firm will train supervisors, enumerators and data entry operators together in one central location, so that they receive the same training using standardized Powerpoint presentations. If training in a central location is not possible, then the survey firm will need to plan sufficient time and budget to provide standardized training in different locations. During budget negotiations with potential survey firms, the IE team should ensure that the survey firm is budgeting for travel, food (lunch and coffees in all cases, dinner for out-of town trainees) and lodging expenses for the supervisors, enumerators and data entry operators during the training. The Principal Investigator and survey firm will need to identify whether training can take place in one plenary group, or if there are too many supervisors, enumerators and data entry operators, should instead be divided into several sub-groups. In this case, the survey firm will still need to standardize training across sub-groups by using the same training materials among trainers.
- **Duration.** The training should be scheduled for a minimum of 2 weeks.
- **Content.** The training should include the following four main components.
    - **Theoretical review.** The trainers and trainees review the research objectives, the questionnaire content and each question in order to fully understand the objective of each question. This should be done in an interactive way, with ample time for questions and answers.
    - **Classroom practice.** Each trainee should have an opportunity to practice filling questionnaires in the classroom. The trainers could project the questionnaire and have one trainee fill it in front of the classroom while other trainees observe and participate. The trainers could also design case scenarios that are based on typical households

(perhaps those found during the supervisor training or piloting) and have enumerators complete the questionnaire based on the case that is presented. Another idea is to film a pilot interview and have the trainees fill in a questionnaire for the interview to test consistency across the trainees.

- **Field exercises.** After the theoretical review and classroom practice, the trainees should go to the field to administer the full questionnaire to a small number of households (outside the study sample). The following day, the team will meet together to discuss the results of the field exercises. This is an opportunity to clarify concepts, how to deal with "difficult" examples, skip patterns and communicate any additional issues related to questionnaire format and translation.
- **Evaluation.** Following the training, enumerators, supervisors and data entry operators should be evaluated based on their understanding of the questionnaire and their ability to correctly record data using the same test scenarios as used in the classroom practice. Enumerators, supervisors and data entry operators who do not meet the minimum requirements should ideally not be allowed to continue to participate in the survey.

**Field Team Training materials and proposed curricula, and Field Manuals** are available in the Toolkit.

## Pilot Test

The pilot test has the same key objectives as the pre-test, but the difference is that the pilot involves the entire field team after training has been completed. The objectives remain:

- Ensuring that questionnaires are properly adjusted to the local context
- Reviewing the translation of the questionnaire in order to ensure it is adapted to the local context.
- Identifying any adjustments that need to be made to the questionnaire in order to minimize "don't know" responses and refusals.
- Testing field work organization and management: eg. the division of labor between enumerator in the facility questionnaire, and the role of the team supervisor
- Testing the gathering of the biometric measures and adjusting procedures and equipment if necessary.
- Evaluating enumerators' ability to administer the questionnaires.
- Estimating how long it takes to administer each questionnaire in the field. The IE team and the survey firm need to understand the average duration of each data collection component in order to properly plan field work. See **Duration of Interview Tracking Sheet.**
- Documenting the changes and adaptations suggested by the team.

At this stage, it is best to keep edits to the questionnaire to a minimum. The IE team and survey firm should plan at least 2 days to pilot test the questionnaire and to finalize all the logistics required before initiating field work.

## Managing Field Work

Once data collection starts, it is the responsibility of the survey firm to ensure data collection is in accordance with the research protocol, sampling plan and field work plan. The survey firm should budget for a two to three day meeting for all field team supervisors, enumerators and data entry operators to meet after field work begins. This meeting would give the team an opportunity to discuss and correct any problems related to supervision, field work organization, questionnaire format or content, and data entry issues. We suggest that the meeting occur 2-3 weeks after the initiation of field work.

Enumerators' performance also needs to be checked by field supervisors. An **Enumerator Evaluation Form** is included in the Toolkit, along with other resources such as a **Supervisor Tracking Form** that allows managing progress or a **Cash Management Sheet**.

## Reporting

Communication between the IE team and the survey firm is critical at all stages of the data collection in order to ensure that any issues are addressed as soon as possible. However, experience has shown that survey firms often have a hard time putting together reports that are truly useful for the IE team. Therefore, the Global RBF IE team has developed a standardized **Survey Progress Report** that survey firms can use to report progress in the preparation, training, data collection and data management activities. The structure of the Progress Report is as follows:

1. Date and Phase.
2. Summary of Progress.
3. Key Challenges.
4. Next Steps.
5. Status of Data Collection and Entry.
6. Status of Deliverables.
7. Status of Payments.

# Module 6

# Storing and Accessing Data

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

# Module 6. Storing and Accessing Data

| Main Recommendations and Available Tools for this Module | | | |
|---|---|---|---|
| **Recommendations** | **Critical** | **Important** | **Nice to have** |
| • The TTL should plan for and coordinate comprehensive and complete documentation of impact evaluation activities.<br>▸Include updated Concept Note, Research Protocol, Questionnaires, Training Manuals, etc.<br>▸Decide on what information needs to be removed for respondent confidentiality. | ✓ | | |
| • The Principal Investigator should prepare (a) separate ID control file(s) that establishes the link between the geographical ID codes and the field ID codes. | ✓ | | |
| • The Principal Investigator should decide on any variables that cannot be released publicly (e.g. sensitive personal information). | ✓ | | |
| • Confidential files (ID control file and other non publicly available data) should be stored in a secure location, preferably a data enclave. | | ✓ | |
| • Impact evaluation teams should allocate sufficient time for documenting and uploading the data, in order to guarantee data access continuity within the team, ease future data sharing and analysis process | | ✓ | |
| • Impact evaluation teams should refer to the Memorandum of Understanding (or other data sharing agreement) when documenting, storing and sharing the data. | | ✓ | |

| **Tools** |
|---|
| • 6.01 Data Deposit Form – IE Micro-data Catalog<br>• 6.02 Nesstar Data Storage Templates<br>• 6.03 Login to Micro-data Management Toolkit<br>• 6.04 How to Access the Data Catalog and Data |

## Module Contents

The average RBF impact evaluation is expected to cost about US$ 1.5 million, and we expect that approximately 70 to 80 percent of that budget will be spent on data collection. So data are expensive and highly valued, yet in many cases they are greatly underutilized, and in many cases this is because they are not publicly available. In addition, there are many instances where data were collected but later lost or improperly documented because computers failed or people moved to a different job. Finally, impact evaluation data often includes sensitive information such as ways to locate and/or identify respondents. Such sensitive content needs to be protected and the confidentiality of the responses preserved.

Bearing this in mind, the RBF IE team is committed to helping teams to ensure that

- Data collected with HRITF funding are safely stored and properly documented;
- Teams have access to the data in a way that meets their needs;
- Data ultimately become publicly available in accordance with the World Bank's Open Data Initiative, while respecting confidentiality.

All three objectives can be achieved using a Micro-data catalog. In this module, we explain (i) why use a Data Catalog to store data, (ii) the types of data, (iii) how to catalog the data, (iv) the features of the RBF data catalog and (v) the different kinds of data access that can be set up in the RBF data catalog.

## Why use a Data Catalog?

IE teams will need to document and store their data either in the RBF Data Catalog or in the World Bank's central data catalog. Both catalogs were set up specifically for RBF teams and follows the Dublin DDI standards for storing and documentation of data. Storing data and documentation in these data catalogs has the following advantages:

- Team members, including Principal Investigator, TTLs, IE team, and Government counterparts, can easily share country data and their corresponding documentation. Team members are not dependent on the presence of a particular team member in order to safeguard data and documentation.
- Improve data and documentation security. Data and documentation that are stored and shared using email, computer hard drives, and portable data storage devices are vulnerable to hardware malfunction, hacking and viruses. These threats are minimal in a protected data catalog.
- The catalog follows international standards on data documentation and storage, and in doing so it makes it easier to ensure that the data are labeled, accurate and anonymous.

The RBF data catalog is currently a closed access catalog which is password protected. By contrast, the World Bank's central data catalog is publicly visible and searchable from the web without a password. As

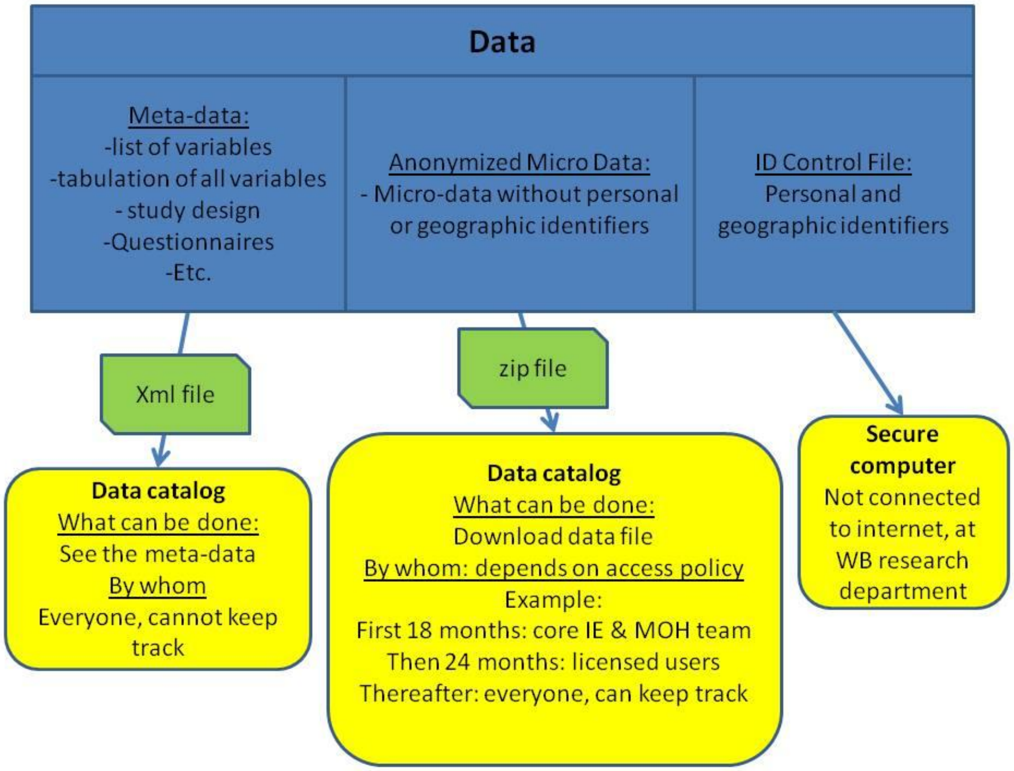
RBF data become publicly available, they would be migrated from the password protected RBF data catalog to the World Bank's central data catalog.

## Types of Data

Typically, an impact evaluation survey will produce three types of data:

- **Macro-data:** Macro-data include the list of variables, the tabulation of all variables, the study design, questionnaires, etc. In short, macro-data is the "data on the micro-data" or all of the information that is needed to be able to correctly interpret the micro-data.
- **Anonymous Micro-data:** Micro-data is the data at the level that they were observed, for example, income information on individual households, height and weight measurements on individual children, and financial data on health centers. This data should be anonymous, and should not contain the geographical codes, names or other individual identifying information. However, they should contain field IDs that allow researchers to analyze the data without identifying the units that were surveyed. In certain cases, principal investigators may decide not to release certain micro-data *publicly* if they are very sensitive in nature (eg. HIV test results).
- **ID control file:** The principal investigator must safeguard the file that establishes the link between the geographical ID codes and the field ID codes, as well as any names that were collected with the survey. Without it, (s)he will not be able to recreate the sample in the endline data collection if the survey firm must return to the same sample areas. In order to protect anonymity of the data, the ID control file should NEVER be included in the data catalog. However, it should be kept in a secure location; the World Bank has a secure server that is not connected to the internet where teams can safeguard and store their identification key files. (see also under Module 5, Defining Unique Identification Codes)

Figure 11: Managing IE Data


Data
Meta-data:
-list of variables
-tabulation of all variables
- study design
-Questionnaires
-Etc.
Anonymized Micro Data:
- Micro-data without personal or geographic identifiers
ID Control File:
Personal and geographic identifiers
Xml file
zip file
Data catalog
What can be done:
See the meta-data
By whom
Everyone, cannot keep track
Data catalog
What can be done:
Download data file
By whom: depends on access policy
Example:
First 18 months: core IE & MOH team
Then 24 months: licensed users
Thereafter: everyone, can keep track
Secure computer
Not connected to internet, at WB research department

## Steps to Cataloging Survey Data

The Micro-data Management Tool is a tool that allows cataloging data in two steps:

**Step 1: Organize the data using Nesstar software**

The first step is to process and organize the data using the Nesstar software. This software is free, and can be downloaded from: (http://nesstar.com/software/download.html). Nesstar allows the user to extract, organize and store the macro-data using international standards for data documentation, preservation, security and storage[16]. According to international standards, all final documentation should be in English; if parts of the activities are carried out in another language, the team should budget resources to translate the materials into English. The software stores two types of information:

[16] For more information, please visit http://www.ihsn.org

(i) “Documentation” stored under a Data Documentation Initiative (DDI) file, which gives basic information about the survey such as team members, dates of the survey, design of the survey, etc.; and (ii) “External resources” stored under a Dublin Core (DC) file, which describe resources shedding light on the data generation process, such as questionnaires, concept notes, research protocols, etc. Note that the actual resources are not stored within Nesstar.

Both DDI and DC files are turned into two “.xml” files that contain all of the macro-data of the study: in short, an overall description of the impact evaluation survey methodology and instruments. The whole project is also saved into a .Nesstar file, which can be reopened and modified with Nesstar.

**Step 2: Upload the data to the data catalog**

The second step is for the two .xml macro-data files and the anonimized micro-data[17] to be uploaded to the data catalog. Other documents such as questionnaires, survey methodology and research protocols should also be uploaded, because they are essential to understanding how the data were generated and how they should be interpreted. This can be done using the data catalog’s Resource Manager.

After they are uploaded the macro-data will be visible to anyone who searches the catalog. Access to the anonimized micro-data files can be restricted by using the “unavailable” or “licensed use” options in the catalog (Cf. below). As a reminder, the ID control file should never be uploaded to the data catalog.

While we understand that processing and uploading the data can be a slightly tedious process, the HNP IE team has partnered with the World Bank’s research department to obtain help when needed. Please see the contact page of the Toolkit for the email address. Teams will be asked to populate a **Data Deposit Form** to facilitate uploading of the data**.**

## The RBF Data Catalog: Features

The RBF Data Catalog includes a detailed and searchable list of all data and associated documentation that has been collected in the context of the RBF program. Please note that the website is password protected and is made accessible only to RBF teams. It can be found at: http://www.ihsn.org/apps/hritf/index.php/auth/login/?destination=

The following features are available in the catalog:

[17] In a zip file format

- **View and download all corresponding documentation**: Team Members can view their up-to-date Impact Evaluation Concept Note, Research Protocol, Final English (and local language) Questionnaires, Sampling Plan and any other documentation required to understand how the data was generated.
- **View Macro-data**: View RBF teams' macro-data, including questionnaires, summary responses, variables, research design, etc. Access to the micro-data depends on the Data Access Policy.
- **Compare Macro-data**: Compare the frequencies for the same variables across countries.
- **Track Citations**: Track the publications which utilize the datasets in the database.
- **Control Access to the Data**: Country teams can control who has access to the micro-data and over what time periods. Data access should be clearly defined by each country team's **Data Access MOU** (Cf. Module 4 for the tool and discussion).

## Access to Data

All data and documentation should be stored and documented in the RBF Data Catalog within 6 months of completion of the data collection. Data that is stored in the data catalog can be labeled as either "not accessible", "licensed use" or "public use". Data that become "licensed use" or "public use" would be migrated to the World Bank's general data catalog, since the RBF catalog is password protected and does not provide real "licensed use" or "public use" access. The World Bank's general data catalog also has "not accessible", "licensed use" and "public use" modalities.

- **Not accessible.** Access to the micro-data can be limited to the country impact evaluation team and core counterpart team. The usual arrangement is that this is will be the case for two years after the completion of data collection or six months after publication of the first report, whichever comes first.
- **Licensed use.** Within 2 years of completion of data collection or within six months after publication of the first report, whichever comes first, data should normally be available under licensed use. Under licensed use, external visitors can submit an online request to access the data, including their research topic, variables of interest, timeline and dissemination plans. The request is sent through the catalog to the Data catalog manager, who works with the country team to approve or reject requests. Once a request is approved, the data manager sends a username and password to the requester.
- **Public use.** Within 4 years of completion of data collection, the micro-data should be available for public use. Under public use, external visitors can submit an online request to access the data, including their research topic, variables of interest, timeline and dissemination plans. The request is immediately approved by the system, and the individual is provided a username and password.

2012/02/05

### Country Spotlight: Defining Data Access Policy Among Several Stakeholders
### Afghanistan Strengthening Health Activities for the Rural Poor Project

The impact evaluation of RBF is a collaborative work between the World Bank, Johns Hopkins University (JHU) and the Government of Afghanistan. JHU collected baseline data in late 2010. The intervention is still ongoing, and the follow-up survey is scheduled for February 2013. There is no explicit Memorandum of Understanding between the parties that defines access to data. However, there are contractual agreements between JHU and the government defining access to data.

After baseline data was collected, the three stakeholder team shared the data amongst themselves. Since the IE is constituted of two cross sections, there was no need to include information identifying households or health facilities to find them again in the second round within the datasets. JHU, as the survey firm, has access to the nominative data contained in the paper questionnaires.

Even though data access within the team was a smooth and obvious process to the team, it is not yet clear how the data will be made available to outsiders. According to the World Bank Open Data Initiative, the data collected and financed by the World Bank would have to become public at some point. The IE team discussed which approach they should adopt to comply with the initiative, while benefiting from a privileged access to the data in the first stages of baseline data release. The team unequivocally agreed that the data would become public. However, the timing of this public release is still being debated.

Full story available: see Country Spotlights section of the Toolkit.

# Module 7

# Analyzing Data and Disseminating Results

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

# Module 7. Analyzing Data and Disseminating Results

| Main Recommendations and Available Tools for this Module | | | |
|---|---|---|---|
| **Recommendations** | **Critical** | **Important** | **Nice to have** |
| • Data analysts should keep a record of any alteration and statistical analysis performed on the data. | ✓ | | |
| • The original data must absolutely be kept intact. Any alteration must be saved as a different dataset. | ✓ | | |
| • Prior to baseline data analysis, the data analyst should refer to international and national guidelines on how to calculate indicators. (eg. WHO) | | ✓ | |
| • The data analyst can help identify errors that occurred during baseline data collection or entry. This can then allow for adjustments in training and supervision during future rounds of data collection. | | | ✓ |
| • Data cleaning, analysis and dissemination of results take time. It helps to plan ahead in terms of manpower and funds. | | | ✓ |
| • Ex-post power calculations are a part of the internal validity checks of the impact evaluation. If need be, data analysts can recommend ways to increase power at follow-up. | | | ✓ |
| • The analysis should be developed keeping in mind the best way of ultimately disseminating results and informing policymakers. | | | ✓ |
| • Impact evaluation data are typically very rich: while analyzing the impact of RBF may be the primary goal, other analyses can be conducted to inform policymaking. | | | ✓ |

**Tools**

- 7.01 Household Baseline Report
  - *7.01a Handbook Household Baseline Report*
  - *7.01b Indicators Rwanda Household Baseline*
  - *7.01c STATA do files Rwanda Household Baseline*
  - *7.01d Ex post Power Calculations Rwanda Household Baseline*
- 7.02 Health Facility Baseline Report
  - *7.02a Suggested detailed outline of health facility baseline report*
- 7.03 Community Health Worker (CHW) Baseline Report
  - *7.03a STATA do files Rwanda CHW Baseline*
- 7.04 STATA ado file for Baseline balance table
- 7.05 WHO Anthro calculation package
- 7.06 STATA training
- 7.07 STATA Training Design Validation

## Module Contents

The ultimate goal of the IE is to evaluate the impact of the RBF intervention and share the evidence with policymakers in order to inform policy decisions. The dissemination of the results is a crucial step in the IE, yet teams often underestimate the time and financial resources required for producing quality data analysis and dissemination products. Teams should identify ahead of time qualified data analysts available following baseline, mid-term and endline surveys (See timeline in Module 3) in order to produce timely results.

Teams should also consider the capacity building aspect of impact evaluations: given the technicality of impact evaluations, many low income countries lack the capacity to fully understand and implement randomized controlled impact evaluations, as well as analyze the impact of interventions. Working in collaboration with local researchers and analysts can build the foundation for further championing of impact evaluations among policymakers, and build local capacity to conduct and analyze impact evaluations.

Several reports should be produced along the course of the IE:

- For the different units surveyed: households, health facilities, health workers, community health workers, etc.
- For various stages of the IE cycle: baseline, mid-term, endline.

This module gives IE teams an overview of the main recommendations for (i) baseline data analysis and report writing, (ii) impact analysis and (iii) dissemination. The table below briefly summarizes the main purposes of baseline, mid-term and endline reports.

**Table 16: Impact Evaluation Reports and their Functions**

| | |
|---|---|
| **Baseline Report** | • Validate IE design: conduct tests of difference in means between treatment and comparison groups to assess the balance of the sample. Conduct external and internal validity checks<br>• Produce descriptive statistics on the units surveyed<br>• Give recommendations for the implementation of follow-up survey(s) |
| **Mid-term Report** | • Issue first assessment of impact and cost-effectiveness (if enough exposure)<br>• Give recommendations for the implementation of the intervention if unexpected impacts or operational issues are detected (e.g. input shortages, delays in disbursements of incentives, etc.) |
| **Endline Report** | • Evaluate impact and cost-effectiveness<br>• Produce recommendations for the scale-up, continuation and possible improvements of the intervention. Provide general policy recommendations. |

## The Baseline Report

**Goals:** The baseline report has three main goals:

- Checking the internal and external validity of the evaluation design (see below)
- Presenting descriptive statistics on surveyed units or population in order to document sample characteristics
- Provide recommendations for future survey rounds if need be.

There may be additional objectives that need to be integrated in the content of the report, particularly at the request of the project TTL and/or Government. For example, the baseline report could look at issues such as access to care, equity, etc.

**Recommended outline:** Outlines can vary depending on the surveys and on whether the baseline report has any additional objectives beyond validation and presentation of descriptive statistics. We have found the following two outlines helpful: Rwanda Community PBF 2010 household baseline report and *Impact Evaluation in Practice* (Gertler et al. 2011, p.213).

**Figure 12: Examples of Outlines for a Baseline Report**

**Outline 1 (Rwanda Community PBF 2010)**

1. Overview<br>
  1.1 Introduction<br>
  1.2 Project Background<br>
  1.3 Project Components<br>
  1.4 Objectives of the Study<br>
2. Methodology<br>
  2.1 Randomization<br>
  2.2 Study Design<br>
  2.3 Sample Size and Strategy<br>
  2.4 Variables for Data Analysis<br>
  2.5 Instruments for Data Collection and Data Quality Insurance<br>
  2.6 Data storage, Management & Access Policy<br>
3. Sample Representativeness and External Validity<br>
  3.1 Geographic Representativeness<br>
  3.2 Comparison between Baseline Study and Country Population<br>
4. Findings (organized by section of the questionnaire or relevant topics)<br>
5. Internal Validity of the Study<br>
  5.1 Ex-post power calculations<br>
  5.2 Threats to internal validity<br>
  5.3 Sample Balance: Summary of tests Results

**Outline 2 (Gertler et al 2011)**

1. Introduction<br>
2. Description of the Intervention (Benefits, Eligibility Rules, etc.)<br>
3. Objectives of the Evaluation<br>
  3.1 Hypotheses, theory of changes, results chain<br>
  3.2 Policy questions<br>
  3.3 Key outcome indicators<br>
4. Evaluation Design<br>
  4.1 Original design<br>
  4.2 Actual program participants and nonparticipants<br>
5. Sampling and Data<br>
  5.1 Sampling strategy<br>
  5.2 Power calculations<br>
  5.3 Data collected<br>
6. Validation of Evaluation Design<br>
7. Comprehensive Descriptive Statistics<br>
8. Conclusion and Recommendations for Implementation

## Data Cleaning

*To begin, please note the Principal Investigator should not share non anonymous data with other data analysts. The ID control file linking geographical and any other identification with field IDs should be stored safely and data analysts should only be given the data with field IDs (see Module 6).*

The IE team will need to clean raw data files prior to conducting data analysis and producing dissemination materials. Data cleaning is the process of identifying inconsistencies and inaccuracies in the raw data and reducing these errors based on defined assumptions. During the data cleaning process, data analysts should document any recurring error, along with the solutions reached. This should be passed on to the IE implementation team in order to avoid incorrect questionnaires or data entries during follow-up survey rounds. Data cleaning is time consuming and IE country teams should allow for sufficient time before starting the analysis and writing the baseline report.

*Please note: The data cleaning process is directly linked to the quality of data collection and entry. Data errors are introduced due to a lack of training, supervision and quality assurance mechanisms in place during the preparation and implementation of the impact evaluation. When strong data quality assurance mechanisms are in place throughout the preparation and implementation process, fewer errors will be introduced in the raw data files, thereby reducing the resources required to clean the data.*

## Creating Variables for Analysis

Data analysts can create two types of variables to be included in the analysis:

- **Indicators of interest on inputs, process, outputs and outcomes**: These are the most important variables, since they will allow analysts to evaluate the efficiency, cost-effectiveness and impact of the RBF intervention after follow-up survey(s). These variables should be defined according to international and national definitions (please refer to Module 3 on indicators). IE country teams should also consult with national health and/or RBF experts to identify other country-specific indicators required. They should pay particular attention to including the appropriate population in their calculation, with correct age range and gender according to guidelines and definitions.
- **Covariate variables**: these do not assess the impact of the program, but document the main characteristics and behavior of the population included in the survey.

Raw data will consist of continuous variables already present in the dataset (e.g. respondents' age or income) or dichotomous 0-1 variables. For some variables, the data analyst(s) will need to convert responses to dichotomous 0-1 variables for the analysis.

- For example, household survey data contains the following question: "What was NAME mainly suffering from?". The responses are coded from 01 to 20, each number representing a different disease. In this case, the analyst will need to create twenty dichotomous 0-1 variables that take value 1 if the individual suffered from this specific disease and 0 if not. The data analyst will then be able to calculate the mean of each variable, which will correspond to the percentage of respondents who mainly suffered from that disease.

*It is crucial that the data analyst(s) document the data cleaning and variable construction process, especially through annotated **Stata© do files**, for two main reasons: (i) replicating results and (ii) improving efficiency for future analysis. The IE country team should be able to replicate results and provide definitions of assumptions made and variables constructed. Additionally, since survey instruments and Data Entry Programs will be similar at baseline and follow-up, work on the baseline variable construction and analysis bears the fixed costs of writing the analysis code, which can then be used for future rounds of data.*

## Validating the Evaluation Design

The baseline report must assess the external and internal validity of the IE.

### External Validity

The assessment of external validity should answer the following question: To what extent can the findings of the study be generalized to a broader population? The sampling strategy defines who was targeted by the survey and how this population was selected, but the analysis must answer how comparable it is with national or sub-national populations. The assessment of external validity is mostly based on sampling design, completed by results from actual sampling in the field at baseline. Results from the IE survey can be compared to other available data (e.g. Demographic and Health Surveys) to assess the comparability of the data to other measures.

### Internal Validity

The assessment of internal validity is one of the main goals of the baseline report. Most of the internal validity checks consist in comparing the theoretical design of the IE to the result from field work. The validity checks should answer the following questions:

- Is there sufficient power in the experiment? Is the sample size for each unit (facilities, health workers, households, beneficiaries) sufficient?
- Are the treatment and comparison groups balanced? Are means of the treatment and comparison groups balanced?

- Is there any other threat to internal validity? Does any pre-existing characteristic or intervention threaten the internal validity of the study?

### *Is there sufficient power in the experiment?*

Once the team has collected baseline data, they can use the baseline values of key variables, the chosen sample size, the desired level of power, and the desired level of statistical significance to determine the minimum detectable effect for a given outcome of interest, that is, the smallest effect that the experiment will be able to detect. If this smallest detectable effect is smaller than the expected effect of the intervention, then the experiment will likely detect the expected effect of the intervention (e.g. have high statistical power) at the chosen level of statistical significance.

Ex-post power calculations with baseline data are useful for policy dialogue during the course of the intervention. For example, the size of the minimum detectable effect may affect how long the evaluation will need to last: if the minimum detectable treatment effect is high, then the intervention will need to have a large impact for this impact to be detected in the impact evaluation. For the impact to be high, one may need to let the intervention run for a longer time in the treatment group while maintaining the comparison group, assuming that the intervention effect will keep growing over time. Therefore, a higher minimum detectable treatment effect will mean that the evaluation will need to last longer.

Power calculations can be run in software such as Optimal Design[18], made freely available by the William T. Grant Foundation (http://www.wtgrantfoundation.org). The software and its documentation are available online from the Foundation's website or at:

http://sitemaker.umich.edu/group-based/optimal_design_software

Code for running power calculations for cluster-randomized trials in R (http://www.r-project.org/), an open source language and environment for statistical computing and graphics, is also made freely available from the authors of the Optimal Design software.

[18] Raudenbush, S. W., et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from www.wtgrantfoundation.org or from sitemaker.umich.edu/group-based.

An example of ex-post **Power calculations** is available in this Toolkit in Module 3, based on the results from baseline data collected on the Rwanda Community PBF intervention. Users can also refer to Module 3 for references on power calculations.

### *Are baseline characteristics balanced?*

The baseline report should contain statistical tests of the difference in characteristics between treatment and comparison groups. When there IE has more than one treatment arm, we recommend the use of the following two statistical tests:

- General F-test of difference in means of key indicators and covariate variables across all study arms

- T-tests of difference in means of key indicators and covariate variables from one study arm to another, with an emphasis on the difference between each treatment group and the comparison group.

With only one treatment arm, T-tests are enough.

The baseline report should contain an overall assessment of the percentage of unbalanced indicators and covariates, distinguishing results from the F-tests and T-tests. The results of T- and F-tests obtained for each section of the questionnaire should also be discussed throughout the report as descriptive statistics are presented.

### *Are there any other threats to the internal validity of the study?*

The baseline report should evaluate other potential threats to the internal validity of the study:

- Do formal or informal interventions that already exist within the country imitate, or conversely counteract the intervention? Are they likely to produce unintended responses to the RBF intervention in treatment or comparison groups?
- Did respondents mention already receiving rewards that may influence their behavior in the same or opposite direction as the RBF intervention? If yes, is the magnitude of those incentives sufficient to threaten RBF incentives and are those incentives different between treatment and comparison groups?

- Compliance to the sample design during field work: data analysts should compare theoretical and actual number of units included in the survey (geographic units, households, health facilities, etc.) and report on non-response rates. They should check whether noncompliance or non-response are different between treatment and comparison groups. If noncompliance or a high non-response rate is detected, data analysts should provide recommendations to the survey team to improve follow-up surveys.

2012/03/27

**Country Spotlight: Validating IE Design and Engaging Policy Discussions with Ex-Post Power Calculations**
**Rwanda Community Performance-Based Financing Project**

As part of the IE design validation, the team conducted ex-post power calculations on the baseline data collected. The goal was to ensure the actual sample sizes and sampling from the field allowed to detect a reasonable effect size of the intervention for a given power and a given confidence level, and assess to what degree the results from baseline matched the results expected at the design stage. The research team used a set of core outcome indicators calculated from the household baseline data. The 14 binary indicators were selected based on their likeliness to be impacted by the program:

- ANC coverage (1+ visit),
- Timely ANC (prior to 4th month of pregnancy),
- ANC coverage (4+ visits),
- TT2 coverage during pregnancy,
- 90-day iron supplementation during pregnancy,
- Skilled delivery,
- Delivery in a formal health facility,
- Low-birth-weight newborns,
- Timely initiation of breastfeeding,
- Exclusive breastfeeding (0-6 months),
- Timely PNC visit in a formal health facility,
- Postnatal supplementation with vitamin A,
- Modern contraceptive prevalence,
- Unmet need for Family Planning

The statistical model was defined based on the design of the study: a blocked 2-level cluster randomized trial, where sectors were blocked by poverty level and the data clustered at the sector level. Treatment was allocated at the sector level. Type 1 error rate was defined as 0.05 and desired power as 80%. For the set of outcomes studied, the minimum detectable effect ranged from 0.06 to 0.12. Since ex-ante power calculations were based on a minimum detectable effect of 0.2, the team concluded ex-post minimum detectable effect sizes were within range and could be reached by the intervention.

The team started two important policy discussions based on those results.

- For indicators that were already high, the results of the power calculations raised the question of the ability to reach those minimum detectable effects. For example skilled delivery had to increase from 89% at baseline to 96% at endline for the increase to be detected by the IE. The team assessed how existing non-Results Based Financing Community Health Workers packages, defined to improve maternal health, could contribute in impacting those harder to reach indicators.
- The magnitude of the minimum detectable effect sizes was confronted to the relatively slow progress in increasing key indicators showed by monitoring data. As a result, the team concluded the duration of the experiment had to be extended for those minimum effects to be produced - and hence detected by the IE. The Government decided to maintain treatment and comparison groups until January 2013, as opposed to the initial June 2012 planned.

Full story available: see Country Spotlights section of the Toolkit.

## Descriptive Statistics

**Conducting the analysis and breaking it down by category:** The majority of the baseline report is typically devoted to producing statistics on the characteristics of the surveyed units. These statistics are drawn from core indicators of interest on input, process, output and outcome, as well as covariate variables. The report should present estimates of the means of these variables for:

- The complete sample interviewed in the questionnaire section

- The sample interviewed in the questionnaire section in each study arm

Means can be generated according to specific demographic or geographic criteria: they can be presented for urban versus rural households, health facilities or CHWs; for male versus female respondents; for different age categories; for specific individuals such as the head of household, children under 1 year old, children under 5, etc.; for governmental versus non-governmental health facilities, etc. Depending on the policy objective and whether the statistics will be compared with other surveys, a break-down of statistics by category may be useful.

**Presenting descriptive statistics:** The descriptive statistics obtained can be presented in the main body of the report or in an annex to the report. The Toolkit includes **Stata© code** to produce tables and graphs that can be used to illustrate key characteristics of the respondents. Graphs are especially useful when the team wants to disseminate results and make them accessible to a broad audience and policymakers.

## Impact Analysis

The impact and resulting cost-effectiveness of the RBF intervention can only be assessed once midline and/or endline surveys are completed.

A midline survey can provide early analysis of the impact of the program, in the sense that it can evaluate changes in inputs, process, outputs, and outcomes after a limited amount of time. However depending on the amount of time that has passed, there may not have been enough time for outcomes to have changed. Midline results can be useful in that they also provide recommendations for adjusting the intervention if unexpected changes are observed, such as unexpected negative impacts of the intervention, inputs shortages, inaccurate distribution of incentives, costs escalation, and noncompliance with treatment and comparison group assignment. In the last case, the impact analysis method should be re-determined in order to account for the fact that individuals eligible to the intervention may not have been treated, or that individuals ineligible to the intervention may have benefited from it.

### After the Endline Survey: Determining the Impact of the Program

Impact analysis should be conducted according to the identification strategy that was laid out in the IE design paper. *Impact Evaluation in Practice* (Gertler et al. 2011) provides an in-depth discussion on how impacts may be assessed depending on treatment and comparison group assignment. As with baseline data, endline impact analysis should be performed on clean data once the appropriate indicators of interest and covariates have been created. If the intervention includes more than one treatment group, then the impact of the intervention should be evaluated in each treatment group with regard to the comparison group. Treatment groups should also be compared to each other in order to identify the best RBF strategy.

Other dimensions may be considered in the impact analysis. In particular, the distributional impacts of RBF and whether it benefits the poor is a key aspect teams can report on. Equity analysis can inform national policy decisions and feed into the global evidence on whether RBF is a strategic instrument to improve health systems while preserving or improving access to care for the poor.

*We recommend that IE teams use survey instruments and data entry programs that are similar between baseline and endline surveys: when instruments are similar, endline data cleaning and analysis will be similar to baseline data cleaning and analysis. This will save significant time and effort at the time of writing up the results.*

### The Endline Impact Report

The main results of the impact evaluation will most likely be written up as a paper by a team led by the Principal Investigator. However, in many cases it is more helpful for the Government to have a more comprehensive report that outlines all of the results from the evaluation, including those that may not be very appealing in a publication. Therefore we recommend that the IE team also prepare a comprehensive endline/impact report, in addition to any papers for publication. In addition, the team should think of the best way to present and disseminate results to policymakers: highlighting results and providing policy recommendations should be the ultimate goals of the report. The table below proposes an outline for the endline report.

### Figure 13: Example of an Outline for an Endline Impact Report

**Impact Report Outline**

1. Overview<br>
    1.1 Introduction<br>
    1.2 Project Background<br>
    1.3 Project Components<br>
    1.4 Objectives of the Study<br>
2. Methodology<br>
    2.1 Randomization<br>
    2.2 Study Design<br>
    2.3 Sample Size and Strategy<br>
    2.4 Variables for Data Analysis<br>
    2.5 Instruments for Data Collection and Data Quality Insurance<br>
    2.6 Data storage, Management & Access Policy<br>
3. External and Internal Validity<br>
    3.1 External validity<br>
    3.2 Internal validity<br>
4. Limitations of the study<br>
5 Key descriptive statistics by study arm<br>
6. Findings<br>
    6.1 Impact analysis (organized by input, process, output, outcome, or impact indicators and comparing treatment groups)<br>
    6.2 Cost effectiveness (organized by input, process, output, outcome, or impact indicators and comparing treatment groups)<br>
7. Aspect of interest of RBF – To be defined, e.g. Distributional impacts of RBF<br>
8. Recommendations and Policy options

## Dissemination

Timely and appropriate dissemination is key to ensure that the impact evaluation will have an impact on policy. We recommend the following means of dissemination:

- **Written document**: Baseline and endline/impact reports should be shared with the Government, Bank management and with other key stakeholders before being sent for publication. The World Bank's Research Working Paper Series has long published preliminary results from impact evaluation. It is important to keep in mind though that a number of medical journals (such as the Lancet and the New England Journal of Medicine) may reject a paper if the results are contained in another (prior) report that is available online, including working papers. Please check the Journal policies before making results available online.
- **Country-based presentation and dissemination workshops**: It can be useful to disseminate results to national and local stakeholders in a workshop format. Local stakeholders can help interpret unexpected results, and a broader audience can be reached nationally.

- **Policy briefs:** Executive summaries of the program design, evaluation design and results can be useful for policy dialogue in other countries, if they are available in a non-technical, easily accessible way. The “En Breve” and “HD Brief” series are both useful instruments for this type of dissemination.
- **International workshops and events:** The compilation of national evidence will ultimately allow assessing whether RBF is an efficient and cost-effective mechanism to improve health systems and deliver services. Sharing results across countries is also extremely valuable for countries that are less ahead in the impact evaluation and project cycle.

We highly recommend involving local researchers and analysts in the dissemination process to facilitate future evidence-based policy dialogue.

*Please note: The IE team should plan in advance the type of dissemination activities required in-country and in headquarters in order to ensure these activities are included in the IE budget (see Module 3).*

2012/04/03

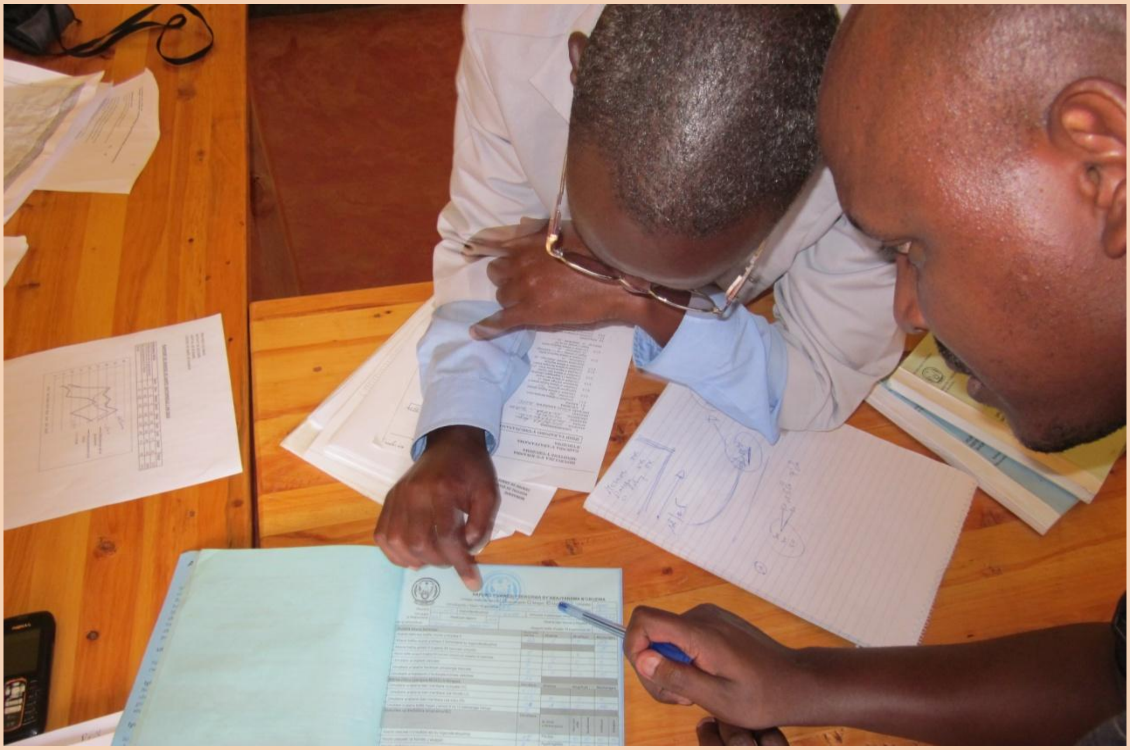
### Country Spotlight: Endline and Baseline Results Dissemination and Policy Dialogue
### Rwanda Health Facility and Community Performance-Based Financing Projects

In September 2011, the Government of Rwanda held a three-day workshop on the health facility PBF and the CPBF programs. The goal of the workshop was to disseminate available results on both the health facility PBF and the CPBF programs, so as to foster policy dialogue among community level, district level and central level representatives. Ultimately, central level policy makers wanted to identify implementation issues and bottlenecks, and design solutions in collaboration with sub-national level representatives. Additionally, the Ministry of Health wanted to disseminate results from both the PBF and the CPBF programs in order to use the lessons learned from the first PBF program for the benefit of the implementation of the second CPBF program. Finally, the workshop fostered the participation of the students from the local Research partner (National School of Public Health): students presented results from the data analysis of health-related interventions in the country.

Results disseminated on the health facility PBF program were based on a publication of endline results on the impact of the program on child health outcomes in a peer reviewed journal (Basinga et al., Lancet 2011). The results showed a positive impact of the program on various child health outcomes, and helped supporting the advocacy of the Government for PBF, among local stakeholders, donors, and the general public with the presence of the press.

The more recent and ongoing CPBF program was the main focus of the workshop. The first results presented on the program were based on the analysis of the baseline data collected within the IE. The team used the friendliest vehicles of results for the audience, such as simple descriptive statistics tables and mostly graphs. Participants in the workshop were invited to comment and question the results as often as possible. Group sessions were organized, where participants could use the results of the baseline (including tables of summary statistics on relevant topics) to reflect on a specific aspect of program implementation (e.g. user fees, monitoring and verification, etc.). They identified issues in the implementation that were reflected in the data, or on the contrary spotted discrepancies between the data and their own experience in the field, and came up with potential solutions to these issues to be designed and implemented throughout the course of the program. Participants then gave feedback to the central level by presenting their analysis and proposed solutions to the audience.

Full story available: see Country Spotlights section of the Toolkit.



# Module 8

# Monitoring and Documenting RBF Programs

Impact Evaluation Toolkit
Measuring the Impact of Results-Based Financing on Maternal and Child Health
Christel Vermeersch, Elisa Rothenbühler, Jennifer Renee Sturdy

**www.worldbank.org/health/impactevaluationtoolkit**

# Module 8. Monitoring and Documenting RBF Programs

| Main Recommendations and Available Tools for this Module | | | |
|---|---|---|---|
| **Recommendations** | **Critical** | **Important** | **Nice to have** |
| • Monitoring and documenting project activities are a crucial complement to the impact evaluation because they provide information on the actual interventions on the ground, and therefore, on the intervention that is being evaluated. | | ✓ | |
| • Impact evaluation teams will want the program to identify two major risks to the impact evaluation: (1) compensation of the comparison group through an alternative intervention or program; and (2) imitation of the treatment by the comparison group. | | ✓ | |

**Tools**

- 8.01 Monitoring Indicators Rwanda Example
- 8.02 Field Supervision Visit Rwanda Templates

## Module Contents

Impact evaluation is one element of a broad range of complementary methods that support evidence-based policy, which includes monitoring and evaluation. A general description of monitoring systems can be found in Görgens and Kusek (2009). In this module, we will focus on those aspects of monitoring that are directly relevant to impact evaluations of RBF projects, i.e. (i) Monitoring RBF interventions, (ii) Adherence to evaluation design and how to minimize risks to internal validity during operations, (iii) Using data sources that are complementary to the IE.

## Monitoring RBF Interventions

In this module, we will use the following definition of monitoring: "The systematic collection of information on a program's inputs, activities, and outputs, as well as the program's context and other key characteristics" (Centers for Disease Control and Prevention, 2008). This description of a programs mechanics and function over time is also commonly referred to as "process evaluation." Monitoring or process evaluation should be an integral component of the overall evaluation of any RBF scheme.

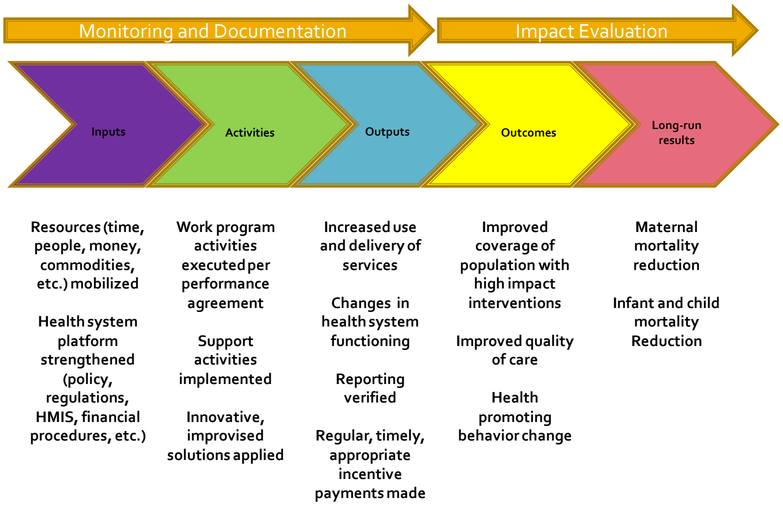
An effective monitoring system involves many steps:

- Establishing **indicators of program inputs, activities, and outputs**
- Setting up systems to collect information relating to these indicators
- Collecting and recording the information
- Analyzing the information
- Using the information to inform day-to-day management

**Purpose of Monitoring and documentation.** Continuous monitoring and documentation of RBF scheme implementation serves multiple purposes. First, monitoring identifies operational and other problems that can be addressed by corrective actions early in implementation. Second, monitoring serves to keep government officials and development partners regularly informed of the progress toward the interventions objectives. Third, information from monitoring and documentation will help IE teams better interpret and understand their evaluation findings, particularly if the intended effects are not achieved. Finally, by providing rich description of the mechanics of RBF implementation—inputs, activities and outputs—and the context in which this unfolds, monitoring helps answer the question of *why* such interventions succeed or fail. This information is critical to assessing the extent in which the interventions can be replicated in different settings and to responding to requests for detailed information about how to implement them.

**Core Monitoring and Documentation Questions.** A large amount of information could be collected through monitoring. It is important, therefore, to be strategic about which information to collect. Designers of RBF schemes should try to distinguish between essential and complementary information to guide the choice of indicators, methods to be used, and budget decisions. We propose a practical strategy that is built to answer three core questions. Complementary areas of investigation that are commonly included within the broad scope of monitoring, or which find themselves at the intersection of monitoring and evaluation for RBF, are described later.

Figure 1. A common Monitoring and Evaluation Framework for RBF


Monitoring and Documentation
Impact Evaluation
Inputs
Activities
Outputs
Outcomes
Long-run results
Resources (time, people, money, commodities, etc.) mobilized
Health system platform strengthened (policy, regulations, HMIS, financial procedures, etc.)
Work program activities executed per performance agreement
Support activities implemented
Innovative, improvised solutions applied
Increased use and delivery of services
Changes in health system functioning
Reporting verified
Regular, timely, appropriate incentive payments made
Improved coverage of population with high impact interventions
Improved quality of care
Health promoting behavior change
Maternal mortality reduction
Infant and child mortality Reduction

***Q1: [Inputs] What inputs were used to secure implementation of RBF, including human and financial resources, policies, and procedures?***

The monitoring and documentation strategy should be sufficiently robust to capture the nature, amount, and timeliness of the various resources that are mobilized by the RBF intervention, as well as the kinds of decisions and actions taken to facilitate implementation, including, if possible, their associated costs. The following are examples of inputs that may need to be monitored and documented:

- Up-front investments in the form of inputs: Although "paying for results" is at the center of RBF interventions, facilities will require some up-front investment in the form of inputs.
- Modification of existing laws, regulations, and/or health policies and procedures.
- Signing of performance agreements, quasi-contracts, and contracts (Loevinsohn, 2008).
- Upgrading of the Health Management Information System (HMIS)
- Revisions and upgrades to the financial management procedures
- Opening of bank accounts by health facilities.

***Q2. [Activities] What activities were undertaken under the RBF intervention? What were the facilitating and constraining factors encountered in executing these activities?***

Implementing RBF will require a range of activities to take place, both at the central level and at the local level. The monitoring strategy should collect information on whether, and to what extent these intended activities, which are often summarized in work plans with timelines, have been implemented as planned. It should also document any facilitating and constraining factors. Here are some examples of activities that teams would want to monitor:

- Training of health workers on what to expect from their participation in the RBF scheme, including potential rewards and sanctions;
- Additional supervision in the health centers;
- Education and communication activities, as well as outreach efforts that are carried out in communities and households in order to increase demand for services;
- Transport subsidies are provided to women to help them overcome obstacles to accessing services;
- Technical assistance provided, from both internal and external sources, at the national and sub-national level.
- Furthermore, RBF schemes can generate innovative, often improvised solutions to obstacles that impede service provision and use. These actions may be simple, low-cost efforts undertaken at the point of service delivery or in households and communities. Since these obstacles are unpredictable and it is impossible to specify their solutions in performance agreements or contracts, the IE and project teams will need to rely on monitoring documentation to understand exactly how the RBF scheme resolved obstacles at the local level.

***Q3. [Outputs] To what extent were the services linked to performance targets delivered and used? To what extent was the accuracy of reporting verified? Were the financial or non-financial incentives provided and received as planned?***

Outputs are the direct products of the RBF activities. The credibility of RBF interventions depends crucially upon (i) the availability of reliable and valid information of services provided and used (i.e., on the outputs), as specified in performance agreements or contracts; and (ii) whether timely disbursement of payments were made to the right providers and/or beneficiaries, for the right reason, at the right time. To be able to explain why an RBF scheme worked or didn't work, the IE team will need this crucial information. Factors that may have enhanced or impeded the provision or usage of services should also be documented. The following are a number of outputs that teams may want to monitor:

- Delivery of targeted services or achievement of targets for pre-determined output indicators as specified by performance agreements or contracts. These outputs may include a range of nutritional, child health, maternal and newborn service indicators, particularly in schemes that are focusing on MDGs 4 and 5.
- Explicit, expected changes in health system functioning.
- Frequency, timeliness and appropriateness of payments to beneficiaries, whether providers or households. When payments are not made in a timely fashion or in the amount intended, the credibility of the RBF scheme, which is grounded in the intimate link between performance and incentives, may be at risk.
- Any sanctions, such as withholding a portion of payments, when achievements are not reached at the level intended.

- Any sanctions for cases of “gaming” or fraud, such as removing providers from the scheme or suspending beneficiaries.

A summary of each step in the monitoring process as reflected in the three core questions, and the relationship between monitoring and evaluation, are presented in Figure 1.

## Monitoring and Documentation in the Context of Impact Evaluation

In a prospective impact evaluation, the evaluators design the evaluation before the start of program implementation, when the program itself is still at the design stage. As the program gets rolled out, the implementers may make changes to the original design of the program. A sound monitoring and documentation of these changes will inform evaluators of what changes have occurred and when, and allow them to decide whether they need to account for the changes in the analysis. If revisions to program design are significant, such as an expanded beneficiary age range or a revised incentive structure, then evaluators may need to make significant changes to the evaluation methodology.

As discussed in Module 3, the internal validity of the impact evaluation rests on the ability to compare the treatment and comparison groups without any differences aside from the RBF intervention. We discussed in Module 5 how the internal validity of the impact evaluation design could be threatened by certain data collection practices. The internal validity of the evaluation design can also be threatened during the implementation of the RBF intervention, and mitigating these threats requires sufficient monitoring and documentation, and intense in-country collaboration with the Government and other partners. We discuss two common examples:

### Risk #1: Compensation to Comparison Group

Although there may be complete agreement between the project team and Government counterparts on the evaluation design, it is possible that the Government, or other partners, wish to compensate the comparison group (at the district or health facility levels) for being excluded from the RBF pilot. For this reason, other programs targeting maternal and child health may be specifically introduced in the comparison group. The following is an example of how this affects internal validity:

- Say there is agreement between the project team and the Government on the evaluation design; the treatment and comparison groups are clearly defined and the comparison group will wait 24 months to start receiving the RBF intervention. The baseline data collection is completed successfully and the baseline data show that the groups have similar means on key outcome indicators. At month 12, the Government receives pressure from representatives in the comparison group districts to compensate them for the fact that they are not included in the RBF pilot. Another development partner is planning a large-scale, non-RBF maternal and child health project, and the Government decides to allow the partner to scale the project up immediately in the comparison group so as to alleviate some of the political pressure. This is not discussed with the RBF project or IE teams as it is thought this does not violate the agreements made on the RBF project and evaluation design. Endline data are collected at month 24 as agreed, and shows large improvements in MCH outcomes across the evaluation sample. However, the results of the impact analysis show no difference between the treatment and comparison groups, suggesting that the RBF intervention had no impact on maternal and child health outcomes. In reality, both the RBF intervention and the other MCH program probably improved MCH outcomes; however, it is impossible to estimate the true impact of the RBF program since the comparison group is no longer an accurate estimate of the counterfactual. Therefore, the impact evaluation ends up underestimating the impact of the RBF intervention. With no positive results to show, the Government now struggles to defend its investment in the program and experiences a hard time finding donors willing to support it.

So does this mean that the Government cannot roll out ANY other program on MCH during the impact evaluation? Not really. Additional programs can be rolled as long as they benefit the treatment and comparison groups from the evaluation equally.

- In the example above, the Government could have allowed the other donor to roll out the other MCH program as long as it did this in all facilities. If that were not possible, it would need to ensure that the same number of treatment and comparison facilities would benefit from the program. In technical language, the Government should have ensured that new programs were rolled out "orthogonally to the assignment to treatment and comparison groups for the RBF evaluation".

### Risk #2: Imitation of Treatment by the Comparison Group

In some circumstances, it is possible that the districts/health facilities assigned to the comparison group never learn about the RBF intervention and continue to operate in a business-as-usual fashion. However, given the accessibility to information and in some cases the very public promotion of innovative programs like RBF, individuals working in the comparison group may be entrepreneurial and motivated enough to implement their own form of RBF. While this may be difficult for some complex health system reforms like PBF, it is possible for a demand-side incentive strategy. The following is an example of how this affects internal validity:

- Say there is agreement between the project team and the Government on the evaluation design; the treatment and comparison groups are clearly defined and the comparison group will wait 24 months to start receiving the RBF intervention. The baseline data collection is completed successfully and the baseline data show that the groups have similar means on key outcome indicators. However, at month 12, there is a national conference on maternal and child health attended by health facility managers from all over the country. At this conference, health facility managers from the comparison districts learn about a new strategy to increase utilization of prenatal, delivery and postnatal services which provides in-kind incentive packages to women to attend the facility for these services. The health facility managers from the comparison group recognize that this may be a very useful method for them to increase utilization of their own facilities. Using their input-based budget from the Ministry of Finance, some health facilities located in the comparison districts begin offering women these same incentive packages. Endline data is collected at month 24 as agreed.  The results of the impact analysis show that there were improvements in MCH outcomes in both the treatment and comparison groups, and that the improvement was slightly larger in the treatment group than in the comparison, suggesting that RBF had only a small impact on MCH outcomes. Unfortunately, because the RBF intervention (in this case a demand-side incentive strategy) was also (partially) implemented in a sub-sample of comparison districts, comparing the outcomes between the treatment and comparison facilities at endline will lead us to underestimate the true impact of the RBF intervention.

## Complementary Data Sources

We estimated that the cost of survey data collection represents around 80% of the total cost of an impact evaluation, and therefore complementary data sources can therefore serve as important means of saving financial resources and ensuring timely results.[19] The following are examples of complementary data sources that may be useful in the context of RBF schemes:

- Regular surveys (census, Demographic and Health Survey (DHS))
- Regular monitoring (annual achievement tests)
- Administrative records (health records, school enrollment)

[19] Before using complementary data sources for impact evaluation, we recommend that IE teams assess the quality of these sources.

2012/03/29

**Country Spotlight: Using administrative data to monitor the impact of RBF and advocate for local capacity building**

**Argentina Plan Nacer**

The national IE [of Plan Nacer] was rolled out in two phases with nine provinces in Phase I and the remaining provinces in Phase II. In 2007, the program was expanded to cover the rest of the country which in effect contaminated the comparison group for the phase I provinces, while no follow-up data had been collected yet due to a long survey firm contracting process. (…) The World Bank team sought alternative strategies for generating interim results in order to further support the policy dialogue: (…) administrative data. In two provinces, Misiones and Tucuman, data sources were found that included all health services provided at public sector Primary Care Centers (PCCs) and at maternity clinics (…) for the 2005-2008 period in Tucumán and the 2007-2009 period in Misiones. An important feature of these databases is that records include the complete universe of care provided, both for Plan Nacer participants and non-participants. All told, databases cover more than 2,750,000 services during the period of analysis. The vast amount and high quality of the data offer excellent statistical potential in terms of identifying impacts, even with respect to very low incidence indicators such as neonatal mortality.

The data from the two pilot provinces were analyzed in half a year, which allowed the team to generate intermediate results and organize a dissemination event in early 2010. The positive results served as an important endorsement of the program and raised the interest of the national and provincial Governments to explore the use of monitoring systems to further track the intervention and generate additional results.

Full story available: see Country Spotlights section of the Toolkit.

## References

Basinga P., P.J. Gertler, A. Binagwaho, A. Soucat, J. Sturdy and C. Vermeersch. 2011. “Effect on maternal and child health services in Rwanda of payment to primary health care providers for performance: an impact evaluation”. *The Lancet,* Vol. 377(9775), pp. 1421-28.

Bessinger, R. E. and J. T. Bertrand.2001. "Monitoring quality of care in family planning programs: A comparison of observations and client exit interviews". *International Family Planning Perspectives*, Vol. 27(2), pp. 63-70.

Bloom, E., I. Bhushan, D. Clingingsmith, R. Hong, E. King, M. Kremer, B. Loevinsohn and J.B. Schwartz. 2006. Contracting for Health: Evidence from Cambodia. Unpublished manuscript.

Das, J. and J. Hammer. 2007. "Location, location, location: residence, wealth, and the quality of medical care in Delhi, India." *Health Affairs,* Vol. 26(3), pp. w338-w351.

Das, J. and J. Hammer. 2004. “Strained Mercy: The Quality of Medical Care in Delhi”. *Policy Research Working Paper,* No. 3228, The World Bank, Washington, D.C.

Donabedian, A. 1980. “The definitions of quality and approaches to its assessment”. *Health Administration Press, School of Public Health, The University of Michigan*, Ann Arbor.

Donabedian, A. 1988. "The quality of care. How can it be assessed?" *JAMA*, Vol. 260, pp. 1743-8.

Doran T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland. 2006. “Pay-for-performance programs in family practices in the United Kingdom.” *New England Journal of Medicine*, Vol. 355, pp. 375-384.

Fiszbein, A. and N. Schady. 2009. “Conditional Cash Transfers: Reducing Present and Future Poverty”. The World Bank, Washington, DC.

Fleetcroft R., N. Steel, R. Cookson, S. Walker and A. Howe. 2012. “*Incentive payments are not related to expected health gain in the pay for performance scheme for UK primary care: cross-sectional analysis*.”, *BMC Health Serv Res.*, Vol. 16, pp. 12-94.

Franco, L. M., C. Daly, D. Chilongozi and G. Dallabetta. 1997. "Quality of case management of sexually transmitted diseases: Comparison of the methods for assessing the performance of providers." *Bulletin of World Health Organization,* Vol. 75(6), pp.523-32.

Franco, L. M., C. Franco, M. Kumwenda and W. Nkhoma, W. 2002. "Methods for assessing quality of provider performance in developing countries." *International Journal of Quality in Health Care*. Vol. 14, pp. 17-24.

Gertler P.J., S. Martinez, P. Premand, L. Rawlings, and C. Vermeersch. 2011. "Impact Evaluation in Practice." The World Bank, Washington, D.C.

Gertler P.J. and C. Vermeersch. 2012. " Using Performance Inventives to Improve Health Outcomes," *Policy Research Working Paper Series*, Forthcoming, The World Bank, Washington, D.C.

Glickman, S., F. Ou, E. DeLong, M. Roe, B. Lytle, J. Mulgund, J. Rumsfeld, B. Gibler, M. Ohman, K. Schulman, E. Peterson. 2007. "Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction." *JAMA*, Vol. 297(21), pp. 2373-2380.

Görgens, M. and J.Z.Kusek. 2009. "Making Monitoring and Evaluation Systems Work." *The World Bank, Washington, DC*.

Hermida, J., D. D. Nicholas and S.N. Blumenfeld. 1999. "Comparative validity of three methods for assessment of the quality of primary health care." *International Journal of Quality in Health Care,* Vol. 11(5), pp. 429-33.

Huntington, D., H. Zaky, S. Shawky, F. Fattah, and E. El-Hadary. 2010. "Impact of a Service Provider Incentive Payment Scheme on Quality of Reproductive and Child-health Services in Egypt," *Journal of Health, Population and Nutrition*, Vol. 28(3), pp. 273-280.

Jha, A., K. Joynt, J. Orav and A. Epstien. 2012. "The Long-Term Effect of Premier Pay for Performance on Patient Outcomes," New England Journal of Medicine, Vol. 366(17), pp. 1606-1615.

Leonard, K. L. and N. C. Masatu. 2005. "The use of direct clinician information and vignettes for health service quality evaluation in developing countries." *Social Sciences and Medicine,* Vol. 61(9), pp. 1944-51.

Leonard, K. L. and N. C. Masatu. 2006. "Outpatient process quality evaluation and the Hawthorne effect." *Social Sciences and Medicine,* Vol. 63(9), pp. 2330-40.

Levine, R. and R. Eichler. 2009. "Performance Incentives for Global Health". *Center for Global Development, Washington, DC*.

Lindenauer P., D. Remus, S. Roman et al. 2007. "Pay for performance in hospital quality improvement," *New England Journal of Medicine*, Vol. 356, pp. 486-496.

Martinez, A., S.W. Raudenbush and Jessaca Spybrook. 2009. “The Design of Blocked Cluster-Randomized Trials,” Unpublished Manuscript.

Meyer, C., N. Bellows, M. Campbell, M. Potts. 2011. *The Impact of Vouchers on the Use and Quality of Health Goods and Services in Developing Countries: A systematic review*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Musgrove, P. 2010. “Rewards for Good Performance or Results: A Short Glossary of RBF”. www.rbfhealth.org

Olken, B.A., J. Onishi and S. Wong. 2011. “Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia”, unpublished working paper.

Peabody, J. W., J. Luck, P. Glassman, T.R. Dresselhaus and M. Lee. 2000. "Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality." *JAMA,* Vol. 283(13), pp. 1715-22.

Petersen L. A., D. L. Woodard, T. Urech, C. Daw and S. Sookanan. 2006. “Does pay-for-performance improve the quality of health care?” *Annals of Internal Medicine*, Vol. 145, pp. 265-272.

Raudenbush, S. W., H. Bloom, R. Congdon, C. Hill, A. Martinez and J. Spybrook. 2011. Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from www.wtgrantfoundation.org or from sitemaker.umich.edu/group-based.

Raudenbush, S. W.. 1997. “Statistical Analysis and Optimal Design for Cluster Randomized Trials,” *Psychological Methods*, Vol. 2(2), pp. 173-185.

Spybrook, J., H. Bloom, R. Congdon, C. Hill, A. Martinez and S. Raudenbush. 2011. Optimal Design Plus Empirical Evidence: Documentation for the “Optimal Design” Software.”

Witter, S., A. Fretheim, F.L. Kessy, and A.K. Lindahl. 2012. Paying for performance to improve the delivery of health interventions in low- and middle-income countries (Review). *Cochrane Database of Systematic Reviews 2012*, Issue 2. Art. No.: CD007899.

World Health Organization. 2011. World Health Statistics 2011: Indicator Compendium. Geneva, Switzerland.