

DRAFT

H N P D I S C U S S I O N P A P E R

Monitoring and Evaluating Projects:

A step-by-step Primer on Monitoring, Benchmarking,
and Impact Evaluation

Rebekka E. Grun

November 2006



Monitoring and Evaluating Projects:

*A step-by-step Primer on
Monitoring, Benchmarking, and
Impact Evaluation*

Rebekka E. Grun

November 2006

Health, Nutrition and Population (HNP) Discussion Paper

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

Enquiries about the series and submissions should be made directly to the Managing Editor, Janet Nassim (jnassim@worldbank.org). Submissions should have been previously reviewed and cleared by the sponsoring department, which will bear the cost of publication. No additional reviews will be undertaken after submission. The sponsoring department and author(s) bear full responsibility for the quality of the technical contents and presentation of material in the series.

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

For information regarding this and other World Bank publications, please contact the HNP Advisory Services at healthpop@worldbank.org (email), 202-473-2256 (telephone), or 202-522-3234 (fax).

© 2006 The International Bank for Reconstruction and Development / The World Bank
1818 H Street, NW
Washington, DC 20433

All rights reserved.

Health, Nutrition and Population (HNP) Discussion Paper

Monitoring and Evaluating Projects: *A step-by-step Primer on Monitoring, Benchmarking, and Impact Evaluation*

Rebekka E. Grun^a

^a Europe and Central Asia Human Development Department, the World Bank, Washington, DC, USA.

This manual was created to support the Evaluation of the Egyptian Health Sector Reform Program and was partially funded by Middle East and North Africa Human Development Department of the World Bank.

Abstract: This manual attempts to be a practical step-by-step guide to prepare and carry out benchmarking and impact analyses of projects. The audience is policy makers, mainly in the research or strategic departments of ministries. However, more operational departments may also have an interest in the concepts explained. Examples are mainly drawn from the Egyptian Health Sector Reform Project (HSRP), but the manual is written to be of cross-sectoral use.

We assume the perspective of a policy maker that wants to deliver the best Value for Money to citizens and therefore has to select projects based on robust quantitative evaluations. We attempt to present analytical tools solidly grounded in economic theory all while focusing on the practical questions of evaluations. Little space is given to theory, in order to spend more time on the actual steps involved, trying to make this book as easy-to-use as possible.

The manual is divided in two parts: the first part gives a rationale for quantitative evaluations and provides detailed operative guidance for benchmarking and impact evaluation (Chapter 1-4); and the second part provides templates for the tools needed in an evaluation, such as a data collection checklist, a focus group guide, and an example Terms of Reference (ToR) to subcontract an impact evaluation (Chapter 5).

Keywords: Monitoring, Benchmarking, Impact Evaluation, Focus Groups, Surveys

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

Correspondence Details: Rebekka E. Grun, The World Bank, 1818 H Street N.W., Washington, DC 20433, USA. Tel: 1-202-473-4984. Email: rgrun@worldbank.org. www.worldbank.org/eca

TABLE OF CONTENTS

GLOSSARY	vii
HISTORY AND ACKNOWLEDGEMENTS	xi
EXECUTIVE SUMMARY	xii
1. VALUE FOR MONEY.....	1
2. MONITORING: INDICATORS TO MEASURE THE ECONOMICS, EFFECTIVENESS AND EFFICIENCY OF A PROJECT.....	3
WHY DO WE WANT TO MEASURE THESE INDICATORS?.....	3
WHAT DATA DO WE NEED?	4
<i>Expenditure</i>	4
<i>Inputs</i>	5
<i>Outputs</i>	5
<i>Outcomes</i>	6
<i>Environment</i>	7
<i>Dimensions of the data</i>	8
WHAT DO WE DO WITH THE DATA? – CALCULATING THE INDICATORS	10
<i>Arranging the database</i>	10
<i>Calculating the indicators</i>	12
WHAT CAN WE DO WITH THE RESULTS?.....	12
<i>Comparison across units, or project sites (cross-section)</i>	13
<i>Comparison over time</i>	14
<i>Rankings</i>	16
LIMITATIONS OF THIS EXERCISE	16
3. EVALUATING THE IMPACT OF A SPECIFIC PROJECT.....	18
WHAT DO WE WANT TO MEASURE?.....	18
WHY IS THAT DIFFICULT?	18
SO WHAT CAN WE DO?.....	19
<i>A word of caution</i>	22
HOW DO WE GET THE NECESSARY DATA?	23
WHO COULD DO AN EVALUATION?	23
WHAT RESULTS WILL WE GET FROM AN EVALUATION?	23
HOW CAN WE ASSESS THE QUALITY OF THE RESULTS?.....	26
<i>'Goodness of Fit'</i>	27
<i>Are the assumptions valid?</i>	27
4. CALCULATING THE COST-EFFECTIVENESS OF A PROJECT.....	29
COST	29
OUTCOME.....	29
THE FORMULA AGAIN.....	29
PRACTICAL EXAMPLES	30
5. TEMPLATES FOR DATA COLLECTION	31
AN EXAMPLE OF A DATA REQUEST FOR EXISTING DATA	31
A CHECKLIST FOR COLLECTING PRIMARY QUANTITATIVE DATA	33
A CHECKLIST FOR CONDUCTING A QUALITATIVE FIELD VISIT OR A FOCUS GROUP	37
A TEMPLATE FOR TERMS OF REFERENCE FOR AN IMPACT EVALUATION	39
REFERENCES	43

GLOSSARY

Baseline

A baseline is the data collected on project-participants, and outcome indicators *before* any interventions have taken place.

Control Group

A control group is formed by people who haven't received the intervention under the project, but are comparable to those who have, in their characteristics. Most approaches to evaluate the impact of an intervention need data from both project participants and a control group, in order to establish a comparison.

Counterfactual

The most common counterfactual asks 'what would be the situation if the project hadn't taken place', in other words, 'what would the participants feel / be like if they hadn't participated'. Ideally, we would have data on the counterfactual to evaluate the impact of a project. However, this cannot be evaluated in practice as people cannot both participate and not participate at the same time. Therefore, some modern impact evaluation methods seek to approximate a counterfactual through a comparable control group, and/or baseline data for the participants. - (Apart from the counterfactual illustrated here, there are other counterfactuals such as: what would be the situation if we had implemented a different program. The reasoning is identical.)

DEA

Data Envelopment Analysis, a linear programming approach, is concerned with comparing the efficiency of organizations (e.g. health facilities, local authority departments, telecom districts, schools, retailers). It is applied where there are many fairly similar units each of which has multiple inputs and multiple outputs. DEA constructs a series of efficiency indicators, i.e. outputs / inputs, for all relevant outputs and inputs, and then evaluates holistically, (i.e. taking all indicators into account) the performance of the units against each other. There is no universally applied weighting of indicators; each unit has its indicators weighted as is appropriate in its specific environment.

DiD

Difference-in-Differences compares the change in outcomes over time for participants and non-participants. (The name is due to us analyzing the difference over time, *and* between the two groups.)

Endogeneity

In an economic model, an endogenous change is one that comes from inside the model and is explained by the model itself. For example, in the simple supply and demand model, suppose that there is a change in consumer tastes or preferences (an exogenous change). This leads to endogenous changes in demand and thus the equilibrium price and quantity.

In the context of impact evaluation, it is problematic if the participation in the project is endogenous, i.e. influenced by factors which are important for the project otherwise. For example, poorer people are usually also less healthy. So if the HSRP has been assigned to the poorest first, a very simple comparison might look like participating in the program makes you less healthy.

Impact

We say a program had an impact if it had the desired effect on its objectives for e.g. individuals, households or institutions, and if the effects can be attributed to the program.

IV

The Instrumental Variable (IV) approach attempts to correct the endogeneity problem of participation. It does so by finding measures that influence participation, but are uncorrelated with the outcomes, conditional on participation (e.g. health outcomes under the HSRP.) That means, they are uncorrelated with any latent variables that may influence outcomes besides participation, but are not yet accounted for in our evaluation.

Matching

The Matching approach is used to select comparable individuals from a control group. For this we need a group that did not receive the intervention, but has similar characteristics to those who did (including characteristics that influenced the targeting for the participants). We can then run a regression that predicts the likelihood of participation among eligible people, and use the same model to predict likelihood of participation among those who could not participate. This likelihood measure is called the ‘propensity score’. Another possibility to ‘match’ the treated and control individuals is by including the variables that would enter the participation regression (as discussed above) directly in the evaluation regression.

The benefit of the intervention is then calculated as follows: for each person receiving the intervention, we find one person from the control group with the closest propensity score, and calculate the difference in their respective outcomes. The average difference in outcomes is then called the ‘average treatment effect’.

Merging

Merging of data in statistical software packages such as Stata or SPSS commonly refers to joining up two databases at a label they have in common. For example, suppose we have a Stata file of all personnel by facility, with a unique facility code, in one dataset, and another file with the percentage of patients with a certain disease, by facility. Then merging means linking the two datasets at the facility level, so that all information is displayed in one single dataset - with all information correctly assigned to the respective facility. Stata’s help function for example explains the necessary commands in detail.

Panel

A panel data set contains data across two dimensions: time, and a cross-sectional unit, for example individuals. It is always the same individuals that are tracked in each point in time, without fail. Some very useful statistical methods are only possible in a panel data set.

Paragon

A ‘paragon’ is a unit that emerges as the most, or among the most successful in a benchmarking exercise. Basically a ‘leader’ unit others can learn from. Once paragons are discovered in a benchmarking exercise, they can be analyzed in depth, quantitatively and qualitatively, to derive lessons for other units. The term is mostly used in a DEA Benchmarking, but in this manual we use it more widely.

Pooled data

A pooled dataset also has two dimensions, but the cross-section need not always contain the same units over time. For example, the individuals we have data for might change in each time interval. This dataset typically does not allow us to use more statistical methods than a simple cross section. But through the additional dimension it adds more data points, which helps for a more robust evaluation.

Randomization

Randomization is the intentional random allocation of a program to its beneficiaries (be it regions, individuals or institutions), for example by lottery. This has the benefit of automatically generating comparable treatment and control groups, because anybody could or could not be participating, with equal probability. It also has the benefit of being transparent and non-judgmental.¹ Evaluations of a randomized project are called ‘prospective’, and of a targeted project ‘retrospective’. Only retrospective evaluations run into the endogeneity problem.

Regression

Statistical regression analysis models the relationship between one or more dependent variables (usually named Y), and the explanatory variables (also called independent variables, explanatory factors or regressors - usually named X_1, \dots, X_p).

Sample Size

When collecting primary data, it is important to decide how many people (or schools, or health facilities, or other units) to collect data from. Usually, including every unit exhaustively will be too costly, so that we have to consult a sub-sample. Estimating the sample size is important, because without these calculations, sample size may be too high or too low. If sample size is too low, the experiment will lack the precision to provide reliable answers to the questions it is investigating. If sample size is too large, money will

¹ However, randomization also has some limitations. First, often it is politically unfeasible to allocate policies literally ‘by lottery’. Second, a lot of policies are too complex in structure to be ‘administered’ via a lottery-like approach. Third, pure randomization will be difficult because people can always opt out. See IEG (2006).

be wasted collecting data that are not needed. So, we need the sample size just big enough to support the analysis of our parameter of interest without too high an error. So the exact sample size depends on the parameter to be analyzed, and its typical standard deviation. If we want the average of the parameter to be within a confidence interval of the true parameter of 95%, then we need a certain sample size to give us that security. Given that typical standard deviations are known for a few typical parameters economists are interested in, some rules of thumb have been established over the years. One is that the minimum sample size to analyze consumption should be 500. This does not necessarily refer to the overall sample size, but to the sample within each 'cell' of data that we need, for example: consumption in entire Egypt, then for rural and urban Egypt, then for households with more than 4 members in rural Egypt...in each of these cells that we want to cut, we would need 500 data points. – For other parameters, their respective standard deviation and the desired confidence level decide about the sample size.

Selection Bias

Often, a policy has not been randomly assigned to different people. Usually, a policy has been purposefully allocated to certain people according to criteria. This can make comparisons difficult. The non-participants cannot just be taken as a control group, because the participation may depend on criteria that are relevant for the outcomes. For example, poorer people are usually also less healthy. So if a program has been assigned to the poorest first, a very simple comparison might look like participating in the program makes you less healthy. Then a simple comparison of impact between participants and non-participants is not unbiased; it has a *selection bias*.

Treatment group

The treatment group is the group with effective access to the project or policy to be evaluated. Note that not necessarily everyone who is eligible to be treated will participate – if they have a choice over this. The treatment group is formed by the people who actually participate. The name is used to distinguish this group from the control group; a typical evaluation sample is thus divided into the *treatment* and the *control* group.

HISTORY AND ACKNOWLEDGEMENTS

This Manual is a product of the evaluation phase of the Egyptian Health Sector Reform Project (HSRP) at the World Bank. Until then, very few in-house evaluations had been done for the Middle Eastern region of the Bank, and even fewer included a substantial client involvement and capacity building element.

The Egyptian HSRP promised to be different. From the beginning, the client brought a qualified interest and commitment to the evaluation, and the case team at the Bank nourished a strong ambition of producing a thorough evaluation of the impact of the different reform interventions, at least of the substantial infrastructure components.

Given that Bank resources were tight, and client commitment was keen, it was necessary to generate a tool that would allow the client to actively participate in and shape the evaluation of the reform. It is in this context that the case team decided to produce a 'manual' that would explain the background of impact evaluation, and provide transparent and practical tools for data gathering, data analysis and data presentation. It was seen as crucial that the manual would be relevant and respond to the demands of the client, while serving as a reference and toolkit for any future evaluations.

The author has to thank various inspirers, encouragers and contributors. Alaa Hamed from the World Bank, for recognizing the importance of developing a tool for the HSRP and future evaluations, Nagwan El Sammak, Rania Fares, Mohamed Nouh and Marwa El Shorbagy at the Egyptian Ministry of Health and Population, for hard and thorough evaluation contributions and feedback on the manual, Dr Isaac El Mankabadi for coordination of the Egyptian evaluation team; Martin Ravallion, Laura Rawlings, Sebastian Martinez and Ayo Akala, at the Bank, for proof-reading and valuable feedback and last, but not least, my former mentors Michael Ridge and Christoph Riechmann from Frontier Economics Ltd and Costas Meghir and Orazio Attanasio from University College London, for teaching me how to evaluate policies and programs.

November 2006, Rebekka Grun

EXECUTIVE SUMMARY

This manual provides analysts with a practical step-by-step guide to prepare and carry out benchmarking and impact analyses of projects. The audience is policy makers, mainly in the research or strategic departments of ministries. However, more operational departments may also have an interest in the concepts explained. Examples are mainly drawn from the Egyptian Health Sector Reform Project (HSRP), but the manual is written to be of cross-sectoral use.

We assume the perspective of a policy maker that wants to deliver the best Value for Money to citizens and therefore has to select projects based on robust quantitative evaluations. We attempt to present analytical tools solidly grounded in economic theory all while focusing on the practical questions of evaluations. Little space is given to theory, in order to spend more time on the actual steps involved, trying to make this book as easy-to-use as possible.

The manual is divided in two parts,

- The first part exposes a rationale for quantitative evaluations and provides detailed operative guidance for benchmarking and impact evaluation (Chapter 1-4);
- The second part provides templates for the tools needed in an evaluation, such as a data collection checklist, a focus group guide, and an example Terms-of-Reference (ToR) to subcontract an impact evaluation (Chapter 5).

Chapter 1 explains the economic rationale behind Value for Money, and the overall framework of the proposed analysis. Chapter 2 presents a simple process to benchmark projects against each other, and chapter 3 provides the background to understand and subcontract a full impact evaluation. Chapter 4 explains how to calculate the cost-effectiveness of a project or policy and chapter 5 provides templates for all the tools mentioned in the previous chapters.

1. VALUE FOR MONEY

Policy makers face the task of providing services in a world where resources are limited. For example for health care, even in the wealthiest country it will not be possible to provide every beneficial medical service to all citizens. Health care, like other services, has to be rationed. This means that choices need to be made in the allocation of resources, i.e. where to ‘put the money’.

If resources must be allocated, then how would citizens want them to be allocated? – They would want an allocation that provides the best health improvement. Or, in other words, they want the best value for their money (colloquially, the most ‘bang’ for the ‘buck’.)

The question is how to measure ‘value’ and ‘money’. In order to measure true use of resources, we do not only need monetary values, but actual consumption of the resources themselves, doctors’ time, hospital beds, equipment etc. We can then value these at the cost of their provision. The ‘value’, i.e. the health improvement, is more difficult to measure, and should encompass the objectives the policy makers are aiming for, such as better survival rates, increased coverage of a health service, increased take-up, and better customer satisfaction, for example². These outcomes are usually brought about by a series of direct and measurable outputs, such as number of patients treated, number of files produced among others.

In this manual we want to provide tools to measure the value for money of specific projects or investments. We propose a framework that measures the ‘value’ or outcomes of an investment as defined by its objectives; as well as its outputs, its resource inputs, and the monetary cost of the inputs. This allows us to put the outcomes and outputs in relation to the inputs and expenditures and in that way derive the ‘bang per buck’ of a project.

The following graph summarizes the framework.

² Compare Weinstein (2003).

Figure 1: Value for Money

$$\text{Value for Money} = \frac{\text{Outcome}}{\text{Expenditure}}$$

$$= \frac{\text{Input}}{\text{Expenditure}} \times \frac{\text{Output}}{\text{Input}} \times \frac{\text{Outcome}}{\text{Output}}$$

“Economics” **“Efficiency”** **“Effectiveness”**

The relation of the Inputs to their costs is commonly called the ‘Economics’ of a project and answers the question “How cheap did we shop?”. The relation of the Outputs to the Inputs of a project is called the ‘Efficiency’ of a project, and answers the question “How productively did we use our resources?” And the relation of Outcomes to Outputs defines the ‘Effectiveness’ of a project and provides information about how effective each outcome was in bringing about our objectives.

Combining all three measures results in a formula to calculate the ratio of Outcome to Expenditure, which answers our initial question: “What value did we get for our money?”

In order to fill and use this formula, we need to

- Measure Inputs, Expenditures, Outputs and Outcomes, and put them in relation to each other,
- Make sure the change in outputs and outcomes derives from the project to be evaluated, and from nowhere else.

Chapter 2 deals with the first and chapter 3 with the second point. Chapter 4 will bring the content together and apply the formula to derive the cost-effectiveness of a specific project.

2. MONITORING: INDICATORS TO MEASURE THE ECONOMICS, EFFECTIVENESS AND EFFICIENCY OF A PROJECT

WHY DO WE WANT TO MEASURE THESE INDICATORS?

Although we are ultimately interested in measuring the Value for Money of a project, for example, how much it cost to improve children’s diarrhea survival rate, it makes sense to examine the contribution of a project more profoundly. Every investment or project has a value chain of delivery,

- starting with procuring the inputs,
- organizing them to produce a service, and finally,
- obtaining the desired impact of the service.

Every step in this value chain is important to achieve a good Value for Money, and failures along the line can influence the final result.

So in order to provide a good instrument of diagnosis, it makes sense to examine the chain of production from beginning to end. If we recall graph 1, we do this by analyzing indicators for

- the procurement stage, looking at how economically inputs were bought;
- the production stage, looking at how efficiently inputs have been employed to produce service outputs, and,
- finally, looking at how effective provided services have been at bringing about desired results.

Figure 2: Value for Money (repeated)

$$\text{Value for Money} = \frac{\text{Outcome}}{\text{Expenditure}}$$

$$= \frac{\text{Input}}{\text{Expenditure}} \times \frac{\text{Output}}{\text{Input}} \times \frac{\text{Outcome}}{\text{Output}}$$

“Economics”
“Efficiency”
“Effectiveness”

An example for an 'Economics' indicator would be the average cost of staff IT training provided, or the average cost of a PC bought. An example for an 'Efficiency' indicator would be the number of computerized records achieved per number of PCs or per staff hours of IT training. An example of an 'Effectiveness' indicator would be the increase in vaccination coverage per number of computerized records, or per staff hours of training.

When formulating the Economics or Efficiency indicators, it is important to bear in mind to only relate items that logically belong to each other, e.g. the input and the relevant cost of that input; the output and the input needed to provide it (and not an input not needed to provide it). For the Effectiveness indicator, we cannot at this stage attribute outcomes to certain outputs, so here, all outputs are valid denominators. (We shall return to this point in chapter 3.)

Now that we have clarified the concept of Value for Money measurement, the question is, what data do we need, and where do we get it.

WHAT DATA DO WE NEED?

Recalling our framework above, we need data for Expenditure, Inputs, Outputs and Outcomes. - Note that existing data sources sometimes use these terms already, but in a different way. Therefore, in each case, data has to be reviewed carefully according to our conceptual framework, and for each data entry it needs to be defined, in which of our categories it falls.

Below we discuss the basic data needs of a Value for Money comparison. In doing that we make frequent reference to the Egyptian HSRP. For a specific data needs discussion for other sectors, such as utilities, see for example the Chapter 6 in the book of Coelli, Estache et al. (2003).

Expenditure

Data on expenditures or costs are usually held by the controlling department, in current costs of national currency. If we know the procuring unit within the organization, it may be quicker to obtain the data from there. As an example, if we want to evaluate investments under the Health Sector Reform, we can obtain the expenditure on each investment from the unit responsible for the investment, e.g. expenditure on equipment and remodeling of healthcare facilities from the Engineering unit, and expenditure on training of facility staff from the Human Resources unit.

If we collect expenditure data over time, inflation may have distorted the real cost of the later data. Therefore, if more than a year has passed between any two expenditures, all expenditures should be normalized to their value in national currency of one year.

Inputs

Inputs into a project are different from expenditures in that they are physical inputs, e.g. number of chairs or number of PCs, or hours of training, and not their cost.

Data on these are usually trickier to get than data on expenditure. Sometimes, the central procuring unit would know the actual number of inputs and the date they were provided, so again HR for any training and staffing, and Engineering for any equipment.

At other times, however, the point of decision is de-central, in other words, the Governorate or District level, or even the place where the project is being implemented, decides on the amounts and dates of inputs. For example, in the Egyptian HSRP, the District level decides on when to conduct training for the staff in Primary Care facilities. In these cases, some of the data will only be held at the de-central level, and has to be collected from there.

In each case, data should be collected on actual inputs and the date from when they were employed.

Outputs

Outputs are direct products or services derived from a running project. In most cases, these are not recorded by the standard accounting system. In some cases, the unit of implementation (e.g. the Primary Care Facility in the case of the HSRP) will have some records of a few standard outputs, e.g. number of patients treated, the number of scans made etc.

However, the best way to obtain robust output data is to put a monitoring system in place from the first day of a project. The outputs of interest can then be defined as monitoring indicators and collected in regular intervals from the place of implementation. - Chapter 5 provides a detailed checklist for a primary data collection to this end.

Three issues are important when collecting output data, whether from existing sources, or through a new collection system:

- content: what measures should we collect;
- timing: when and how often should we collect; and
- level: how aggregated should our data be?

Content

What the best output measures are depends on the nature of the project to be evaluated. They should reflect the typical physical output that is produced by the project in question, e.g. number of ultrasounds made for a new ultrasound machine, number of staff applying new examination protocols after protocol training etc. Outputs are the immediate 'product' of an activity, not yet the intended effect. So, it is number of ultrasounds made, rather than % of babies delivered well.

Timing

If the output measure in question is measurable *before* the project has been implemented, it should be measured at least once before implementation. In that way a comparison *before and after* the implementation can be made. The data collected before implementation are called the ‘baseline’ data.

Level

Monitoring data should be collected at as de-central a unit as possible. That means, it shouldn’t be collected at the national level, but rather at the level of project implementation.³ In the case of the HSRP, it should be collected at the facility, rather than at the District, Governorate or National level. Afterwards it is always possible to aggregate up, but we cannot disaggregate, if data have been collected at a high level. Also, a more de-central level means we have more points of comparison, which makes for a more robust evaluation.

Outcomes

Outcomes are different from Outputs. They are our final variable of interest. So not the number of patients treated, but the number healed. Not the number of scans made, but the improvement in the related health outcome.

These are usually trickier to measure than outputs, and more long term in nature. As before, a monitoring system should be put in place. And here again, the content, timing and the level of data collection are important.

Content

The outcomes of a project correspond to the intended effect, therefore reaching further than the immediate output. E.g. % of babies delivered well rather than number of ultrasounds made. (This example makes clear why outcomes are usually more long-term in nature than outputs.)

In most cases, the outcomes simply are the objectives initially set for the project. So in the case of the HSRP, the coverage of the population (%) with health care, and the change in children’s diarrhea fatality rate, for example.

³ Note that this preference is only valid if we are talking about the implementation of a ‘micro-economic’ project, i.e. one that is directly directed to individuals, facilities, schools or companies. If we want to monitor a project that is primarily directed at entire sectors (‘meso-economic’) or an entire country (‘macro-economic’), then obviously, both implementation and outcomes would best be observed at the relevant levels, and not less aggregate.

Timing

Here as well, it is vital to start measuring *before* project implementation has started. This is because the impact of the project is more difficult to evaluate if we cannot do a before-after comparison. We need to measure the *change* in the outcome variable. It is vital to collect at least one spell of baseline data of outcomes.

If no baseline has been collected under the management of the project in question, we have to think about constructing one. In this case, we should consult secondary sources, such as household surveys, which are routinely collected in the country. We have to study the questionnaires of all relevant surveys, in order to see whether they happen to collect outcome variables of our interest.

Also, some outcomes are known to have long transmission times. Therefore, the follow-up collection of data, after implementation of the project, needs to leave a time-lag big enough for transmission effects to come through.

Level

If we join various sources of data, such as monitoring data collected under the project, and household surveys routinely collected, we need to make sure the data measures at the same level. That is to say, if our project data measures for example at the facility level, the survey would also need to allow measurement at the facility level. If the survey measures for example at the household level, it would need to be aggregated up to the facility level.

Environment

The success of a project is not only determined by Inputs, Outputs and Outcomes. It also depends on the environment in which the project is implemented. For example, reaching out to the population is more difficult in rural, sparsely populated areas. And areas that already have a lower level of basic services, such as water access, already present a bigger health challenge to begin with.

A good evaluation therefore takes care of environmental factors as far as possible and relevant. Relevant environmental factors for health could be: water access and quality, climate, humidity, endemics, public transport networks, health hazards such as chemical plants, poverty incidence etc.

Environmental factors are different from Outcomes, Outputs, Inputs and Costs, in that they are outside the control of the project implementing site.

There are principally three sources for this environmental data,

- The monitoring system, as before, i.e. primary quantitative data⁴;
- Secondary quantitative data; and/ or
- Primary qualitative data.

Ideally, environmental data would be collected under the monitoring system, from before the implementation of the project. If we cannot collect it under the monitoring system, we can try to get it from secondary household surveys in the way described above.

If no quantitative surveys provide these data, a limited collection of primary qualitative data is in order. Environmental data are not as essential to the evaluation as the other measures; therefore a comprehensive primary data collection (unless already implemented under a monitoring system) is usually not justified.

But it is advisable to back up any quantitative data with a few field visits of project implementation sites (maybe between 2 and 10) in order to understand in what way project environments can differ, and in what way this influences the project outcome. The sample should encompass at least one rural and one urban site, a wealthy and a poor environment and two sites of differing climates (if climate variations play a role). Other categories may be added if they are deemed relevant to the project.

During the field visits, the surveyor should understand the interventions received on the site, what impact they had, and in which way the environment might have influenced their success.

Chapter 5 provides a checklist of points to discuss on a field visit. The same checklist can be used as well as basis of a focus group guide.

Dimensions of the data

Our discussion above reveals that three dimensions are important in our data collection:

- the time;
- the cross-section; and
- the level.

The time dimension of the data means how frequently it has been collected, e.g. annually or quarterly, from before (baseline) or only during and after the project. And how many quarters, or years of data exist.

The cross-sectional dimension asks, for which type of unit we are collecting data, e.g. for towns, for Primary Care facilities, for schools, for companies, or for individuals.

⁴ If data is 'primary' we have collected it ourselves, or designed the collection ourselves and subcontracted it, if it is 'secondary', we have obtained it consulting existing data-sources.

The level then specifies at which degree of aggregation we can obtain data from the cross-sectional units. For example, the national level, the Governorate level, the District level, the facility level, or the individual level.

For each of these dimensions, it is advisable to obtain as many data points as possible. That means:

- the more spells of data over time we have, the better. So collecting quarterly rather than annually is an advantage.
- the more cross-sectional units we get data from, the better – i.e. the more schools, the more care units or the more towns, the better.
- the more disaggregated the level, the better. So, ideally, we collect data at the individual level. If that is not an option, it is still usually preferable to collect at for example the facility level rather than at the Governorate level.

Following any of the three recommendations will augment the number of data points in our data base. The more data we have, the more robust our evaluation will be. Ideally we would want to work from a sample in the 1000s, so for example 400 Primary Care Facilities X 7 quarters of data = 2,800 data points, would be a good sample. The minimum rule-of-thumb for a good evaluation would be a sample of about 1,000. (See chapter 5 for more on this.)

WHAT DO WE DO WITH THE DATA? – CALCULATING THE INDICATORS

If the necessary data have been gathered, it needs to be put into a format that allows an easy calculation of the indicators in our formula. First, we need to arrange our database.

Arranging the database

Before calculating our indicators, it is important to create a consistent database. If we gather data from various sources, for example Inputs and Outputs from Monitoring data, but Outcomes from household surveys, we need to be able to ‘merge’ the sources into one table, so that all the relevant data for one unit, e.g. a certain facility, clearly belong to that facility. That means merging needs to be accurate along all dimensions: cross-section, level and time.

Cross-section

We need to make sure that the cross-sectional units of the data are clearly marked. For example, if the Monitoring data come at the facility level, the facility needs to be clearly marked with a unique name or code. The same is true for the household survey: the units, for example health facilities’ catchment areas, need to be marked with the same name or code as in the other dataset.

Level

Further, we need to gather all the data at the same level of aggregation, i.e. if the Inputs are measured at, say, the level of the health facility, then Outputs and Outcomes should also be measured at the level of the facility. If they are measured at the level of the individual, then all other types of data should be measured at the level of the individual as well.⁵

Time

Also the years and quarters need to match. If one source gathers data at the quarterly level, and another at the annual, and both sources are vital to our exercise, we need to collapse the quarterly data to annual intervals by taking averages. - Ideally however, we would keep all data at the highest frequency (time dis-aggregation) possible.

⁵ However, it is worth briefly mentioning one exception. Sometimes it won’t be possible to gather all measures at the same level. This depends on where the decision on them is being taken. For example, some Inputs, like houses, or computers, may be decided upon and purchased at the National level, although they are subsequently employed at a more de-central level. In this case, the Economics calculations on these figures should be done at the national level, i.e. the average cost of the Inputs is only available at the national level. Efficiency and Effectiveness however, can then be calculated at more de-central levels, as long as the employment of the Inputs can be measured at the de-central level.

Merging

When merging, the dimensions of the data need to match. Cross-sectional units need to match, their level needs to match, and years or quarters need to match. If cross-sectional units and time units all overlap perfectly, we have a ‘panel’ of data. A panel usually describes a dataset that has two dimensions: time and cross-section, and is complete in both.

If, however, some dimensions do not match, there may be remedies.

- If the level does not match, the more disaggregated dataset can be aggregated to the level of the other one. However, this should be avoided as far as possible, as it means losing data points.
- If the time units do not completely overlap, the dataset can be merged as far as possible, with a few years hanging over in different data sets. In this case, we would not be able to work in a complete panel, but in a so-called ‘pooled’ dataset.
- This last point is valid as well if cross-sectional units do not overlap.

Contemporary software-packages, such as Stata or SPSS have specific ‘merge’ commands that allow combining different datasets that wholly or partially overlap in their dimensions. Smaller datasets could also be merged by hand in Excel, but we advise against using Excel for huge sample sizes.

The result of merging data from different tables into a consistent database with the same dimensions could look like the following. We provide an example for the Egyptian HSRP:

Table 1: Example of a merged panel data base

Facility	Quarter	No of computers	Received staff training (yes=1)	No. of community activities	No. of computerized records	% in catchment area vaccinated
FHU A	2003 Q 4	0	0	2	0	40
	2004 Q 1	1	0	4	10	45
	2004 Q 2	1	1	4	50	45
FHU B	2003 Q 4	0	1	1	0	40
	2004 Q 1	1	1	3	20	50
	2004 Q 2	2	1	8	60	60
FHU C	2003 Q 4	0	0	5	0	30
	2004 Q 1	0	0	5	0	50
	2004 Q 2	3	0	6	50	80
	2004 Q 3	4	1	7	120	100
....
...

Stata for example orders data in this format, with the cross-section (here facility) the first criterion of order, and the time dimension following.

Compatibility with other databases

When designing or merging a database, it is highly desirable (usually indispensable) that the data labels be compatible with existing national databases. This means that the same unique identifiers should be used for e.g. geographical areas (postcodes, regional codes), census segments, individuals etc.

Once our table is merged, complete and read into a statistical software package, we can proceed to calculate the indicators.

Calculating the indicators

The very exercise of calculating indicators is now the easiest bit. Once the data are arranged in the format described above, we can easily extract the different ratios suggested by our Value for Money formula.

For example, taking the above table, we can calculate the no. of computerized records (Output) per computer (Input) in the facility, and thereby extract an Efficiency ratio, for each facility, and for each quarter. This tells us about the utilization of the computer equipment, e.g., for FHU B, in 2004 Q 2, this ratio is 30 records/ computer.

Also, we can calculate an Effectiveness ratio, dividing the % of patients vaccinated (Outcome) by the number of community activities (Output). This tells us about the effectiveness of community activities in reaching out to people and mobilizing them. E.g. FHU C in 2004 Q 1 vaccinates 50% of the population with 5 community programs, i.e. 10% per program.

The actual calculation process depends on the software. In both Stata and Excel, the calculation can be written like one would do on paper:

- In Stata, the result feeds into a new variable, so for example, if we have named the relevant variables `%_vaccinated` and `no_of_activities`, the formula will read *Gen Outreach effectiveness = %_vaccinated/ no_of_activities*, and the program will generate a new variable ‘Outreach_effectiveness’ for each facility and quarter, containing the above described Effectiveness indicator.
- In Excel, the ratio can be calculated in a new cell, linking back to the cells with the outcome and output variables, for example in the format “=F1/C1” for the efficiency indicator from the above table, and “=G1/E1” for the Effectiveness indicator; for FHU A, first quarter of data. The new cell will then contain the result for the indicator. This process has to be copied for all facilities and quarters, it is not automatic as in Stata.

WHAT CAN WE DO WITH THE RESULTS?

Once all Outcomes, Outputs, Inputs and Costs are put in relation as suggested by the Value for Money formula, we have generated a new database of Economics, Efficiency

and Effectiveness ratios, for each cross-sectional unit, and each point in time with data available.

We can now work with this database, and derive visual summary statistics and rankings to give us an overview over the performance of the sector.

As for rankings, the ratios allow us principally to conduct comparisons across two dimensions:

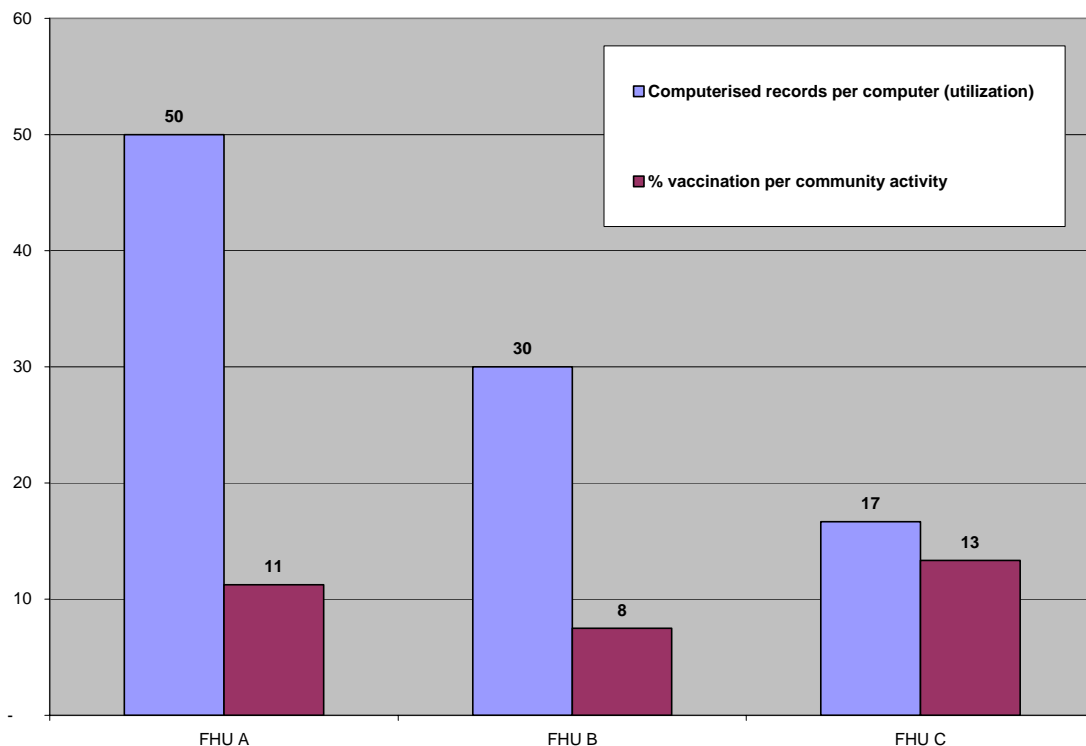
1. Cross-sectional, that is for example between facilities, or project sites; and
2. over time.

Comparison across units, or project sites (cross-section)

The first comparison would benchmark units, for example health facilities, against each other, at the most recent point in time. This comparison could show, which facility is most effective in achieving the project objectives with the project outputs. Or it could show, which facility is most productive (efficient) at employing the project inputs. In this way, we could derive a paragon facility for each dimension of comparison.

We show an example calculating ratios from our data table above:

**Figure 3: An example of benchmarking indicators at one point in time (artificial data):
Computer utilization and % vaccinated per community program in 2004 Q 2**



In this example (with artificial figures), we compare the three facilities from the data table above across two ratios: Computerized records per computer (an efficiency indicator) and the % of population vaccinated per community outreach program (an effectiveness indicator).

We see that FHU A processes 50 records on one computer per quarter, and therefore achieves the highest utilization in this category. We also see that FHU C manages to have 13% of the population vaccinated for each outreach program they run, compared to only 8% for FHU B. This latter result may be an indicator for FHU C's programs being more effective at bringing people to the facility. But it may also be due to other factors.

Paragons in the cross-sectional comparison

In a comprehensive comparison, some facilities might turn out as winners in many dimensions. If one or a few facilities are outstanding for many ratios, it is worth examining them further. This could help identifying success factors, and thereby lessons for other facilities who do not achieve the same standard.

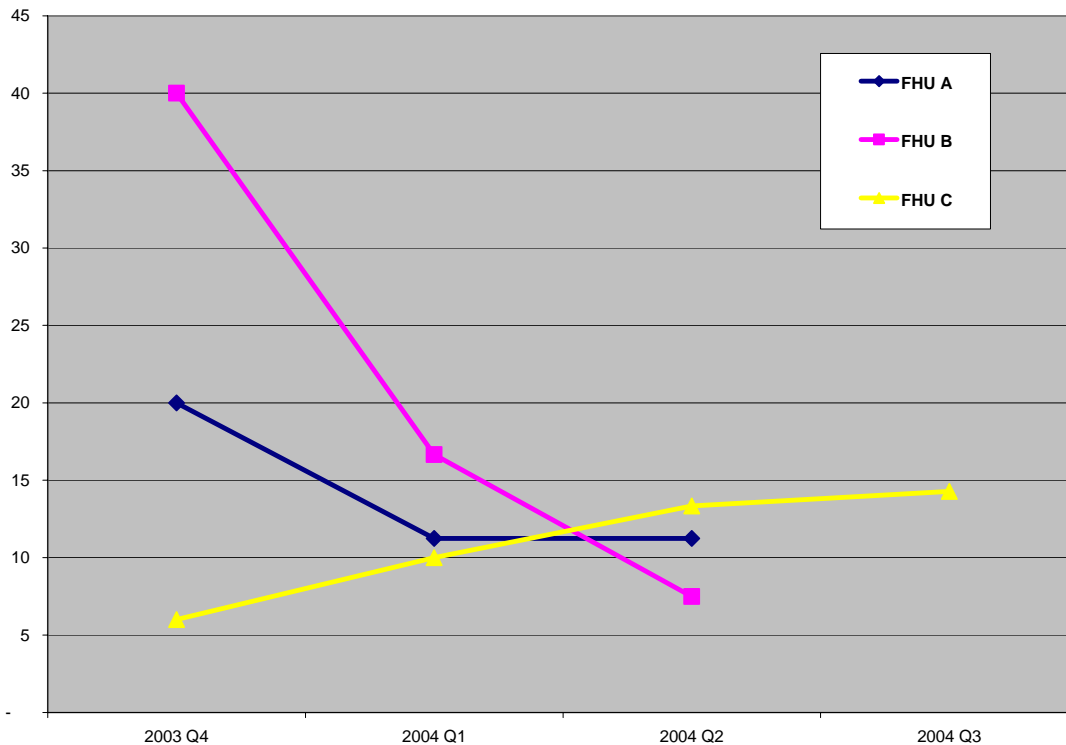
In order to do this, first one should consult the environmental data available from quantitative sources, or collected under the qualitative sources. – Is there an obvious environmental reason why this facility is doing so well? Does it have a very healthy or educated patient base?

If environmental factors are taken into account, but still do not seem to explain the success of the facility fully, further information should be gathered. A field visit or consultations with the directly managing authority (for example the District) are in order. This could reveal interesting local success stories, which could be replicated in other facilities. In this way, lagging facilities can learn from the leading ones, the so-called 'paragons'.

Comparison over time

The second comparison would plot the performance ratios for each facility over time (e.g. over the quarters of data available from 2003, in the case for the HSRP facilities). This comparison allows us to see, whether a facility improved or deteriorated over time, and which facilities improved most. We provide an example from our data table above:

Figure 4: An example of benchmarking indicators over time (artificial data); % population vaccinated per community program from 2003 Q4 to 2004 Q 2 or 2004 Q3



We see that FHU A and B start at a very high level of vaccinations per outreach program, but then decline. In contrast, we see that FHU C improves markedly from an initially very bad performance.

Two remarks are in order here. First, the share of vaccinated people that you can reach is limited: it can never be higher than 100%. But the number of outreach programs can be expanded without a natural limit. This is one of the reasons why we see the curves converge. Second, the data need to be put into context.

Paragons in the time-series comparison

As in the cross-sectional comparison, we should explore cases of extreme performance. On the one hand, drastic improvements or deteriorations in the performance of some facilities may be due to changes in the environment. This should have been captured in the field visits, or in the additional environmental data.

On the other hand, if environmental factors can be confirmed as not relevant, an explanation for exceptional performance has to be sought from the staff or the management involved. As before, lagging facilities can learn from the fast improvers. Also, the underlying factors of deteriorating performance maybe explored and addressed in this way.

Rankings

Both perspectives, cross-section and time, allow us to rank the project units in order of performance. We can establish two useful rankings for each calculated indicator,

- one of the highest absolute performance at the most recent point in time, and
- one of the fastest improvement over the time range observed.

In each of these rankings, performances that can be explained through very unusual environmental characteristics, should be excluded.

As discussed, the main benefit of the rankings is the identification of high performers in the different elements of the value chain of the project in question. Less successful performers can then learn from the leaders.

That said, this simple benchmarking, of one dimension at a time, has some limitations.

LIMITATIONS OF THIS EXERCISE

Above we described how we can compare projects, for example health facilities, with Economics, Efficiency and Effectiveness indicators, for one indicator at a time. This approach has chiefly three limitations:

1. we always only compare one dimension at a time;
2. we only compare one project site to another, without providing an absolute evaluation of the project success; and
3. we cannot attribute the outcomes measured for one project site with security to the project. Other factors may have influenced.

We address each limitation in turn. The first limitation is an issue, because project sites may trade off high performance in one dimension with low performance in another. For example, a facility may be good at employing personnel to do frequent outreach; the same personnel may then not be available to file records, although trained in the filing system. So, a ranking in one dimension does not usually tell us how good the project site is overall.

There is a remedy against this issue, Data Envelopment Analysis, DEA, which combines various dimensions of performance at once, and allows a holistic comparison of project sites against each other. However, the application of DEA is technically fairly sophisticated and goes beyond the objectives of this manual. If this topic is of interest, useful references would be Coelli, Estache et al. (2003) for an introduction to the concept, and Hirschhausen and Kappeler (2004), or Burns, Huggins and Riechmann (2002) for an application. User-friendly software for DEA is available in the market. If there is further interest, the World Bank can provide expertise.

The second and the third limitation go together. In order to evaluate the success of one specific project as a whole, we need to be able to attribute changes in outcomes to the

project in question. This requires a careful comparison both across units and over time, and the consideration of environmental factors.

This is the focus of the next chapter. It deals with attributing the measured outcomes directly to a project, or its various elements, and thereby measuring the impact of a project.

3. EVALUATING THE IMPACT OF A SPECIFIC PROJECT

Stakeholders and policy makers are increasingly asking for hard data on the success of policies. Future budget allocations, or aid money are often dependent on evidence of ‘project impact’. – What is this project impact? And how can we measure it?

WHAT DO WE WANT TO MEASURE?

If a project has an impact, it achieves its original objectives. And we can trace the success back to the project in question, and not to other factors. The objectives of a project are actually its outcomes, as first defined in chapter 2. The purpose of the exercise presented in this chapter is to check how much of the change in the outcomes can be attributed to our project in question.

In other words, we are asking: “How much did our specific project inputs impact on the outcomes?” The outcomes will vary with each project as each project has different objectives, in the case of the HSRP they are for example the diarrhea fatality rate, the % coverage of the population and other health indicators. The Inputs are the different interventions under the reform, for example staff training, remodeling and new computers.

We want to measure the difference our project made to the outcomes in question. That means, we must be able to measure the difference in the outcomes that is attributable to the project.

WHY IS THAT DIFFICULT?

Evaluations of programs like the HSRP typically face the problem of finding a ‘control group’, i.e. a group to compare against, of people that have not received the intervention.

Often, a policy like the HSRP has not been randomly assigned to different people, so that anybody could or could not be participating. (Actually, if it had been assigned randomly, that would make our evaluation work a lot easier, and it could also have political benefits. We will talk about this later on. Here we talk about the more frequent case of a targeted intervention.)

Usually, a policy has been purposefully allocated to different regions, according to criteria. This can make comparisons difficult. The non-participants cannot just be taken as a control group, because the participation may depend on criteria that are relevant for the outcomes.

For example, poorer people are usually also less healthy. So if a program has been assigned to the poorest first, a very simple comparison might look like participating in the program makes you less healthy. This problem is called ‘endogeneity’. The participation is ‘endogenous’. A useful solution therefore either has to generate a new control group or to remedy the endogeneity.

Only in rare cases will we find a ‘natural experiment’, that means, genuinely comparable individuals are equally eligible for the program. This is for example the case if a program has been rolled out in comparable communities at different points in time, with the delay not being caused by criteria relevant to the outcomes. It is also the case, if a program has been purposefully allocated randomly, for example by lottery. This so-called ‘Randomization’ of an intervention has the benefit of automatically generating comparable treatment and control groups. It also has the benefit of being transparent and non-judgmental.⁶ Evaluations of a randomized project are called ‘prospective’, and of a targeted project ‘retrospective’. Only retrospective evaluations run into the endogeneity problem.

The optimal solution for the endogeneity problem is determined by the project in question. Depending on the exact way and timing of the interventions, one analytical approach usually emerges as the best one. Below we present various potential solutions.

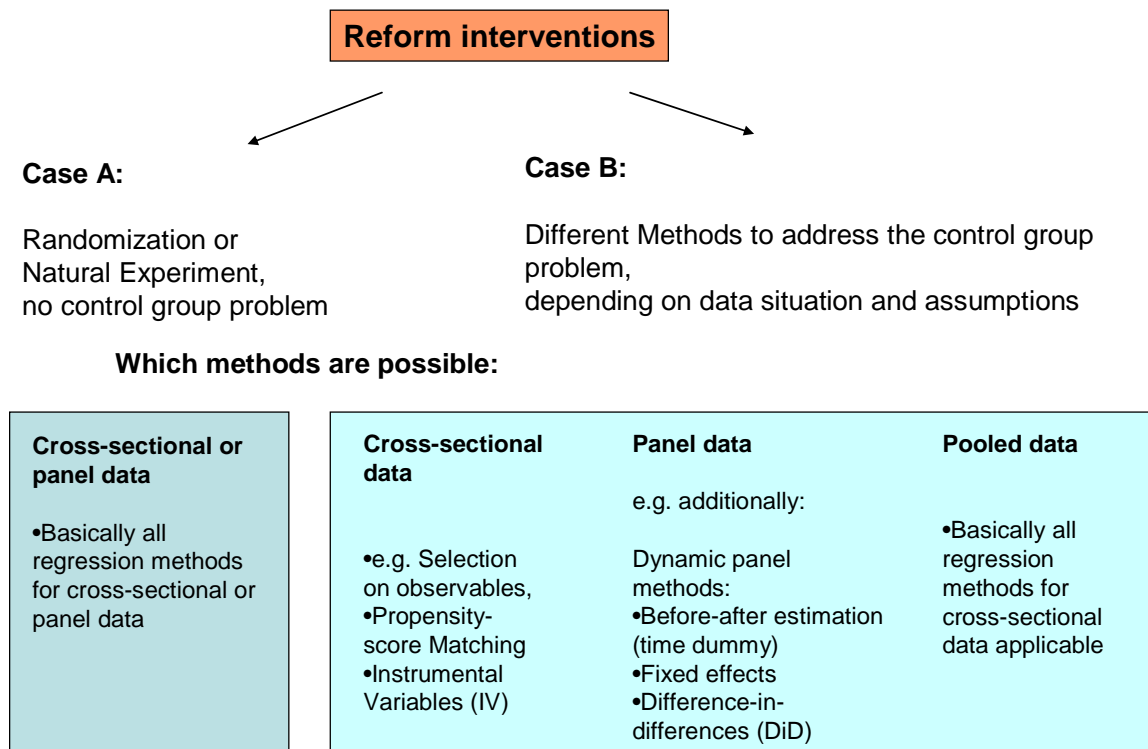
SO WHAT CAN WE DO?

The problem we have is to find a timewise (before-after), or cross-sectional control group. As we want to keep the theoretical discussions to a minimum, below we summarize the most important solutions in a diagram. An explanation will follow after the graph.

]

⁶ However, randomization also has some limitations. First, often it is politically unfeasible to allocate policies literally ‘by lottery’. Second, a lot of policies are too complex in structure to be ‘administered’ via a lottery-like approach. Third, pure randomization will be difficult because people can always opt out. See IEG (2006).

Figure 5: Econometric Measurement of Impact in the Evaluation Literature: Overview



We discuss both cases, A and B in detail.

If the project has been randomized or implemented as a ‘natural experiment’ (Case A), we are lucky. In this case we can assume that, under the perspective of our outcomes (so for example, under a health perspective), the assignment of the intervention has been random. – This case is very rare in reality.

If, however, there is an endogeneity problem (Case B), we principally have four possible methods to measure the impact of the intervention:

1. The **Differences-in-differences (DiD)** approach compares the change in outcomes over time for participants and non-participants. (The name is due to us analyzing the difference over time, *and* between the two groups.) This approach assumes that the control group problem may exist, but that it doesn’t change over time. Comparing the change in outcome for the two groups, the change of the difference between the two groups can be interpreted as the effect of the reform.
2. The **Matching** approach is used to select comparable individuals from a group without the possibility of participation. For this we need a group that did not receive the intervention, but has similar characteristics to those who did (including characteristics that influenced the targeting for the participants). We can then run a regression that predicts the likelihood of participation among

eligible people, and use the same model to predict likelihood of participation among those who could not participate. This likelihood measure is called the 'propensity score'. The benefit of the intervention is then calculated as follows: for each person receiving the intervention, we find one person from the control group with the closest propensity score, and calculate the difference in their respective outcomes. The average difference in outcomes is then called the 'average treatment effect'.

3. The **Instrumental Variable (IV)** approach corrects the endogeneity problem of participation. It does so by finding measures that influence participation, but are uncorrelated with the outcomes. (For the curious, a good example is Card (1992).)
4. The **Regression Discontinuity Design (RDD)** approach is a technique similar to the Matching approach. Matching can be used if people in the program and not in the program are comparable. If they are not, which is often the case in well targeted programs it may be possible to identify the program effect by comparing those who joined to those who almost joined. For example, if a program is allocated to the poorest 20%, and anybody richer did not get it, then this approach would compare the richest among the poorest 20% to the next richer people, who are just above the group of the poorest 20%.

Each of these approaches should control for other factors of influence, environmental or in time, as far as possible.

A word of caution

The above mentioned remedies for the endogeneity issue rest on assumptions, which may sometimes not hold in practice. We discuss the limitations of each technique.

- **DiD** assumes that the trend in outcome variables does not change over time for the treatment and the control group. However, this is a strong assumption and can sometimes not be tested from the data. One also needs to assume that no other program was implemented with the same targeting criteria at the time, which is not always the case.
- The **Matching** technique assumes that we can measure all relevant characteristics of participants and non-participants, and that there are similar people who received and did not receive the program. Both assumptions may not hold in practice. First, if our database is not very rich, it is likely that we do not control for some relevant characteristics, which influenced allocation of the program in question. This omission will bias our results. The effect of the ‘left-out’ variables will be picked up by the variable measuring program participation and bias it in the direction in which the omitted variable would have influenced the outcome. Second, the people who received the program may be entirely different from the people who did not receive it, i.e. all their characteristics relevant for program participation are different. In this case, we say that the treatment and control group ‘do not have common support’ and a comparison is misleading. This assumption can be tested from the data.
- The **IV** approach hinges entirely on finding adequate instruments, which can be very difficult in practice.
- The **Regression Discontinuity** approach rests on the assumption that a program has targeting criteria that create a sharp cut-off point (‘discontinuity’) where people cannot join any more. However, in practice, this is often not the case, and targeting of programs is made more loosely. While there is a RDD technique, ‘fuzzy RDD’, see Hahn et al. (2001) that allows for some discretionary implementation, the discontinuities found in practice are often not enough to be used for identification. – Further, RDD generates results that cannot be generalized to the entire population, because it only analyzes the population at the cut-off point.

Whenever the assumptions of the above techniques do not hold in practice, an impact evaluation using them cannot identify the effect of the program or project in question. Because of this, the ‘gold standard’ of evaluating is using randomization in the implementation of projects wherever possible. - There are tools available to design a randomized intervention, for example ODS, Optimal Design Software, which can be downloaded from the University of Michigan webpage. Also see Duflo and Kremer (2003) in the references for a more in-depth description of randomization and its advantages.

Often, however, evaluation is first discussed when a program has been irrevocably targeted already and randomization is not an option. In this case, the best evaluation approach depends on the actual design of the intervention, and the data availability.

For example, under the Egyptian HSRP, it needs to be determined whether the sequence of interventions at the facility level was due to criteria that have nothing to do with the health outcomes – that would be Case A, a natural experiment; or whether it was due to intentional targeting that can be traced back to observable characteristics - that would be Case B.

In Case B, we would need to ask whether data were available that

- included a control group of non-treated individuals (we would need that for Matching, RDD and DiD),
- included a baseline (we would need that for DiD), and/or
- included variables that are correlated with participation, but not with the outcome (we would need that for IV).

Depending on the structure of the available data, we could then choose the best method.

HOW DO WE GET THE NECESSARY DATA?

Ideally we would design a primary data collection from *before* the start of the project. This data set could then be collected in order to allow different approaches of analysis. In this way we can check the robustness of the result.

Chapter 5 provides a checklist to collect primary data that satisfies the requirements of all the evaluation approaches above.

WHO COULD DO AN EVALUATION?

Because carrying out an impact evaluation is technically- and time-intensive, we believe that policy makers' time is better spent on the conception rather than the technical realization of an evaluation. We propose that policy makers do the strategic thinking behind an evaluation, and prepare Terms of Reference for a consultant or academic to do the technical work. - This further has the advantage to have an independent third party offer a regard on the success of the policy.

Given that the policy makers know best how the project or program was implemented, they can pass this knowledge on in a ToR as far as relevant to the evaluation. From our explanations above, they will be able to sub-contract the necessary analysis, and also evaluate submitted proposals with regards to their feasibility and usefulness.

As a practical reference, Chapter 5 provides a template for an Impact Evaluation ToR.

WHAT RESULTS WILL WE GET FROM AN EVALUATION?

A robust impact evaluation, taking into account our recommendations above, will come up with an estimate of the change in output per change in input. In other words, the

evaluation will say what % change in the objectives, e.g., coverage rate of the population, is due to a change in each of our project inputs.

And the change can be attributed to the respective input, because other influencing factors have been controlled for. This information can later be used for a cost-effectiveness calculation.

Below, as an example, we list the results obtained from a Difference-in-differences estimate of the impact of the Egyptian HSRP on the coverage with the measles vaccine, the share of couples using contraception and the female anemia rate (see columns from left to right.)

Figure 6: Results from a DiD-Regression of health indicators

Dependent Variable	Vacc coverage (measles)		% couples using contraception		Anemia rate (women)	
	Coefficient	z-stat	Coefficient	z-stat	Coefficient	z-stat
<u>Interventions</u>						
HSRP participation (dummy) year 2005	-0.05	-2.3	-0.06	-3.34	0.01	0.35
HSRP in 2005 (dummy)	0.02	1.9	0.05	4.74	0.12	6.93
	0.04	1.5	0.04	1.43	-0.07	-1.53
<u>Controls</u>						
lower urban	-0.04	-1.6	-0.01	-0.6	-0.07	-1.9
lower rural	-0.004	-0.2	-0.022	-1.1	-0.004	-0.1
upper urban	-0.03	-1.0	-0.02	-0.9	-0.07	-1.8
upper rural	-0.06	-2.4	-0.11	-4.7	0.11	2.6
frontier governorates	-0.04	-1.4	-0.13	-5.7	0.02	0.6
district size	-7.54E-09	-0.2	8.11E-08	2.7	-3.33E-08	-0.6
pc consumption	1.83E-05	1.2	-2.4E-05	-1.9	-1.1E-05	-0.5
av hh size	0.005	0.4	0.027	2.3	0.007	0.3
% illit hh heads	-0.0002	-0.3	-0.0002	-0.3	0.0021	1.8
dependency	0.009	0.5	-0.057	-3.4	-0.115	-4.0
unemployment	-0.020	-0.1	0.145	1.2	-0.271	-1.3
crowding	-0.019	-0.6	-0.133	-4.5	0.108	2.1
access water	0.0002	0.5	-0.0015	-3.6	-0.0012	-1.6
washing mach	-0.0002	-0.3	0.0025	5.0	0.0012	1.4
constant	0.73	6.8	0.94	10	0.94	10
<u>Observations</u>	435		449		449	
left censored	1		0		18	
right censored	13		4		2	
LR chi2(17)	46.73		395.46		104.04	
Prob chi2	0		0		0	

This table summarizes the results from a DiD regression made with a Tobit model⁷, from district-level Egyptian data, 2000 and 2005. This is a very specific approach, but it can serve as a more general example of the presentation of regression results.

The first column lists all the potential driving factors of the dependent variables, starting with the typical factors used in a DiD: a dummy (a variable that is 1 if “yes” and 0 if “no”) to indicate whether or not a district participated in the HSRP, a dummy for the year in which the impact should be felt (2005) and an interaction term using both dummies (HSRP in 2005). This last term is the one we read the project impact from. – Further down the column also lists all the ‘control variables’, i.e. other factors that may be different between treatment and control districts, and relevant for the impact. And even further down it lists the number of statistical observations used for each regression, (and how many of them were right- respectively left censored – this only relevant for the Tobit estimator; it is about how many of the observations hit the upper or the lower limit of the 0 to 1 range in the dependent variable), and a few test statistics which we will discuss in the next section.

The next two columns, which are repeated for each dependent variable (“coefficient” and “z-stat”⁸) are the ones where we read the strength of the influence of each factor. The coefficient tells us how much each factor matters. The z-stat tells us, if a factor matters at all.

The z-stat is calculated by dividing the coefficient through its standard error (that is the square root of the coefficient’s variance), and therefore measures, how many standard errors the coefficient is away from zero. This tells us if the effect is significantly different from zero at all; that is the case

- with a 90% likelihood if the absolute value of the stat is >1.5,
- with a 95% likelihood if the stat is absolutely >2, and
- with a 99% likelihood if it is >2.5.

So, for example the interaction dummy (HSRP in 2005), which measures the project impact, is significantly different from zero with 90% likelihood (we say ‘at the 10% level’ of significance) for vaccination coverage and the anemia rate. The HSRP has a significant positive impact on vaccination coverage, and a significant negative one on the female anemia rate – which is a result we want to see for this type of program. It is not statistically significant for couples using contraception.

⁷ A Tobit model is an adequate approach if the dependent variable is a % or a share between 0 and 1.

⁸ For most estimation models, this column would actually be called ‘t-stat’, but for a Tobit it happens to be a ‘z-stat’. Technical details of this difference would go beyond the objective of this manual.

Standard Errors can be tricky!

Note that standard errors (SE) are not always straightforward to obtain. Depending on how the data are structured, there may be correlations we need to take into account to calculate SE properly. This is very important, because the SE is so crucial in determining whether program effects are statistically significant at all. The following is an overview of complications that may arise with SE. There are remedies available for each of the complications, the description of which would go beyond this manual. However, if the data for an evaluation suggest one or more of these complications, the consultant carrying out the evaluation should demonstrate how he overcame it.

- Serial correlation – the behavior of individual units (people or health facilities) is correlated over time; that means the SE are correlated as well.
- Heteroskedasticity – this means, the variance (and SE) of a variable is not constant but varies across observations. For example, the variation in utilization of a health facility can vary more strongly in a larger facility. – Standard software packages can control for this challenge by making SE ‘robust’ through additional calculations⁹.
- Cluster-effects – these occur, if the behavior of individuals (or other cross-sectional units) is correlated between them. For example, people living in the same town may orient themselves in what the others do (‘peer effects’) so that individual variances are linked. This increases the SE of coefficients and has to be taken into account to calculate the significance correctly. Note that cluster-effects are frequent in evaluations looking at individual units as data points (such as patients, students, health facilities or schools – rather than entire regions or countries), but are also frequently forgotten by researchers. Standard software packages give options to correct for this when running a regression.

Now to the ‘Coefficient’. We have only regard to the coefficient after we have determined through the z-stat that it is significantly different from zero at all. - The coefficient measures the *size* of any factor’s impact. For example in the above table, you can see that if a district participates in the HSRP, measles vaccination coverage rises by 4 percentage points, and the female anemia rate declines by 7 percentage points by the second year. Likewise, over time (dummy 2005, compared to 2000), measles vaccination and the spread of contraception increased, by 2 and 5 percentage points, respectively.

HOW CAN WE ASSESS THE QUALITY OF THE RESULTS?

The consultant carrying out the evaluation should be able to demonstrate that the evaluation method chosen is adequate and yielding robust and credible results. There are a variety of tests available to assess the quality of the statistical regression approach.

- We first discuss the statistics for the precision of our estimates; these are typically provided by common statistical software with each regression result, and then
- more specific tests to check the assumptions of the evaluation approaches.

⁹ For example, White- or Newey-West SE are corrected for Heteroskedasticity.

‘Goodness of Fit’

Typically, the software used for regression analysis ‘spits out’ a series of relevant statistics with the regression. These statistics measure what we call ‘Goodness of Fit’, i.e. the prediction power of our regression, or, in other words, how close our assumption that the set of explanatory factors explains our dependent variable comes to reality. The statistics differ with the estimation method used. Some of them are listed in the above table at the bottom of the first column.

- LR chi2(#) and Prob chi2. (The number in brackets gives the number of explanatory factors). The two go together. LR chi2 reports the results of a chi2 (“Chi Squared”) test that assesses if the likelihood that our chosen explanatory factors predict the result of the dependent variable (e.g. vaccination rate) is higher than without including them. Prob chi2 gives the probability that we wrongly assumed the likelihood is higher. So, the lower Prob chi2, the better. If Prob is e.g. 0.05 we say that we reject the probability that our explanatory factors are jointly irrelevant at the 5% level (i.e. with 95% confidence).
- The F-stat, together with Prob(F) is used and read similarly to chi2.
- Another common one is R-squared. This statistic explains how close our regression comes to explaining the variation in the data of the dependent variable. In other words, which percentage of this variation is explained by the movement of the explanatory factors as we express it in the regression.

Apart from the ‘Goodness of Fit’ of the statistical regression, we want to assess whether the evaluation- approach was justified and/or whether it needs additional correction. Below we go through potential challenges to our assumptions, and the tests available in order to see to what extent they are valid or can be remedied.

Are the assumptions valid?

A variety of tests helps us to identify challenges to our evaluation approach, and therefore the right remedies to tackle them. Below we discuss the most important ones, in the order in which they are typically applied:

- No serial correlation in the error term: there are several tests for serial correlation available. The most common one is the *Durbin-Watson* (DW) test¹⁰, with the other tests being a variation on the DW principle. Standard software packages produce the test on the regression with a single command, and emit p-values (like the ‘Prob’ discussed under the Chi-squared and F-tests above). The p-value signals at what level of confidence we can reject the assumption that there is no autocorrelation.
- Validity of Instrumental Variables: so-called ‘over-identification’ tests can be used to test the exogeneity of one potential instrument against another one. The most common one, the *Sargan* test assesses the hypothesis that the IV are

¹⁰ See Durbin and Watson (1951) and (1952).

uncorrelated to some set of regression residuals, i.e. they are not endogenous or reacting to the dependent variable; and therefore they are acceptable, healthy, instruments.

- Validity of the Matching assumption: So-called ‘balancing tests’, such as the *Kolmogorov-Smirnov* (KS) test, assess whether the treatment and control groups were chosen properly, i.e. form comparable groups regarding their predicted probability to participate in the program. The null hypothesis (i.e. the assumption we are testing) for the KS test is of equal balance in the estimated probabilities between treated and control. Standard software packages produce KS with a single command. P-values are read as explained above.

4. CALCULATING THE COST-EFFECTIVENESS OF A PROJECT

This chapter now brings together the content of all the previous chapters.

Once we have identified the impact of each category of input, we can calculate the ‘bang per buck’ or Value for Money.

For this, it is necessary to

- Know the cost of each intervention, and
- Know the corresponding outcome, e.g. an increase in coverage.

COST

We have explained in chapter 2, where we find cost data on each intervention. It is important, as far as possible, to not only include the mere purchasing costs, but as well the costs of accompanying necessary measures, e.g. if a piece of new equipment needs a special installation, a special shipment, or a staff-reorganization, the cost of these must be included as well.

OUTCOME

The change in outcome that corresponds to a certain input is a result of an impact evaluation, as explained in chapter 3.

In the case we don’t have the impact measure from our own evaluation - that is for example, if we plan a project in advance, and would like to know how much impact different interventions are likely to have - it is advisable to search for evaluations made of similar projects in other countries or other sectors. It is legitimate to take their results as a ‘best guess’ to predict an impact ex ante. For example, if after the HSRP, the MOHP were to reform small and medium sized hospitals in a similar way to the FHU, it would be legitimate to use the impact of the HSRP interventions as a best guess for the hospital interventions. This is valid as long as the respective environments are roughly comparable.

THE FORMULA AGAIN

Bringing both measures, cost and impact, together, we can easily calculate the impact per Pound, or Value for Money:

Value for Money = Outcome / Expenditure = Result per Pound

This could be for example, the % increase in pre-natal visits per Pound spent on staff training (if staff training is shown to have an impact on visits in our evaluation).

The nice thing about this measure is that we can calculate it for each intervention with sufficient data, and a shown impact on the outcomes. This allows us to compare the cost-effectiveness of each intervention, and find out, which one brought the most value for our money. This is important information in order to choose between investments or interventions in the future. Which is just the answer to the issues raised in the introduction of this Manual.

PRACTICAL EXAMPLES

For a more practical orientation about Value for Money and Cost-Effectiveness calculations, below I attach a few examples of application of the concept in practice. Note however, that the concrete implementation of the tool varies across examples.

A calculated example of cost-effectiveness analysis in health:

http://www.hsph.harvard.edu/review/review_fall_04/risk_whatprice.html

An article of the use of VfM in health:

www.ruf.rice.edu/~econ/conference/papers/weinstein.doc

The HM Treasury's (UK Min of Finance) guidance on the use of Value for Money (VfM) in PPP procurement:

http://www.hm-treasury.gov.uk/documents/public_private_partnerships/key_documents/ppp_keydocs_vfm.cfm

A website with examples on the use of VfM and benchmarking in education:

<http://www.dfes.gov.uk/valueformoney/>

The HM Treasury's handbook for ex ante estimates (appraisals) of Value for Money:

<http://www.hm-treasury.gov.uk/media/D5E/29/96.pdf>

5. TEMPLATES FOR DATA COLLECTION

Our discussion has highlighted that an evaluation will typically draw on various sources of data. We have discussed

- Existing, or secondary quantitative data;
- Primary quantitative data, collected only for the project in question, such as monitoring data, and
- Primary qualitative data, gathered on a field visit or in a focus group.

Below we present templates to gather data from all three sources. We show

- an example of a data request for data from existing sources;
- a checklist for the collection of primary quantitative data; and
- a checklist to conduct a field visit to gather qualitative data.

AN EXAMPLE OF A DATA REQUEST FOR EXISTING DATA

The data we require for the indicators discussed in chapter 2 can be summarized in a standard data request. If the data are rich enough, it can also be used for the impact evaluation in chapter 3.

The following table gives an example for the Egyptian HSRP:

Data category	Type of data requested	Format	Due date	Contact	Comment
Environmental	Individual level data from the 2002 MOHP survey	Excel, on CD	Wed 30th November	Dr A.B., Tel....	Data need to contain the names of the relevant town, or even a postcode: this allows us to merge the data with other data by postcode.
	Individual level data from the 2004 HIECS Egyptian household survey	Excel, mail or CD	Mid-December	B.C., Senior Economist in the WB office	
Health outcomes and outputs	Monitoring indicators, for each quarter, by facility, outcomes: % of children vaccinated, diarrhea survival rate, maternal mortality rate; outputs: no. of computerized records, no. of families enrolled etc	Excel, CD	7 Dec	Dr C.D., Monitoring Director, Tel...	We need absolute numbers of the outputs (e.g. no. of computerized records) and not already calculated percentages (e.g. % of records complete), but we do get % for the outcomes (% coverage, % MMR etc)
Health outcomes	Health Information System data, by quarter, and for all facilities it was collected for, since 2004	Excel, CD	7 Dec	Dr D.E., Head of IT	
Inputs and Costs: Human Resources	For each training program under the HSRP: Length (total number of hours), number of involved staff, and total cost (derived from the tender documents).	Excel, CD	28 Nov	Dr E.F. Head of HR	
Inputs and Costs: Infrastructure	By facility and category of infrastructure: when it was operational. For all types of infrastructure, at the central level: average cost.	Excel, CD	7 Dec	Dr F.G., Head of Engineering Unit	

A CHECKLIST FOR COLLECTING PRIMARY QUANTITATIVE DATA

In both chapter 2 and chapter 3 we highlighted the need for quantitative data. The analysis in chapter 2 mainly feeds on project monitoring data; the evaluation in chapter 3 needs quantitative micro-data from before the beginning of the project.

We believe that both types of data needs can be satisfied from the same source, and suggest constructing the monitoring arrangements for an intervention so that they can also serve a full impact evaluation. In order for a good and robust evaluation to take place, such a collection of primary data should be built into the design of any project from the very start.

The following checklist summarizes the main points that need to be taken into account for a primary data collection, and can serve as the basis of a ToR.

When should data be collected?

Data should always be collected from *before* the project start. Data collected before an intervention are called the ‘baseline’ and are vital to evaluate the actual impact of a project through a before and after comparison. The survey should cover participants and non-participants (also see below under ‘control group’).

How often should data be collected?

We need one or more follow-up surveys, after the program is put in place. These should be highly comparable to the baseline survey (in terms of the questionnaire, the interviewing etc). Ideally the follow-up surveys should be of the same sampled observations as the baseline survey. If this is not possible, then they should be the same geographic clusters or strata in terms of some other variable.¹¹

How big should the sample be?

The choice of sample size is a trade-off between desired representativity and cost. Primary data collection on the ground, with representative samples, is very costly. So usually a minimum sample size is collected, for a given standard error in the estimation result that has to be achieved.

The lower the standard error needed, the bigger the sample needs to be. For example, a typical labor market analysis with an aspired standard error of 1% might require as many as 8,000 observations. On the other hand, good estimations have been derived from samples of 800 data points.

The exact sample depends on the community and market observed, and the estimation model applied. Market Researchers like to work from the rule of thumb $N=1000$ for quantitative data, academics like to derive the exact sample size from the circumstances.

¹¹ See Ravallion (2001).

Explaining the exact formula would go beyond the scope of this manual, but if a primary data collection is sub-contracted, the contractor should be held to explain the sample size suggested through a transparent formula.

Where should data be collected?

Here, two issues are important:

- the level at which data are collected,
- from whom data are collected.

As discussed in our Data section, the more disaggregate the level of data the better. At least, we should gather at the level of project intervention. So, in the case of the Egyptian HSRP, ideally data would be available at the individual level, marking clearly to which FHU an individual belongs. As this is not possible, we gather data at the level of the facility, which is effectively also the level of project intervention.

Then, our sample collected needs to be statistically representative of the overall project. Ideally, we would gather data from everybody receiving an intervention, exhaustively, but often this is not possible. In these cases, we must make sure the sample collected represents the characteristics of the entire project community in the same shares.

Who is the ‘control group’?

We need a representative sample survey of eligible non-participants as well as one for the participants. Ultimately, our data serve to measure the impact of the project, so data should be collected from a treatment and a control group, as described in chapter 3. The larger the sample of eligible non-participants, the better it will facilitate good matching. If the two samples come from different surveys, then they should be highly comparable surveys (same questionnaire, same interviewers, same survey period and so on.)¹²

A control group can be built in various ways, in principle through randomization, through a natural experiment or through a pilot design.

- Randomization would mean that the program is randomly allocated to a subset of the ultimate target group. This can be politically difficult, but provides an ideal design to be evaluated;
- A natural experiment arises when the program is allocated based on criteria that have nothing to do with the outcome. This can for example be the case if it has been delayed randomly in some areas, so we can compare people who have and who have not yet received the intervention, or if it was assigned based on administrative criteria that have nothing to do with the objectives of the program; and
- A pilot design rolls out the program in only some eligible areas, while leaving out others, or leaving them for later. Usually there is some prioritization based on specific criteria. These criteria need to be recorded carefully.

¹² Compare Ravallion (2001).

A random design, or a natural experiment are to be preferred; they provide a much better ground for a later evaluation. The most important thing is, however, that a control group exists at all.

What kind of data should be collected?

The content of our primary questionnaire is driven by the project to be evaluated. So for example, to monitor and evaluate the HSRP, the following topics would need to be covered:

- Inputs, like no. of PC, staff hours training, no. of other new equipment, and an Input measure for any other reform intervention;
- The respective costs of the Inputs;
- Their time of purchase, and their time of first use;
- Outputs, like no. of computerized records, no. of files, no. of patients with a certain grievance seen, no. of community programs;
- Outcomes, like % of population in catchment area reached by community programs, % of population vaccinated, % of users satisfied with the service.. etc. The outcomes should measure all the objectives of the reform.
- Environmental factors, such as water and transport access, cultural issues, health hazards, rural/urban, population density or any other factor that could influence the outcome of the service supplied.
- If we collect data at the individual level, we need variables to control for 'heterogeneity', i.e. differences between individuals. These data should include at least:
 - age;
 - gender;
 - education;
 - occupation; and
 - employment
- In the light of the evaluation methods described in the main text, it would also make sense to collect Instrumental Variables, i.e. variables that are connected with participation in a project or program, but are not connected to the outcome.¹³

When designing the content of a primary questionnaire, as far as possible each question should be formulated in a way that allows quantifying, so e.g. 'no of PCs' rather than 'do you have a PC', and 'no. of hours training' rather than 'did you receive training', as well as '% of houses with fresh water connection', rather than 'does this region have a fresh water connection'. This allows for a more precise evaluation.

Finally, it is worth having regard to a few rules of thumb:

- Any questionnaire should ideally not last longer than 30min. Shorter is better.
- A questionnaire should start with general, non-intrusive questions, and ask the personal details, like age, contact details etc at the end;

¹³ Also see Ravallion (2001) for a data checklist.

- Any questionnaire should be first tested in the field ('pre-test') to see which questions work, and how they are understood. Then, the final questionnaire should be adapted to a workable format.
- If the MOHP is sub-contracting a primary data collection, some key staff should sit in on a few pre-test interviews, and take control of the final shape of the questionnaire.

For further reference on these questions, consult for example *Grosh, M. & Muñoz, J. 1996. A Manual for Planning and Implementing the Living Standards Measurement Study Surveys. LSMS Working Paper #126, The World Bank*

A CHECKLIST FOR CONDUCTING A QUALITATIVE FIELD VISIT OR A FOCUS GROUP

As discussed above, we suggest conducting field visits for two reasons:

- to develop a practical understanding of the value chain of a project and the linkages between inputs, outputs and outcomes on the ground; and
- to capture environmental factors influencing the project outcome, that have not been taken into account quantitatively.

Below we propose a brief checklist that serves as a first orientation for a field visit. It will need to be adapted and detailed according to the scope of the project in each case, but it highlights the headings which need to be discussed. We give concrete examples from the context of the HSRP.

The issues we suggest can be covered

- in a personal interview with people involved in project implementation, or
- through focus groups with people affected by the project: staff, clients/patients.

The advantage of personal, possibly 1-2-1 interviews, is that individuals will generally be more open than in the quasi-public atmosphere of the focus group, so that also sensitive issues may be raised.

The advantage of a focus group is that it draws on various opinions and the group dynamic generates new associations, coming from the combination of a variety of perspectives.

In any case, it is important to identify a target group of potential interviewees ahead of the field visit. The target group should comprise at least:

- One person with a leadership function (and therefore a certain overview of the site). If there are various layers of management, e.g. project site and district, it would be ideal to have a manager from each level;
- One or various persons working on the site day-to-day in the key service provided, so e.g. a doctor or a nurse in a FHU;
- If possible, one or various recipients of the main service provided.

Qualitative interviews serve to put the quantitative findings into context, without aspiring to representativity. Therefore, sample sizes of total people interviewed can be small, and samples around N=10 are common practice.

Below we suggest a few headings that would usefully be covered in a field visit for a Value for Money evaluation.

Issue	What do we need to cover?
Site visited	Description: name, region, type (e.g. San Stefano, Alexandria, FHU4 and centre)
Catchment area	Description of number and type of people served by the service, e.g. number of patients, % of coverage of catchment area, their typical illnesses, their average age, their typical occupations
Utilization	The extent to which staff and equipment capacity is used up
Interventions	Interventions that have been happening under the project, e.g. remodeling a building, installing new equipment
Impact on clients	Clients' view of changes, e.g. how do the patients of a reformed facility under the HSRP think the interventions benefited them, and what disadvantages did they have. Why? What are the links?
Impact on staff	Staff's view of changes, e.g. how do the patients of a reformed facility under the HSRP think the interventions benefited them and what disadvantages did they have. Why? What are the links?
Impact on environment	Did the project have an impact on the environment, its catchment area, neighboring sites or services? E.g. did the facility lose patients to, or win patients from a nearby hospital? Why? What about attractiveness of the area to residents, pollution, co-operation with other services? Why?
Environmental factors	Are there any factors that might influence the project success? E.g. is there a health hazard near the facility, a polluting plant for example? How is the access to water, transport and other services? How wealthy is the community of the catchment area? Are there established channels for information and outreach, e.g., newspapers, mosques?
Recommendations by staff	What would the staff change about the project, its implementation and management?
Recommendations by clients	What would the clients, e.g. patients, change about the project, its implementation and management?
Recommendations by implementors	Are there any recommendations by the people managing and implementing the project?

A TEMPLATE FOR TERMS OF REFERENCE FOR AN IMPACT EVALUATION

Below we give a possible outline for Terms of Reference, explaining under each heading what information is needed, and why.¹⁴

I. Background

A ToR should start with giving the consultants a bit of the history behind the project: why it was started, what its focus is, and what the project objectives are. (A very short para is enough.)

II. Project Objectives

The more background that can be given on the project objectives, and their rationale (i.e. why the Government thinks specifically this project can reach them), the better. This section could cover about two medium-size paras and dwell on project outputs and their relation to outcomes.

III. Project Approach

It would be helpful for the consultant to know how the project was implemented, that is, in general, to know about the involved agencies, decisions taken and their chronology.

If there are various project components, even if only some of them are to be evaluated, they should all be explained with their

- Concept,
- Timeline,
- Any reasons that delayed/affected implementation,
- Actions/interventions and material involved,
- Any variations in interventions and
- Deciding agency.

This section can extend over various pages, as it will cover some of the most valuable information for the consultant.

IV. Project Impact Evaluation

Here you can explain what the task for the consultant will be:

- Which project components do you want evaluated?
- What do you want to find out?

¹⁴ This sample ToR owes much to the book by Judy Baker (2000).

There is little need to describe the methods to be used in detail (chapter 3). A good consultant should know these and be able to explain why they use which. For the consultant, it is most important to know which data are available, so they know which method to choose.

V. Data

This very important section needs to explain, in detail, which data are available from the project data collection system and outside sources. The sources covered should comprise:

- Project Monitoring System;
- Project Accounting System;
- Country's Household Surveys;
- Censuses; and
- Relevant Ministry's accounting data (if relevant: e.g. procurement under the project).

The data would need to be described regarding its:

- content (i.e. what info is available);
- time series (for how many time periods is it available, at what frequency);
- cross-section (who is asked? People, facilities, companies..?); and
- level (at what level is it available? E.g. facility or governorate?).

The data description should keep our theoretical explanations from chapter 3 in mind, because the consultant needs to check the data to know which methods are feasible. Chapter 3 tells you which knowledge about the data is relevant. All these relevant points need to be covered in the data section of the ToR. The section should make clear for example, whether Instrumental Variables are available.

If there are not enough existing data to meet the proposed analysis' data needs (see chapter 3), additional data need to be collected. This section should then also lay out the proposed content, the timeframe, and procedure of additional data collections (see the above checklist for Primary Data Collection). – But additional data collection can also be left as an option to the consultant.

VI. Work Schedule

The ToR should ideally lay out a required work timeline, in monthly detail, and some key milestones for the evaluation. The Milestones could be evaluations for separate project components, or simply the presentation of interim results.

For an Impact Evaluation, it makes sense to give the consultant a little time to mine the data after receiving them, say a month or two, and require a presentation of a detailed research proposal after that, as a first milestone. In this way, the contracting Government agency can give feedback on the proposed research, and also learn from the procedure of analysis.

A second milestone could then be interim results (presentation only) and a third and last milestone the final presentation and final report.

This is just a suggestion for a work schedule. The contracting Agency can also leave this open and specify a more detailed schedule with the consultant once selected. Or it can leave the work planning to the consultant altogether, if there's no time pressure.

VII. Budget

The ToR should close with the proposed budget, if it is fixed.

REFERENCES

Baker Judy: “Evaluation the Impact of Development Projects on Poverty: a Handbook for Practitioners”, *LCSPR/PRMPO*, The World Bank, Washington DC 2000

Burns Philip, Huggins Mike and Christoph Riechmann: “Choice of model and availability of data for the efficiency analysis of Dutch network and supply businesses in the electricity sector”, DTe, Netherlands Electricity Regulatory Services, and Frontier Economics Ltd, London 2002; <http://www.frontier-economics.com/publication.php?id=37>

Card David: “Using regional variation in wages to measure the effects of the Federal Minimum Wage”, *Industrial and Labor Relations Review* 46: 22-37.

Coelli Tim, Estache Antonio, Perelman Sergio and Lourdes Trujillo: “A Primer on Efficiency Measurement for Utilities and Transport Regulators”, World Bank Institute, 2003

Duflo Esther and Michael Kremer: “Use of Randomization in the Evaluation of Development Effectiveness”, Paper prepared for the World Bank’s OED Conference on Evaluation and Development Effectiveness in Washington D.C., July 2003

Durbin, J. and Watson, G.S., "Testing for Serial Correlation in Least Squares Regression I", *Biometrika*, Vol. 37, 1950, pp. 409-428.

Durbin, J. and Watson, G.S., "Testing for Serial Correlation in Least Squares Regression II", *Biometrika*, Vol. 38, 1951, pp. 159-178.

Grosh, M. & Muñoz, J. A Manual for Planning and Implementing the Living Standards Measurement Study Surveys. LSMS Working Paper #126, The World Bank, 1996

Hahn Jinyong, Petra Todd, Wilbert Van der Klaauw: “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design” *Econometrica*, Vol. 69, No. 1, pp. 201-209, Jan., 2001

IEG (Independent Evaluation Group): “Impact Evaluation – The Experience of the Independent Evaluation Group of the World Bank”, The World Bank 2006

Ravallion Martin: "The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation.", *The World Bank Economic Review*, Vol. 15. No. 1, 115-140, 2001

Savin, N.E. and White, K.J., "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors", *Econometrica*, Vol. 45, 1977, pp. 1989-1996.

v. Hirschhausen Christian and Andreas Kappeler: "Productivity Analysis of German Electricity Distribution Companies", DIW, Berlin 2004; <http://www.diw.de/deutsch/produkte/publikationen/diskussionspapiere/docs/papers/dp418.pdf>

Weinstein, Milton: "Getting Value for Money in Healthcare: Can Cost-Effectiveness improve Health?" Center for Risk Analysis, Harvard School of Public Health, Boston 2003



HEALTH, NUTRITION,
AND POPULATION



HUMAN DEVELOPMENT NETWORK

THE WORLD BANK

About this series...

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

Enquiries about the series and submissions should be made directly to the Managing Editor Janet Nassim (Jnassim@worldbank.org) or HNP Advisory Service (healthpop@worldbank.org, tel 202 473-2256, fax 202 522-3234). For more information, see also www.worldbank.org/hnppublications.



THE WORLD BANK

1818 H Street, NW
Washington, DC USA 20433
Telephone: 202 473 1000
Facsimile: 202 477 6391
Internet: www.worldbank.org
E-mail: feedback@worldbank.org