Planning, Monitoring, and Evaluation:

Methods and Tools for Poverty and Inequality Reduction Programs

Poverty Reduction and Economic Management Unit

Poverty Reduction and Equity Unit

Gita Busjeet

THE WORLD BANK
Washington, D.C.

Acknowledgements

This toolkit has been led by Gita Busjeet with comments from Keith Mackay (Consultant, PRMPR), Philipp Krause (Consultant, PRMPR), Helena Hwang (Consultant, PRMPR), Bertha Briceno (Senor Monitoring and Evaluation Specialist, The World Bank), Indu John-Abraham (Operations Officer, LACPRMPR), and participants at the workshop on evidence based policy organized by the Independent Evaluation Group. This work was carried out under the guidance of Gladys Lopez Acevedo (PRMPR). We would also like to thank Michael Alwan for editorial assistance.

Vice President	Otaviano Canuto
Sector Director	Jaime Saavedra
Sector Manager	Jaime Saavedra
Task Manager	Gladys López-Acevedo

Table of Contents

Introduction 3

Ex Ante Distributional Analysis **7**

Poverty and Social Impact Analysis—World Bank *10*

Ex Ante Poverty Impact Assessment—Organisation for Economic Co-operation and Development *11*

Ex Ante Cost Benefit **12**

Agribusiness—International Finance Corporation *15*

Causality Frameworks **17**

The Matrix of Indicators—Mexico *20*

System Dynamics—Bangladesh *21*

Benchmarking **22**

International Benchmarking Network for Water and Sanitation Utilities *24*

The Public Sector Benchmarking Body—Ireland *25*

Process Evaluations **27**

Process Evaluation—Mexico *29*

Process and Implementation Analysis of the Welfare-to-Work Grants Program—United States *31*

Impact Evaluations **33**

Rural Education—Madagascar *36*

Small and Medium Enterprises—Mexico *38*

Executive Evaluations **39**

Avaliação Executiva dos Projetos Estruturadores—Minas Gerais, Brazil *42*

Evaluación Ejecutiva—Department of Planning Colombia *43*

Indicator Evaluations **45**

Avaliação Executiva dos Indicadores—Minas Gerais, Brazil *47*

Evaluación de Programas Gubernamentales—Chile *50*

Evaluation Assessment **50**

Randomized Control Trials Checklist—Coalition for Evidence-Based Policy *53*

Evaluation Report Standards and Rating Tool—United Nations Fund for Children *55*

Introduction

As we enter the second decade of the twenty-first century, governments, international organizations, nongovernmental organizations (NGOs), philanthropic organizations, and civil society groups worldwide are actively focusing on evidence-based policy and increased accountability to stakeholders (Results Agenda[1]).The widespread implementation of the Results Agenda has generated a plethora of books, guides, academic papers, trainings, and case studies, which has enabled an ongoing maturation process in the field. Consequently, specialists are now better equipped to understand what works under which circumstances. Broadly speaking there are two interrelated questions which must be answered when assessing the sustainability of a government Results Agenda. First, is the institutional design and practice of government conducive to evidence-based policy making? Second, are the overarching monitoring and evaluation (M&E) methods and specific tools used appropriate for garnering the evidence demanded by government?

These series of notes aim to make a small contribution to the latter question by summarizing and highlighting a selection of PM&E methods and the tools that governments and international organizations around the world have developed to put these into practice in their own contexts.[2] The central goal of this initiative is to prompt a process of learning, reflection and action by providing practical information to those whose leadership role requires them to understand PM&E methods and their potential for enhancing evidence-based policy making.

Viewed using the technocratic framework of the program cycle, public servants involved with program design and planning, implementation management, and follow-up are continuously faced with decisions and judgments. The Results Agenda aims to ensure that these decisions and judgments are made based on concrete evidence of actual conditions. The tools showcased in this series show how each PM&E method has proved useful in providing that evidence and helped to integrate M&E into the program cycle. The question we would like to emphasize is "To what extent is this PM&E method suitable for my needs?" Each methodology has strengths and weaknesses given a specific context. For example: What type of program are we focusing on? What are the needs of the public sector in terms of results information? What are the resources, data, and time restraints? Table one highlights some of the issues that have arisen when determining if a specific M&E method highlighted here is suitable for a given context.

Figure One: Program Cycle

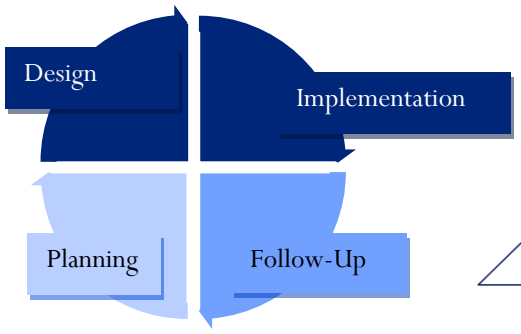
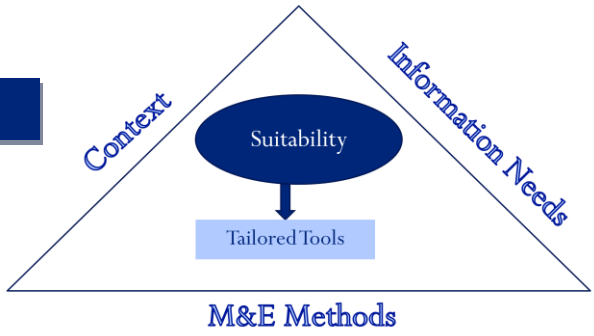


Figure Two: Suitability



[1] In this document we use the overarching term Results Agenda to describe these connected movements

[2] An extensive literature considering the issues surrounding public sector adoption of the results agenda and institutionalization of M&E systems exists. In this context an informative publication is Mackay, K., 2007, *How to Build M&E Systems to Support Better Government*, World Bank, Washington, DC.

Table One: Issues when Determining Suitability

	Questions	Methods	Comments
Design and Planning	If this policy is implemented who will be the winners and losers?	Ex Ante Distributional Analysis	This method is in particular useful for new or redesigned programs with lacking or limited investigation of target populations and other stakeholders. In spite of its upfront costs, investment in this method can be very cost effective in the long run, allowing for the adjustment and refinement of programs before implementation, because programs are likely to be better targeted as a result. Distributional analysis can also provide invaluable information about the political consequences of new programs.
	From a welfare perspective, given limited public resources, should we invest in this program?	Cost-Benefit Analysis	This method is most often used for investment programs where benefits and costs can be easily expressed as a monetary value, such as in infrastructure or agricultural projects. However there have been many innovations in cost-benefit analysis to address this issue. Cost-benefit analysis relies heavily on assumptions and forecasting; it may thus be less suitable for programs planned to be operating in unstable environments.
	What results do we wish to achieve and how do we plan to achieve them?	Causality Frameworks	This method is suitable for all programs; the development of a good causality framework is a vital foundation for good program design and M&E. The process underlying the development of the causality framework is important and often involves multiple stakeholders in discussions and training of program staff if they are not familiar with the method. Therefore developing good causality frameworks can be time and labor intensive.
	Who can provide lessons to improve the program throughout the program cycle?	Benchmarking	This method is suitable for programs that rely on performance indicators to guide management decisions. It is often used by higher-level policymakers to identify well and poorly performing programs that are suitable for comparison. Benchmarking supports the adoption of realistic and challenging targets in programs. It can be difficult to find appropriate benchmarks because of data constraints or lack of cooperation from affected programs.
Implementation and Follow-up	Have operational mechanisms supported the achievement of program objectives?	Process Evaluations	This method is important to inform decision making at both the implementation and follow up stages of the policy cycle. Without accepted standards of quality and its necessary contextual nature (operations vary in each locale) implementing this method can involve high costs in developing an appropriate design and ensuring quality. Process evaluations tend to be very affordable once quality is ensured and can provide excellent value-for-money information.
	Has the program performed from a comprehensive perspective?	Executive Evaluations	This method is suitable in the context of larger evaluation initiatives, driven by central agencies, such as the office of budgeting or the planning department, when these for example have a desire (i) to complement other more focused and in depth evaluations used in government with a rapid evaluation method and (ii) provide overall performance information to stakeholders other than those directly involved in a program such as budget offices, congress, and the public.
	Has participation in the program resulted in planned impacts on target groups?	Impact Evaluations	This method is known to produce very reliable statistical results and has been instrumental in transferring knowledge internationally. Issues have been considerations of the ethical and political consequences of using randomized trials. Budget constraints are also a limitation to the use of this method because these evaluations require a significant time and resource investment. As such the method is most suitable for larger programs with high coverage.
	Is the information from M&E reliable for decision making?	Assessment of Indicators & Assessment of Evaluations	These methods can be very cost effective, helping in particular to enhance M&E capacity in organizations and ensure sustainability of M&E initiatives. A barrier to the use of these methods is that in the context of limited budgets there is often little money left for M&E quality control after evaluations have been completed.

The perspective of the policy cycle remains one of the most useful vehicles for communication and learning today. That said it is also important that we put individual PM&E methods in context of an M&E systems approach. Below four priorities for the advancement and strengthening of the M&E systems in relation to the use of PM&E methods are highlighted.

1. ***Menu of Evaluations:*** Thanks to the maturation process that the Results Agenda has undergone we now have a host of dependable and tested methodologies that are designed to address specific results information needs that arise during the program cycle.1Given this supply of refined methods organizations should take the approach of a Menu of Evaluations, engaging in evaluation planning for their program portfolios aligning different methods, program contexts, results information needs, and budgets for evaluation. Based on this exercise, which evaluation methodology is used, for which program and when during the program cycle, will be determined. A Menu of Evaluations approach is also a call for increased emphasis on evaluation planning and cost-effectiveness in the Results Agenda.

2. ***Reliability:*** For the Results Agenda to be sustainable in the long term it is vital that the evidence that M&E provides for decision making is reliable and leads to real improvements. M&E methods are not always applied to the highest standard due, among other reasons, to a lack of infrastructure (high-quality data systems), inappropriate application of methodology (impact evaluation when a process evaluation was needed), or non-integration of findings into decision-making processes. It is important as we move forward that regular quality control of M&E tools themselves and initiatives to improve the quality of M&E are made integral to the Results Agenda.

3. ***Systematic Integration of Poverty and Inequality Analysis****:* Our understanding of poverty and inequality continues to deepen. Due to innovations in analytical frameworks, data collection, and technology it is possible to understand the poverty context within which a specific policy or program will operate to a higher degree. It is important that moving forward, these advances are benefited from and ex ante poverty and inequality analysis becomes integral to the groundwork for programmatic design and poverty- and equity-centered M&E during the policy cycle. Front-end investments in tailored poverty and inequality analysis will increase the effectiveness of public sector expenditures for policies to reduce poverty and inequality.

4. ***Equipping Programs for M&E throughout the Program Cycle***: The successful use of M&E tools to provide the evidence needed to meaningfully inform decisions made throughout the program cycle depends on many different variables. One crucial step is, where possible, not to approach M&E as an ad hoc activity but from the onset of program design to equip a program with the mechanisms that will allow for high-quality M&E throughout the program cycle. This has not always been possible, given the context of the Results Agenda being adopted by organizations around the world with long-existing policies and programs. Moving forward, however, organizations implementing a Results Agenda should see early adoption of M&E as a priority.

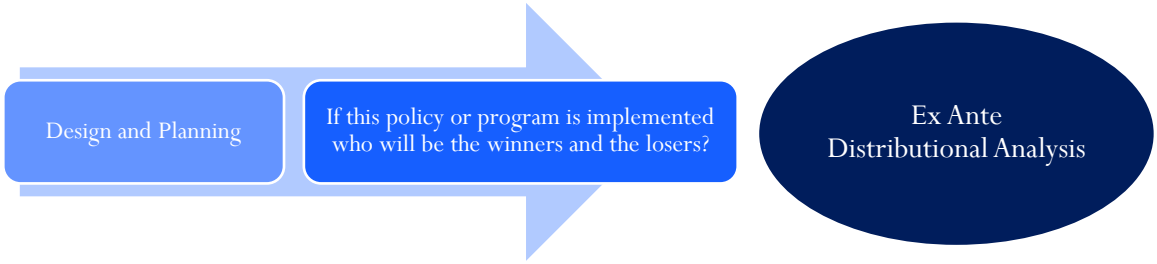
---

[1] Please note that some of the evaluations highlighted such as executive and process evaluations are frequently used for monitoring purposes. Striking a balance between monitoring based on performance indicators and more extensive evaluations for programs that have longer lifetimes going through various cycles is an important part of integrating an evidence focus in programs.

Table Two: Aligning Key Areas of Work, the Program Cycle, Methodologies, and Tools

Areas of Work for the Results Agenda	Stage of Program Cycle	Information Needs/Questions during the Program Cycle	M&E Methodologies	M&E Tools Highlighted
Systematic Integration of Poverty and Inequality Analysis	Design and Planning	If this policy is implemented who will be the winners and losers?	Ex Ante Distributional Analysis	(i) Poverty and Social Impact Analysis—World Bank (ii) Ex Ante Poverty Impact Assessment—Organization for Economic Cooperation and Development (OECD)
		From a welfare perspective, given limited public resources, should we invest in this program?	Cost-Benefit Analysis	Cost-Benefit Analysis of Value Addition to Firms Agribusiness Projects—International Finance Corporation (IFC)
Equipping Programs for M&E		What results do we wish to achieve and how do we plan to achieve them?	Causality Frameworks	(i) The Matrix of Indicators—Mexico (ii) System Dynamics and the Multisectoral Simulation Tool—Bangladesh
		Who can provide lessons to improve the program throughout the program cycle?	Benchmarking	(i) International Benchmarking Network for Water and Sanitation Utilities (ii) The Public Sector Benchmarking Body—Ireland
Menu of Evaluations	Implementation and Follow-up	Have operational mechanisms supported the achievement of program objectives?	Process Evaluations	(i) Process Evaluation—Mexico (ii) Process and Implementation Analysis of the Welfare-to-Work Grants Program—United States
		Has the program performed from a comprehensive perspective?	Executive Evaluations	(i) Executive Evaluation of Structured Projects—Minas Gerais, Brazil (ii) Executive Evaluations—Colombia
		Has participation in the program resulted in planned impacts on target groups?	Impact Evaluations	(i) Small and Medium Enterprises—Mexico (ii) Rural Education—Madagascar
Quality Assessment of M&E		Is the information from M&E reliable for decision-making?	Assessment of Indicators	(i) Indicator Evaluation—Minas Gerais, Brazil (ii) Evaluation of Government Programs—Chile
			Assessment of Evaluations	(i) Randomized Control Trials Checklist—Coalition for Evidence-Based Policy (ii) Evaluation Report Standards and Rating Tool—The United Nations Children's Fund (UNICEF)

Method



Rationale: An Ex Ante Distributional Analysis (EADA) provides an analysis of both the unintended and intended consequences of a planned policy on the well-being of stakeholders. This is considered valuable in numerous scenarios; an EADA can be used to guide program choice among different interventions according to their likely impact on target populations. Another example is contemplating implementing a program that has been very successful in one country into another context; an EADA will in part answer if a given program design will have the same results for the same stakeholders. In the context of limited resources a front-end investment in an EADA can be very cost-effective. Finally, an EADA can serve to clarify policy debates and foster dialogue between by policy makers, and focus discussion on who will benefit or not from a proposed intervention.

Description: There are four analytical components at the core of the EADA method:

1. Objectives: What are the social development priorities?

   A first task in an EADA is to establish which impacts are to be analyzed—that is, distribution of what? The World Bank and Organization of Economic Cooperation and Development (OECD) have been proponents of conducting EADAs and have developed EADA tools which concentrate on a policy or program's impacts on multidimensional poverty in stakeholders (see examples). The definition of multidimensional poverty takes into account traditional income measures as well as variables associated with social capital and environmental sustainability—for example, prescribing to concepts of individual well-being as articulated in the Human Development Index. With this foundation, OECD and World Bank tools posit that identification of the specific impacts to be analyzed in a specific EADA project should for example be guided by National Development Plans, Poverty Reduction Strategy Plans (PRSP), and other policies reflecting government priorities.

2. Stakeholder Analysis: Which stakeholders will influence, benefit, or lose from the program?

   EADA stakeholder analyses test assumptions about the interests of social actors and their possible responses to the intervention. Stakeholders consist of agencies, organizations, groups, or individuals who have a direct or indirect interest in the intervention or its evaluation. The two basic categories are those who influence the intervention (positively or negatively), and those who are influenced by the intervention (negatively and positively). Typically, stakeholder analyses of the target groups of an intervention are the most rigorous and these may be disaggregated by a large number of characteristics such as household type, household size, ethnicity, gender, location, and occupation. The analysis of intra-household effects is also considered important. That said, it is considered very important to analyze the potential 'losers' of a policy or program, to ensure that an intervention does not cause unacceptable damage to specific stakeholders but also to estimate the likelihood of policy success in terms of political ownership and support for reform.

3. Institutional Analysis: What is the role of institutions in influencing impacts?

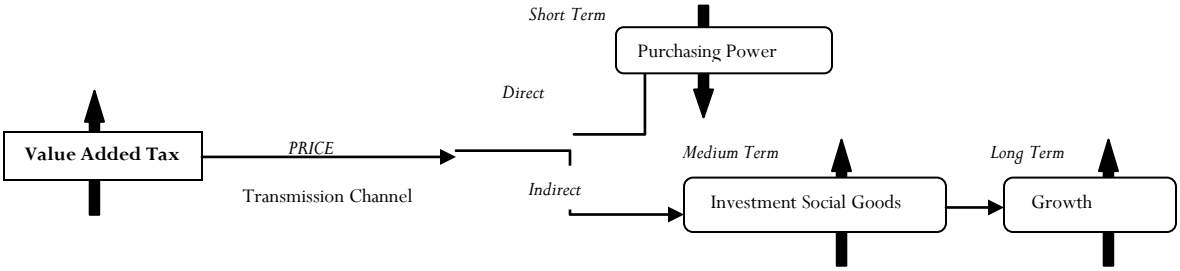
Method

Institutional analysis in EADAs aims to understand how institutions affect the impact of policy on poverty and the welfare of different households or groups. Institutions can be formal or informal and are characterized by organizational procedures and social norms. Put another way, institutions are the sets of rules in a society that govern individual and collective behavior. Institutions can influence impacts through a variety of avenues. For example, the role of formal institutions in implementation of intervention is key, which leads to analysis of the adequacy of institutional capacity and whether or not there is ownership of the intervention. Other interventions may focus on informal institutions such as market structures, which leads to analysis of market type (monopoly, oligopoly, perfectly competitive, and so forth) and whether there are distortions (restrictions to entry, collusion, and so forth). All these considerations will affect the impact of a given policy, and institutional analysis's role is to gather information that can perhaps preempt any challenges or otherwise enhance design and implementation.

4. Transmission Channels: What is the causality framework of a policy or program?

The EADA method posits in its causality framework that any program or policy (intervention) induces changes on stakeholders via transmission channels such as prices, employment, taxes and transfers, access to goods and services, authority, and assets. These changes are triggered by inputs provided through the intervention and lead to short-term and easily determined immediate outputs. These outputs, in turn, lead to intermediate outcomes and final impacts that are normally longer term (see figure 2).

Figure One: Basic Causality Chain in an EADA



In any EADA therefore the transmission channels (usually more than one) of the impacts of a program or policy must be ascertained. Once they are established EADA models analyze impacts along two key dimensions: (i) are impacts direct or indirect and (ii) do they occur in the short or the long term. For example, a direct impact is when government policy to increase the value-added tax translates directly into lower purchasing power for groups with a limited amount of disposable income. An indirect impact of that same policy is if the increase has a positive impact on tax receipts by the government and the new income is used to by the government to invest in social goods such as education, health, and the creation of jobs. Direct impacts on purchasing power will likely be felt in the very short term, while indirect impacts of improved service delivery and higher growth will take more time to materialize. Stakeholders might therefore feel both negative and positive impacts, but at different points in time.

Using these four core analytical components the EADA method allows researchers to make an integral assessment of likely outcomes and impacts on stakeholders, with particular emphasis on the target population of planned policy or program. EADA as a method does not focus on the use of particular techniques for analysis but encourages the use of a variety of analytical tools appropriate to the context,

including mixed methods approaches (quantitative and qualitative), vulnerability analysis, gender analysis, network analysis, and participatory methods.

Bibliography:

World Bank. 2003. A User's Guide to Poverty and Social Impact Assessment. Washington, DC: World Bank.

Organisation for Economic Co-operation and Development. 2007. Promoting Pro-Poor Growth: A Practical Guide to Poverty Impact Assessment. Paris: OECD.

Tool

Poverty and Social Impact Analysis (PSIA)—World Bank

The PSIA was introduced by the World Bank in 2002. It was a response to calls by external groups for the Bank to strengthen its assessment of the potential impact on poor populations living in a country before recommending that governments implement any specific economic reforms. These calls were rooted in the experience of 1980s and 1990s when economic policies—'structural reforms' recommended by World Bank and its partner, the International Monetary Fund---sometimes had been found to have negative impacts on the well-being of the poor. Between 2002 and 2007, 156 PSIAs were completed.

The PSIA tool at the World Bank aims to put the completion of an EADA within a wider framework that fosters policy discussion, longer-term monitoring, and evaluation and capacity building in its partner countries. A successful PSIA is defined as one that not only has provided a rigorous EADA analysis but also has influenced policy decisions. As such, World Bank experience has been that PSIAs can be implemented in a short timeline if limited to an analysis and discussion/feedback with a government, or a longer one if for example the EADA forms the foundation for an M&E framework.

The analytical structure of the PSIA complements the four core analytical components described in the EADA method (objectives, stakeholder and institutional analysis, and transmission mechanisms).

Potential enhancement and compensation measures: Investigation of potential options available to the government that may limit the negative impacts on the welfare of the poor or other target groups of the program or policy. Examples include alternative program designs, direct compensatory mechanisms, and implementation delays.

Risk Assessment: Risk assessment uses different techniques to provide an analysis of the four main types of risk identified by the World Bank: institutional risks, political economy risks, exogenous risks, and other country risks.

Note that researchers are free to choose what they consider the appropriate techniques for each analysis given the context, time, and data restraints of a specific PSIA initiative. To facilitate the production of high-quality PSIAs the World Bank has published a number of workbooks and guides to increase staff knowledge of techniques that can be used.

A recent evaluation of PISAs by the World Bank's Independent Evaluation Group (IEG) highlights that during 2002–07, PSIAs were successful in promoting a more holistic analytical framework for development that brings together economic and noneconomic analysis. Because of the PSIAs overt support for different research techniques, as articulated in the various guides and workbooks published by the World Bank, it is also deemed to have contributed to the use of both quantitative and qualitative evidence in analysis by staff, both in completing PSIAs and in other analytical work.

IEG's evaluation suggests that the greatest challenge facing the PSIA has been its lack of integration in both World Bank and country policy decision-making processes. Reasons for this include that (i) PSIAs are sometimes completed without a prior consideration of their desired objective in terms of concrete policy impact, (ii) there has been an absence of quality control in PSIAs, and (iii) the timing of the completion of the analytical component of PSIAs has not always been in line with both World Bank and government decision-making processes.

Bibliography:

Independent Evaluation Group, World Bank. 2010. Analyzing the Effects of Policy Reforms on the Poor: An Evaluation of the Effectiveness of World Bank Support to Poverty and Social Impact Analyses. Washington, DC: World Bank.

World Bank. 2003. A User's Guide to Poverty and Social Impact Assessment. Washington, DC: World Bank.

World Bank. 2008. World Bank. 2008. Good Practice Guide: Using Poverty and Social Impact Analysis to Support Development Policy Operations. Washington, DC: World Bank.

Method

Ex Ante Poverty Impact Assessment (PIA)—Organisation for Economic Co-operation and Development (OECD)

In 2005 the OECD commenced developing and piloting its own EADA tool. The PIA was rolled out in 2007 and had the overarching aim to promote the organization's pro-poor growth agenda. There were two specific objectives:

1. To harmonize existing frameworks and methods used by both staff and partner countries for ex ante investment appraisal and thereby reduce the burden of over-reporting.
2. To provide OECD staff and partners countries with a rigorous EADA tool that was flexible enough to be used in different contexts, with different resource, data, and time restraints.

An important part of the PIA's harmonization objective was to incorporate analytical components into the basic EADA methodology (objectives, stakeholder and institutional analysis, and transmission mechanisms). This would facilitate an analysis that reflected existing standards and frameworks in relation to poverty as whole endorsed by the OECD, along with more specific elements that the OECD had prioritized as a strategy for poverty reduction. As such the PIA includes an analysis of the following elements.

Stakeholder and Target Group Capabilities: The OECD's definition of poverty lists five capabilities required by individuals or groups to alleviate and overcome poverty: economic, human, political, socio-cultural, and protective-security. Accordingly, the OECD asks that a PIA assess a program or policy by its impact on the five different capabilities in stakeholders and specifically the target group. In the context of an EADA this analysis can be seen as an extension of stakeholder analysis.

Assessment of Results on MDGs and Other Strategic Goals: Here researches are asked to investigate likely contributions of the intervention to strategic objectives such as Millennium Development Goals, which have been endorsed by 192 United Nations member states and more than 23 international organizations. A PIA may also include assessment of a proposed program's contribution to other accepted goals that are of immediate relevance to the sector and stakeholders, such as national Poverty Reduction Strategies.

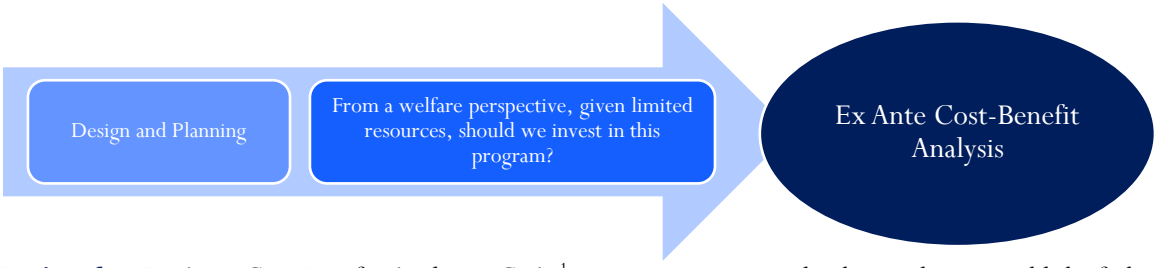
The OECD PIA is ideally completed within three weeks, at a cost of between US$15,000–40,000. It relies mainly on available secondary data. However, a key role of PIAs is to provide an assessment of the quality and availability of data, and as such may recommend more extensive data collection. The OECD recommends that a PIA be completed in the context of an already planned appraisal process, which may reduce both time and costs and enhance synergies between these different analyses. Dialogue with key partner institutions and stakeholders ideally begins with a kick-off meeting and ends with wrap-up meeting at the conclusion of a PIA where recommendations and suggestions are discussed. There may be other points of contact during the PIA completion for data collection purposes (interviews, focus groups, and so forth).

In 2008 the OECD task force mandated to design and pilot the PIA completed a report on the PIA experience over 2006–08 and made recommendations for its continued use as an EADA tool. Between 2006 and 2008, 20 PIAs were completed for programs in 13 different countries and in 6 different sectors. Challenges reported included a lack of incentives on the part of OECD staff to complete a PIA due to time constraints, and a lack of training materials to increase staff capacity in implementing the tool.

Bibliography:

OECD. 2001. *The DAC Guidelines.* Paris: OECD.
OECD. 2006. *Promoting Pro-Poor Growth: Harmonizing Ex Ante Poverty Impact Assessment.* Paris: OECD.
OECD. 2008. "Final Report by the POVNET Task Team on Ex Ante Poverty Impact Assessment." Paris: OECD.

Method



Rationale: Ex Ante Cost-Benefit Analysis (CBA)[1] is a quantitative study that seeks to establish if the benefits expected during the life of a project will exceed its costs. Both benefits and costs are expressed in monetary terms. A key final output of the analysis is the establishment of the net present value (NPV)[2] of a project; simply put this is the expected benefits minus costs of project during its lifetime. If the NPV is positive and high compared to the NPVs of practicable alternatives (such as other projects with the same objectives or not doing anything at all), then the CBA results support an investment in the program. The completion of a CBA is primarily used to establish a solid foundation for funding decisions in the context of limited resources. This ensures accountability to stakeholders that governments and other institutions are making evidenced-based decisions during allocations that are in line with policy priorities. The information collected for CBA analysis and its results can also play an important role in refinement of program design. CBAs in particular have been used frequently in the context of infrastructure investment and other types of economic investment.

Description: Ex Ante CBA has been a key economic appraisal method since the first half of the twentieth century. It has a large research and capacity-building literature and numerous experiences of implementation in different sectors and types of projects. In addition to standalone CBAs, many governments and different institutions have tried to incorporate key CBA concepts into other appraisal practices. The extensive use of CBA, although providing a rich environment for refining the methodology, has also led to concerns about CBA quality. Concentrating on standalone CBAs, we outline the key steps taken by researchers when completing a CBA.

1. Determining alternatives

Part of defining the scope of the CBA is the identification of the alternatives to which you will compare the NPV of the project which you are appraising. A key 'alternative' for comparison in CBA (often considered the minimal for a quality CBA) is what is called the counterfactual—that is, what would happen if the project did not exist, answering the basic question of the worth of intervening at all. Determining project alternatives includes finding those with similar objectives, target groups, operational implementation, and contexts to the project being appraised. An important part of this stage is ascertaining to what extent there is data available for completing the CBA. A lack of data is often the main limit on conducting rigorous CBAs that compare the NPV of the project being appraised to alternatives.

2. Determining whose benefits and costs will be included in the CBA (standing)

The concept of standing expresses the decision that researchers and other CBA stakeholders must make regarding whose benefits and costs will be considered in the CBA. Like Step 1, this is also a scope issue. Will a CBA include analysis of the state/province, local government, national, regional, or global levels? Beyond scope, determining standing is also an important ethical issue that reflects in essence a decision

[1] Ex post cost-benefit analysis and Cost Benefits Comparisons (ex ante CBA compared with ex post CBA) will not be discussed here.
[2] NPV is the one of the most common cost-benefit indicators. Others include ERR (economic rate of return), BCR (benefit cost ratio) (which is the present value of benefits/present value of costs), and NPV/k (where k is the level of funds available).

about whose well-being counts in appraising a project. For example, consider a CBA of an energy plant investment. It grants standing to a local community, which will both benefit from more jobs and be burdened by the cost of pollution. However, it does not grant standing or include an assessment of costs and benefits to the provincial community, which will burdened by the costs of pollution but perhaps not benefit from more jobs. This CBA may not factor in the full environmental cost of the energy project.

3. Determining the benefits and costs that will be analyzed

Determining what can be considered impacts of projects that are costs and benefits is a sector-specific exercise. For example, in the context of a proposed intervention to support agricultural businesses with advisory services, benefits to recipients may include increased profits due to increased production and marketing, and costs may include capital costs in new machinery and training of staff. In the context of the donors' appraisal of projects they fund, benefits may include improvements to donors' reputations and costs donors' contributions to project budgets. Note that costs and benefits will frequently be disaggregated according to the perspective of different groups who have standing, including producers, suppliers, intermediaries, governments, tax payers, and donors. What is a benefit for one group can be a cost for another group.

4. Monetizing benefits and costs

In CBA usually all benefits and costs are represented as a monetary value. This approach can be relatively straightforward if the benefits analyzed are, for example, investments in machinery. However when benefits include, for example, 'improved reputation' or costs are aspects of environmental degradation, then it can be challenging to express the results in monetary terms. In many cases things that are considered benefits and costs of a project will not be exchanged in a marketplace, which would allow researchers to ascertain a price for them based on knowledge of what people are willing to pay for a certain supply. Monetizing costs and benefits appropriately is one of the greatest analytical challenges in CBA. Some researchers assert that there are benefits and costs that cannot reliably by valued monetarily—for example the benefit of prolonged life provided by the health sector. For these, alternative methods such as cost-effectiveness analysis, which does not seek to monetize benefits, have been developed.

5. Determining the benefits and costs be over the lifetime of the project

This is an estimation and forecasting exercise of how costs and benefits will change over time. Often this change is significant. Consider the example of a CBA for a timber development project. In the first three years of the project benefits to producers may significantly outweigh costs because of revenue-producing activities such as logging and more efficient mills. In the following 4–6 years however revenues may stagnate as costs associated with the regeneration of land will increase. Estimation and forecasting is challenging and is based on many assumptions, including the project budget and the political, economic, and environmental climate within which the project will operate. Costs and benefits for projects following designs that have been implemented before in a similar context may be easier to predict. Similarly, forecasting may be easier for projects implemented in politically stable environments.

6. Discounting benefits and costs

At this stage benefits and costs are discounted in CBAs based on a rate that represents how they change value over time to provide present-value equivalents. Discount rates are typically set by a centralized government body and reflect the principle expounded by

economic and business orthodoxy that $1 now is valued more highly than 1$ at some time in the future. Reasons for this preference for consumption now rather than later are unclear. Studies have shown that they include individual impatience; the fact that individuals know they will die and so prefer $1 today than $1 in the uncertain future; and individuals' fear that a future cost or benefit will not occur because of a natural or manmade catastrophe.

7. Comparing project NPV with NPV of alternatives

After completing steps 1–6 the researcher can conduct a comparative analysis of the NPVs of the project being appraised, the counterfactual, and other similar projects and report the alternative with the largest NPV. The alternative with the largest NPV might represent the most efficient allocation of resources in the scope of the CBA. However, this is almost never possible to confirm because the CBA likely does not analyze all possible alternatives.

In addition to key steps 1–7 a fully developed and rigorous CBA will include:

Sensitivity analysis and distributional analysis

Sensitivity analysis is frequently conducted in CBAs because there is often still considerable uncertainty about predicted/forecasted benefits and costs, the monetization of benefits and costs, and the discount rate applied to these. Various scenarios based on fluctuations in key assumptions are presented, which help to clarify for decision makers how uncertainties will affect the NPVs. Distributional analysis is completed in the context of poverty and inequality reduction programs and is enabled in particular by an extensive analysis in Step 2 (see also method note on Ex Ante Distributional Analysis).

Bibliography:

Brent, R. 2006. *Applied Cost-Benefit Analysis*. Edward Elgar Publishing.

European Commission. 2008. *Guide to Cost-Benefit Analysis of Investment Projects.*

Independent Evaluation Group, World Bank. 2010. *Cost-Benefit Analysis in World Bank Projects.* Washington, DC: World Bank.

Jimenez, E., and H. Patrinos. 2008. "Can Cost-Benefit Analysis Guide Education Policy in Developing Countries." Policy Research Working Paper 4568. World Bank, Washington, DC.

Treasury Board of Canada. 2007. *Canadian Cost-Benefit Analysis Guide: Regulatory Proposals*. Ottowa.

Agribusiness Projects—International Finance Corporation

In 2007, primarily with the aim to provide guidance to its own staff, the International Finance Corporation (IFC) published a number of guides on conducting CBAs of investment projects it planned to contribute funding to. The guide discussed here pertains to CBAs of projects where IFC delivers technical assistance to agribusinesses, particularly SMEs, with the aim to improve their performance, corporate governance structure, and access to capital, thereby expanding employment and income generation in local communities.

The IFC guide for Agribusiness Projects articulates the following basic parameters for a good CBA, including standards for determining alternatives, standing, costs, and benefits. We present here a basic summary of the parameters set.

a) **Alternatives**: In a CBA the comparison is made between the anticipated state of the world with an IFC intervention to the anticipated state of the world without IFC intervention (that is, the counterfactual).

b) **Standing**: Cost and benefits are taken into account of those primary producers[1] that participate in the IFC project, primary producers that do not participate in the IFC project, market intermediaries, IFC, donors, and consumers and society.

c) **Total Project Costs:**
- IFC project budget
- Cash fee paid by participating primary producers (e.g. farmers) to IFC for the project
- Capital costs incurred directly by participating primary producers for the project (e.g. new trucks)
- Cash contribution paid by donors to IFC for the producers
- Direct funding to projects from donors

d) **Demographic and Fiscal Assumption:** IFC staff are guided to in CBA to determine:
- Population of the targeted district (or country), to estimate the project's NPV per capita.
- Use a discount rate reflecting the opportunity cost of capital, which can be based on the prevailing nominal interest rate in the country in which the project is planned. The discount rate will be used to discount the benefits and costs back to today's dollars.
- A PPP conversion factor, available from the World Development Indicators database, so NPV calculations can be adjusted for differences in purchasing power across countries.

e) **Project Benefits:** Project benefits are forecasted for primary producers (i) with the IFC project and (ii) without the IFC project (counterfactual). Benefits are forecast along dimensions of output, cost, and price.

Output: Based on the assumption that with the IFC project the average number of productive units is xx amount more than without an IFC project, the CBA forecasts the average number of productive units for the life of the project.

Output	Y	Y
1. Number of primary producers expected to participate in the IFC project		
2. Average number of productive units (e.g. hectares, cows, pigs) that participating primary producers will devote to the target primary good with the IFC project		
3. Average number of productive units that participating primary producers would have devoted to the target primary good without the IFC project		
4. Average annual yield (in output per productive unit) that participating primary producers would have seen without the IFC project		
5. Average annual yield (in output per productive unit, e.g. gallons of milk per cow, kilos of cassava per hectare) that participating primary producers will see with the IFC project		

[1] IFC typically provides technical assistance to primary producers and agro processors. Here we highlight basic guidance for CBA of a project focused on primary producers.

Cost: It is assumed that without the IFC project, cost per unit of output is xx more than with an IFC project.

Cost and Price	Y	Y
1. Cost per unit of output (e.g. cost per gallon of milk, cost per kilo of casava) that participating primary producers will face with the IFC project		
2. Cost per unit of output that participating primary producers would have faced without the IFC project		
3. Price paid by agro processors for the primary good with IFC project		
4. Price paid by agro processors for the primary good without the IFC project		
5. Percent of the price paid by agro processors for primary goods that will be captured by market intermediaries with the IFC project		
6. Percent of the price paid by agro processors for primary goods that would have been captured by market intermediaries without the IFC project		

Price: It is assumed that the price that agro processors would have paid for the primary good would have been xx higher without the IFC project (because of support to producers supply will increase and prices will fall). At the same time, because of the elimination of intermediaries,[2] it is assumed that primary producers capture a percentage xx higher of the price of primary goods as a result of the IFC project.

Based on these estimates of output, cost, and price above, it is possible to calculate the *Additional Profit of Primary Producers' on Primary Goods* (that is, profit with the IFC project minus profit without the IFC project) for each year. To conclude, *Total Project Costs* are subtracted from *Additional Profit of Primary Producers' on Primary Goods;* this produces *Net Benefits.* Net benefits are discounted using the appropriate rate in the country where the project is taking place, to calculate the NPV from the perspective of primary producers participating in the IFC project.

f) **Other Groups:** The IFC guide provides guidance on estimating the NPV for various groups that will be affected by the project, including:

Nonparticipating farmers: There are no benefits. However, the assumption is made that demand for the primary good is fixed and that the additional output sold by participating primary producers necessarily displaces the sales of nonparticipating primary producers. Thus in a CBA displacement rates should be included as costs for nonparticipating primary producers.

Displacement	Y	Y
1. What percent of the additional output sold by participating primary producers displaced sales of nonparticipating primary producers?		

Market intermediaries: There are no benefits. Costs are a reduction in profit as a result of primary producers selling directly to agro processors.

Donors: Costs are IFC and donor shares in project costs. Benefits include reputation benefits for donors.

g) **Project NPV:** Taking the net benefits (costs) of each affected party, the NPV of the project as a whole is calculated. A sample summary is given below.

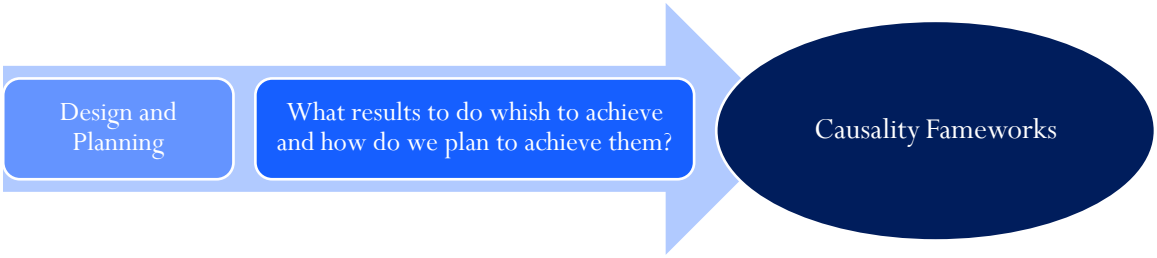
Summary of Project Benefits and Costs	
Affected Party	Net Benefits
Participating farmers	
Non-participating farmers	
Market intermediaries	
Agro processors	
Consumers/Society	
IFC	
Donors	
NPV Of Project	
NPV Of Project Per Capita	
NPV Of Project Per Capita, PPP	

Bibliography:

International Finance Corporation. 2007. *Cost-Benefit Analysis of Value Addition to Firms Agribusiness Projects: A Guide to Ex-Ante Evaluation.* Washington, DC: IFC.

[2] IFC works with participating primary producers to reduce their costs and capture greater value. A key feature is elimination of intermediaries so that primary producers can sell directly to agro processors.

Method



Rationale: The systematic use of a causality framework (CF) is an acknowledgement that every program is an experiment and that desired results cannot be guaranteed. Results depend on elements that are likely to change (variables) and their interrelationships. CFs encourages program managers and policy makers to systematically examine, document, and assign values to objectives and assumed variables, and to examine interrelationships between variables (for example, the connection between more food and better health). These steps bring stakeholders together to clarify the assumptions behind an intervention and enable the definition of indicators. With defined indicators, assumed causal relationships can be monitored and evaluated during implementation and follow up, and adjustments made if needed.

Description: CFS development has existed since the seventeenth century, when CFs were first used in the natural sciences to test hypotheses. In the public sector today, the development of a CF usually involves these steps:

Figure One: Generic Steps in Causality Frameworks

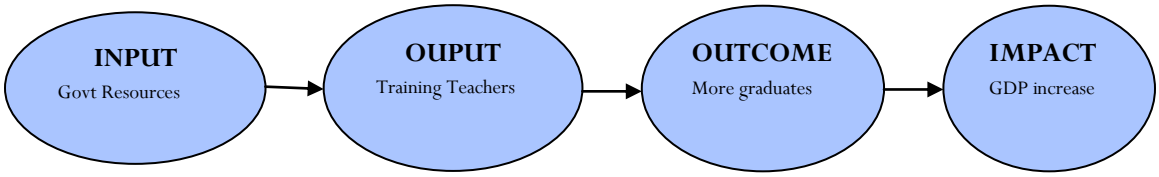
1. Problem Identification	Define the social problem (use of existing diagnostic info). Define the policy priority (link to National Develop Plan for example). Define the people affected (definition of target group).
2. Problem Analysis	Create a detailed explanation of the phenomena behind the social problem using a causal chain. Use a a visual map to make clear assumed causes and effects of the problem.
3. Program Objectives	Define program or policy objectives based on the problem analysis. Articulate desired states/realities. Use a visual map to clarify assumed causes and effects of the solution.
4. Program Structure	Clarify (using steps 2&3) the structure of the public policy intervention. Define the causal relationships the intervention will seek to influence. Define (with what inputs and activities) the objectives of the intervention.
5. Enabling Measurement	Clarify which results of the public policy intervention will be tracked through performance indicators. Show how progress on objectives will be tracked (i.e., through numerical indicators).

Beyond these generic steps there are various approaches to conceptualizing and articulating causality. Two types of CF presented here are the logic framework approach and systems dynamics.

Logic Framework: The logic framework is one of the best-known CF types used globally in public and private sectors and civil society organizations. Though applied slightly differently in different institutions, the it has been pivotal in developing a common language among program and policy managers. A key characteristic of the Logic Framework is that is that it expounds a linear chain of causality and progression of results. Figure One gives an example of a visual mapping exercise using the logic framework language of results (input,

output, outcome, impact) completed by the stakeholders of a policy to clarify a government's intervention structure.

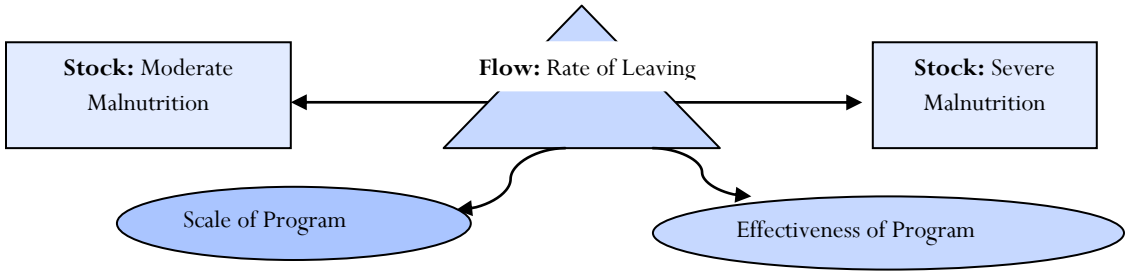
Figure One: Basic Logic Map



This language and linear vision of causality is formalized by production of a "logframe" table or matrix, which is used for management, including monitoring results and implementing changes, throughout the program cycle. In the logframe, rows are used to identify the vertical logic of the program; for example objectives/desired results are hierarchically classified as inputs, outputs, outcomes, and impacts. Columns are used for what is called horizontal logic. In this context, horizontal logic describes how the achievement of the objective will be measured or verified through performance indicators or target values, how this information will be obtained, and which external factors could pose a risk to success.

System Dynamics (SD): The SD approach encourages program and policy managers to consider a large number of variables affecting a social problem. There is no assumption of linearity as in the logic framework approach.[1] SD explicitly acknowledges the dynamic relationships between variables over time and allows for feedback effects and time lags. The language used to describe causality is stocks and flows: stocks describe the state of variables in the system experiencing the problem, and flows are variables that define the changes in the state of stocks.

Figure Two: Basic Causal Link in a System Dynamics Map for Nutrition



SD depends on visual maps, which identify the causality behind a social problem and the role of the public policy intervention in attempting to address this problem. The logic framework also involves many maps preceding a logframe but the matrix is the key summary of the program's causality structure and the basis for future monitoring and evaluation. In SD program and policy managers use maps throughout the program and policy cycle to see how stocks are changing as a result of public intervention directly affecting flows and how and if these public interventions need to modified. SD management and analysis is enabled by the equations underlying the maps, which give numerical values to relationships between variables (established when SD was set up). The equations make it possible to map changes in one variable through the whole system. As such SD relies on advanced computer software to analyze the dynamic relationships between variables over time. Although the underling programming is very complex, the software interface can be

[1] Other non-linear causality frameworks include Theory Based Evaluation, see Monitoring and Evaluation: Some Tools, Methods & Approaches, 2004, World Bank.

user friendly for public servants. Users can input new data online using maps, and the software can run the equations and calculations in the background and output only the results.

Bibliography:

Forrester, J. No date. "Road Maps: A Guide to Learning System Dynamics." Online at: www.systemdynamics.org.

Newman, J., M. Velasco, L. Martin, and A. Fantini. 2003. "A System Dynamics Approach to Monitoring and Evaluation at the Country Level: An Application to the Evaluation of Malaria-Control Programs in Bolivia." World Bank, Washington, DC.

European Commission. 1993. *Project Cycle Management: Integrated Approach and Logical Framework.*

World Bank. 2003. *The LogFrame Handbook: A Logical Framework Approach to the Project Cycle Management.* Washington, DC: World Bank.

Tool

The Matrix of Indicators—Mexico

In 2007 the Mexican government introduced the requirement that all federal social programs use the logic framework method to complete a Matrix of Indicators for Results (MIR). This was done in the context of an initiative to build a complete government M&E system lead by the Ministry of Finance (SHCP), Ministry of the Interior (SFP), and the National Council for the Evaluation of Social Development Policy (CONEVAL). In 2007 most federal programs had not reached the M&E system planning stage (when the logic framework is optimally done). However, it was deemed necessary to establish the underlying hypotheses of the social development portfolio and define the measuring indicators as a foundation for future M&E initiatives. Therefore, MIRs were completed for existing programs as well as new ones.

Since 2007 the MIR has been used

- To inform the federal budget; programs are required to submit an updated MIR to SHCP at the end of the fiscal year
- For program design improvements
- For program management improvements

Introducing the logic framework method in 2007 was a huge task. The Mexican government hired outside trainers to train over 1,620 officials in 65 workshops, and MIRs were set up in 389 programs (covering 70 percent of the federal budget). Since 2007 training has continued but on a smaller scale.

The completion of the MIR is the responsibility of M&E units within each ministry working together with program managers. CONEVAL is the technical leader of the Mexican M&E system in the social development sector and has produced norms, standards, and protocols outlining the requirements and best practices relating to the MIR. These are all available on its website for public servants to use as guidance. Note, however, that although CONEVAL, SHCP, and SFP have strict specifications for MIR content and presentation, there is only generic guidance provided regarding the logic framework method and public servants are encouraged to use other materials. The CONEVAL website contains links to training documents in the logic framework process developed by international entities including the World Bank, the Inter-American Development Bank, and the Chilean government.

Guidance by the Mexican government highlights six basic steps for MIRs:

1. Defining the social problem and the specific population affected (target group)
2. Analysis of the social problem including (i) a visual map of the perceived causes and negative effects and (ii) a diagnosis of the situation based on qualitative and quantitative data
3. Definition of objectives of a program
4. Identification of the activities
5. Elaboration of the analytical structure of the program using information from steps 1–4
6. Establishment of an MIR

Figure One: Basic Construction of the MIR

Matrix of Indicators for Results				
Impacts	Narrative	Indicators	Data Sources	Risks
Objectives				
Components				
Activities				

A finalized MIR asks programs to define in the rows:

1. *Impacts*: how the program is contributing to a high-order objective as defined by the ministry, sector, or National Development Plan
2. *Objectives:* the direct result the program aims to create for its target group/area
3. *Components:* the goods and services it will provide to achieve its objective
4. *Activities:* the principle actions and resources for each of the components

For each row, the program provides in columns:

1. A narrative description
2. Indicators of performance
3. Data sources for the indicators
4. External risks to performance

Bibliography:

Website: www.coneval.gob.mx

World Bank. 2009. "Mexico's M&E System: Scaling Up from the Sectoral to the National Level." ECD Working Paper Series No. 20. World Bank, Washington, DC.

Tool

System Dynamics—Bangladesh

Bangladesh suffers from one of the highest rates of malnutrition in the world. In 2008 the prevalence of moderate or severe stunting in children less than five years of age was reported by UNICEF to be 43 percent. The persistence of high malnutrition rates throughout South Asia despite numerous national and international interventions has brought increased attention to finding new tools that can make a meaningful contribution to addressing the malnutrition challenge.

It is in this context that two international initiatives, the South Asia Food and Nutrition Security Initiative (SAFANSI) and the initiative to End Child Hunger and Under Nutrition (REACH), started working together with the government in two districts of Bangladesh in 2008. The Multisectoral Simulation Tool (MST) founded on System Dynamics (SD) principles is one of flagships of their work. It is hoped that it may allow these districts to reach their nutrition goals in 5 years instead of 10 by answering three questions:

1. Which interventions are likely to have the largest impact?
2. What scale do they have to be operated at?
3. How much would the necessary interventions cost?

The first phase of the MST included the creation of a causal model (expressed visually in a stock-flow diagram) that linked public interventions to nutritional outcomes using the SD method.

Figure One: Main Components of the MST

Stocks	Adequately nourished children
	Moderately malnourished children
	Severely malnourished children
Flows	Scale of public policy programs
	Effectiveness of public policy programs

The stock-flow diagram created shows, taking into account initial conditions of the children's birth (birth weight for example), what affects the behavior over time between the main stocks. Nineteen separate interventions were chosen for their proven effectiveness through evaluations. Malnutrition is internationally recognized as a social ill caused and affected by a large range of variables including income, education, water infrastructure, and cultural practices regarding breastfeeding. As such the creation of the causal model involved the input and collaboration of national and international experts from different sectors.

The SAFANSI-REACH team established the relations between the stocks and flows and initial conditions (including feedback loops and time delays that can affect those relations). They were then able to assign quantitative values to each of these relationships. This was done by (i) using data from existing impact evaluations of nutrition, (ii) calibrating expert responses, and (iii) investing in data collection in the field.

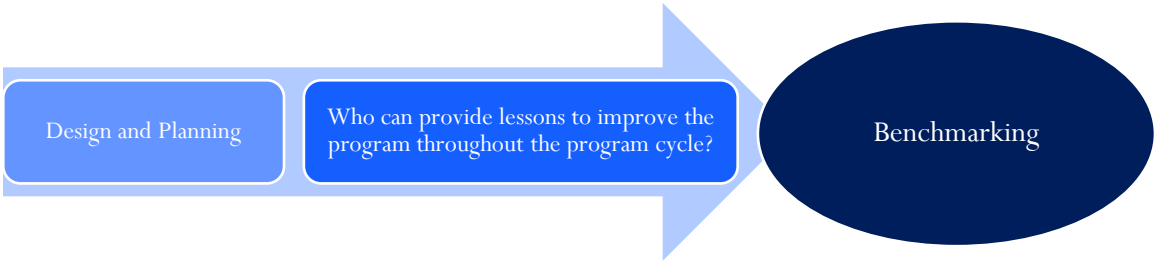
Created using a specialized software package for SD modeling, the MST can simulate different scenarios depending on the values assigned to the stocks and flows. For example, it is able to show how if one intervention (a flow) is scaled, then it will affect the number of adequately nourished, moderately malnourished, and severely malnourished children in one of the two districts. Hence the MST enables local public sector servants to know, for example, what interventions at what scale are required to reach desired targets. In this context the MST has been designed so that it is accessible to all local officials as well as national and international experts. The MST software program uses a Web portal, the stock flow diagram is the main screen for entering any change to a variable, and the software program writes the equations that will be used for the simulations.

Bibliography:

Website: www.worldbank.org/safansi

Newman, J. 2010. "Seven Advances Making It Easier to Work on Results in Development." World Bank, Washington, DC.

Method



Rationale: Benchmarking as a formal M&E tool goes beyond the widespread practice of simple comparison of, for example, performance indicators between different countries, ministries, or localities. To make this information valuable for decision-making, the acknowledged need is for in-depth analysis of the reasons underlying differences in performance. M&E benchmarking can be valuable because using lessons from the experiences and performance of others can inform (i) program and policy design, (ii) the setting of challenging but achievable targets, and (iii) meaningful M&E during the implementation and follow-up stages of the program cycle.[1]

Description: Benchmarking is a term used by almost all public and private institutions to denote a host of activities that involve comparison between different experiences. For example, the highlighting via publications and presentations of best practices that lead to learning and the instigation of change among institutions that aim to achieve similar results is a form of benchmarking. In this note however we concentrate on the use of benchmarking as a formal M&E tool in the policy cycle where it assumes a systemized and analytical nature. Figure One highlights the key steps that a formal benchmarking initiative of this nature would take.

Figure One: Key Steps in a Benchmarking Project

Step	*Description*
1	Selection of the desired results to be benchmarked
2	Selection of the 'points of reference' or 'standards' that will act as benchmarks or points of comparison to your own results
3	Data collection (can include collecting updated baseline information for your case)
4	Analysis of the quantitative identification of results gap magnitude and identification of policies and processes that may explain this gap
5	Implementation, which may involve adjustments to existing programs or support for the design and planning of new initiatives
6	Revisiting benchmarks and possible recalibration: this is done at the monitoring phase of a policy or during ex post evaluation feeding into the planning of the next cycle of the program or policy

In the crucial Step 2, the benchmarking project team must decide whom or what will be considered a benchmark and be the focus of their research and analysis. The process of choosing an appropriate benchmark (sometimes there is active collaboration between two institutions or 'benchmarking partners') includes ascertaining if there is enough data to complete the analysis.

> *Type:* The 'type' of benchmark chosen is usually articulated in the form of performance indicators. A simple dichotomy commonly used for categorization is benchmarks related to **impacts** (for example poverty rates) and benchmarks related to **processes** (for example average service

[1] Focusing briefly on incentives, it has been posited that one of the greatest potential achievements of benchmarking in the public sector is that it motivates public sector servants to improve because it activates employees' professional pride when compared to both internal and external organization with similar tasks, as such credible benchmarking projects can be a mechanism for peer-group control.

delivery times). Many sectors have also established key performance indicators relevant to their field that can be consulted at the beginning of a benchmarking project. (See the tool below, The International Benchmarking Network for Water and Sanitation Utilities).

Level: It is considered best practice to in a benchmarking project use a combination of both internal and external benchmarks. Internal benchmarks refer to results within one's own public sector administration, and external benchmarks refer to results in other countries and regions. Benefits of internal benchmarking include cost effectiveness, because in principle it is easy to gain access to internal institutional or public sector data. However, exclusively internal benchmarking may miss the bigger picture provided by combination with external benchmarks. Even the very best internal practices may not yield innovative ideas and solutions that can be a valuable input of external benchmarking.

Data Availability: After the desired type and level of benchmark have been determined the research team must decide which potential indicators are backed by sufficient quality data to perform analysis. Questions regarding the quality and thus reliability of secondary data used must be answered. A benchmarking project's budget must be reconciled with the need for primary data collection. Data availability can also depend heavily on the willingness to collaborate by different institutions in providing information. It is desirable to collaborate with benchmarking partners at the beginning of the project.

Focusing on step 4, there is no specific analytical method used in benchmarking; researchers have employed different qualitative and quantitative techniques. However, the availability of good baseline information is needed to ensure a valuable analysis stage of the project. Benchmarking assumes that the relevant values an institution wishes to compare as well as any known underlying contributing factors such as social and economic policies are already known. Without these any analysis of the size of the results gap and incorporation of ideas/approaches into programs and policies stemming from benchmarking will be unreliable and possibly lead to unwanted outcomes (Step 5).

Finally, using benchmarking as formal M&E tool involves an iterative follow-up component where benchmarks are revisited and recalibrated as a program or policy continues. Thus, chosen benchmarks become part of monitoring and/or evaluation process. The most useful benchmarking projects are not one-off initiatives. It is important to repeat benchmarking periodically in rapidly changing circumstances, and not only to monitor progress towards a chosen benchmark. Good practices can become dated quickly and the learning to be gained from peers' policy and program innovations should be seen as a continuous process.

Bibliography:

Besley, T., R. Burgess, and I. Rasul. 2003. *Benchmarking Government Provision of Social Safety Nets.* Washington, DC: World Bank.

Regional Environmental Centre for Central and Eastern Europe, 2007. "Guidelines on Progress Monitoring and Benchmarking." World Bank, Washington, DC.

Staplehurst, T. 2009. *The Benchmarking Book: How to Guide for Managers and Practitioners.* Elsevier.

Wynn-Williams, K.L.H. 2005. "Performance Assessment and Benchmarking in the Public Sector: An example from New Zealand." *Benchmarking: An International Journal* 12(5).

Tool

The International Benchmarking Network for Water and Sanitation Utilities

The International Benchmarking Network for Water and Sanitation Utilities (IBNET) is a global network administered by the World Bank's Water and Sanitation Program. Since its establishment in 1996, IBNET has grown into the largest publicly available water sector performance initiative. It collects, analyses, and makes benchmarking information accessible to more than 3,000 water and wastewater utilities (providing services to one quarter of the world's urban population) from 100 countries around the world.

IBNET was established partly because inter-utility performance comparison in the water and sanitation sector, though valuable, is limited. This is because the sector offers limited scope for direct competition. Firms operating in competitive markets are under constant pressure to perform. Water utilities, however, are often sheltered from competitive pressure. As a result, only some utilities are on a sustained improvement track; many others are falling behind best practices. Only efficient, financially viable utilities are able to respond to urban growth, connect the poor, and improve wastewater disposal practices.

The objective of IBNET is to support access to comparative information that will help to promote best practice among water supply and sanitation providers worldwide and eventually will provide consumers with access to high quality, and affordable water supply and sanitation services.

Water utilities are invited to submit their information to the IBNET database using a toolkit which includes :

- A set of core indicators representing industry standards
- A data list complete with data definitions
- A data capture system that also calculates the complete performance indicator set
- A method and public domain to share information on benchmarking

For example, IBNET outlines the performance indicators for service coverage shown in Figure One.

Figure One: Service Coverage Indicators

Water Coverage	Population with access to water services (either with direct service connection or within reach of a public water point) as a percentage of the total population under the utility's nominal responsibility	%
Water Coverage—Household Connections	Subset of Water Coverage	%
Water Coverage—Public Water Points	Subset of Water Coverage	%
Sewerage Coverage	Population with sewerage services (direct service connection)as a percentage of the total population under the utility's notional responsibility	%

IBNET also offers guidance for countries that wish to complete their own benchmarking projects using IBNET information. This guidance includes:

- A checklist of key steps for setting up a benchmarking project
- Examples of Terms of Reference for setting up performance benchmarking on the national or regional level
- Summaries of different more detailed techniques for defining and comparing indicators
- Summaries of different types of benchmarking common in the sector; process benchmarking, customer service benchmarking, and engineering-model company benchmarking

Bibliography:

Website: www.ib-net.org

Van den Berg, C., and A. Danilenko. 2011. "The IBNET Water Supply and Sanitation Performance Blue Book." World Bank, Washington, DC.

Tool

The Public Sector Benchmarking Body—Ireland

In 2002 and 2007, the Ministry of Finance in Ireland commissioned the Public Sector Benchmarking Body (Body) to examine the roles, duties, and responsibilities of jobs in the public service and, to compare these with similar jobs in the private sector, and to make recommendations on the pay rates for the public sector. The work of the Body was seen by the Irish government to address the public sector's need to

- recruit, retain, and motivate staff with the qualifications, skills, and flexibility required to exercise their different responsibilities
- support ongoing modernization of the public service
- underpin the country's competitiveness and continued economic prosperity
- ensure equity between the employees in both the public and private sectors

The 2002 and 2007 reviews each focused on different positions. For example, 109 health sector positions were reviewed in 2007, including Staff Nurse, Public Health Nurse, Clinical Nurse Manager II, Clinical Nurse Manager III, Assistant Director of Nursing, and Director of Nursing.

The specific areas of analysis of the Body were:

- Overall pay levels in the public and private sectors as well as pay rates for particular groups (such as clerical/administrative staff and technicians) and other identifiable groupings (such as graduate recruits)
- The overall pattern of pay rates in the private sector and employments across a range of firm type, size, or sector
- The way reward systems are structured in the private sector
- The value of public service pensions by comparison with pension arrangements available in the private sector

In order to answer these questions the following research was completed by external consultants commissioned by the Body:

i. A job evaluation was carried out using a point scoring system to assess when a public sector job could be compared to a private sector job based on factors related to skills, knowledge, leadership, accountability, and environment. Jobs with the same or approximately the same point score were considered to be comparable.
ii. A survey of 263 private sector companies was carried out covering approximately 36,400 employees and 4,100 jobs for purposes of comparison with similar public service jobs. Included in the survey were questions about annual salary, annual bonus, car or car allowance, medical insurance, other regular benefits or payments, date of salary review, percentage of salary increase at last review, overtime, pension scheme, share options, sick pay, hours worked, annual leave, and performance pay.
iii. A comparison of public service and private sector pensions was conducted.
iv. A comparison was conducted of remuneration in the public service and the private sector, discounting for higher public sector pensions found during (iii).

The 2007 review recommended increases in remuneration for 15 of the 109 positions examined. The comparison exercises showed that the salaries of only a small number of the public service positions examined were below private sector rates. In general, where remuneration was found to be below private sector levels, this was the case of some of the more senior grades examined. The annual cost of the increases recommended by the Body was in the region of 50 million euros on full implementation, or an average increase of approximately 0.3 percent in overall public sector pay costs.

To promote dissemination, stakeholder participation, and accountability, trade unions and other bodies representing employees were asked to review the findings of the Body and submit written comments and questions. These written remarks were discussed in 41 oral hearings.

Bibliography:

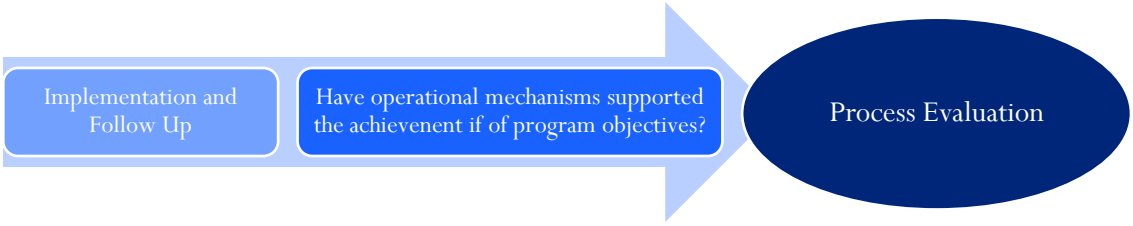
Website: www.benchmarking.gov.ie

Tool

Ministry of Finance, Government of Ireland. 2002. Report of the Public Service Benchmarking Body. Dublin.

Ministry of Finance, Government of Ireland. 2007. Report of the Public Service Benchmarking Body. Dublin.

Method



Rationale: Process evaluations (PEs)[1] aim to assess how and to what extent a program's operational mechanisms as identified in its design are supporting (or not) the achievement of the objectives of a program or policy. PEs are found as standalone evaluations or in broader assessments that seek to give an overall picture of program performance (see the Executive Evaluations method note below). The results of PEs can help program and policy staff to implement direct program improvements. For example, in the context of a program that is operating in different field sites, a PE may provide valuable information about how implementation in these different locales has differed and identify best practices that can be transferred to other sites. PEs can also highlight challenges in the original design of the program. For example if delivery systems are not working as envisioned, a PE may be the impetus for a larger overhaul of a project to ensure longer-term impacts are achieved. Finally, by documenting and assessing in a rigorous manner the operational components of a program, PEs can be an important complement to, for example, impact evaluations, which are more focused on the measurement of specific results and not so much on how these were achieved. In the context of international interest to transfer successful programs and policies it is particularly important that these actors have a view of the possible challenges a program has faced during implementation and what specific operational components have contributed (in perhaps unexpected ways) to outcomes and impacts.

Description: Evaluation of operational processes though one of the performance areas of a program often subject to analysis and reporting, is an evaluative focus which has not been developed with widely excepted norms, standards and research techniques, Notable exceptions include public health. That said, it is possible to identify two overarching evaluative components that characterize the majority of PEs: (i) comparison of the initial design of a program and the reality of implementation, and (ii) assessment of the extent to which operational processes (whether in the initial design or not) are supporting the achievement of program objectives. Focusing here on the second overarching component of PEs, Figure One highlights some of the key areas of analysis that PEs of poverty and inequality reduction programs focus on.

[1] Different names have been used for evaluations which are focused on a program's operations; other names which have been used are implementation evaluations and service delivery evaluations.

Method

Figure One: Key Areas of Analysis

Area of Analysis	Examples of Specific Processes Analyzed	Examples of Evaluative Questions
Processes related to targeting and enrollment of beneficiaries	Introducing program benefits to target groups	How effective are/were communication mediums (such as town hall meetings, pamphlets, websites, mailings, door-to-door visits, and registers from other social programs) in reaching target groups?
	Receiving requests for benefits	Are/were selection criteria for selection clear to applicants? Are/were existing application forms comprehensible and easily available to target groups?
	Selection of beneficiaries	Are/were selection criteria clear to public sector staff involved in selection? Is/was there effective coordination between different sites and/or organizations involved in selection?
	Enrollment of beneficiaries	Are/were existing registration forms comprehensible and easily available for selected beneficiaries? Are/were databases and IT systems able to effectively process beneficiary registries?
Processes related to production and delivery of services	Production of benefits	Are/were there sufficient inputs to produce benefits (machinery, personnel, etc.)? Are/were there appropriate mechanisms to estimate the quantity of output/products needed during implementation?
	Acquisition of benefits	Are/were bidding processes transparent? Are/were there clear protocols for communication and coordination with vendors?
	Distribution of benefits	Are/were benefits delivered in a timely manner to delivery sites? Are/were there sufficient inputs to deliver benefits (trucks, offices, personnel, etc.)?
	Receipt of benefits	Are/were receipt of benefits clearly documented and entered into the program database? Are/were benefits given to targets in a household (programs targeting young female children for example) by those who picked up benefits at delivery sites?
Process related to accountability	Beneficiary satisfaction	Are/were there mechanisms for beneficiaries to be able to communicate complaints? Are/were there mechanisms that actively ask for feedback from beneficiaries?

Issues that transverse individual processes and drive evaluative questions are the assessment of, for example, coordination and communication mechanisms between stakeholders and mechanisms for monitoring (both of which are facilitated by a good IT infrastructure and the existence of protocols). PEs tend to involve extensive descriptions of program operations based on data collection using mixed method approaches. This is in contrast to focus groups, site observation, surveys, and information from existing protocols, which are all data collection tools.

Bibliography:

Bliss, M., and J. Emshoff. 2002. *Workbook for Designing a Process Evaluation.* Atlanta, GA: Georgia Department of Human Resources.

Linnan, L., and A. Steckler. 2002. *Process Evaluation in Public Health Research and Interventions.* Jossey Bass Publishers.

Tool

Process Evaluation—Mexico

In 2009, as part of the Mexican federal M&E system, the National Council for the Evaluation of Social Development Policy (CONEVAL) was charged with the technical leadership of the federal results agenda. CONEVAL launched a standardized process evaluation (PE) tool for social development programs. Which program will be subject to a PE is decided through a negotiation that takes place every year in an annual evaluation planning exercise that includes the participation of CONEVAL, the Secretary of Finance, the Secretary of the Interior, and individual secretaries. Between 2009 and 2011 PEs were completed for programs in the Secretary of Social Development and the Secretary of the Environment and Natural Resources, among others.

CONEVAL has standardized its PE tool by producing a terms of reference (TOR) document. Individual M&E units operating in each secretary are required to use the TOR when commissioning an external evaluator for a PE. The creation and provision of extensive and detailed TORs is a key strategy that CONEVAL has employed since 2007 at the start of the federal M&E system in order to both regulate for and support quality in Mexico's federal evaluation agenda. For example, the TOR for PEs is 47 pages long and includes numerous matrices and tables that the evaluator is required to use throughout the evaluation. A significant consideration in the development of standardized PE in Mexico is the desire to conduct meta-evaluations over the long term that may give insights about implementation processes across government departments and lead to wider reforms.

Basic Infrastructure of the PE

The guiding framework of the Mexican PE tool is called the Model of Processes (Model). This is a representative list created by CONEVAL of the key processes in a social development program (each is described extensively in the TORs): (1) Planning; (2) Introduction of program to stakeholders; (3) Requests for benefits; (4) Selection of beneficiaries; (5) Production and acquisition of benefits; (6) Distribution of benefits; (7) Receipt of benefits; (8) Follow-up to ensure utilization of benefits; (9) Accountability to beneficiaries; (10) Supervision and monitoring; and (11) Others.

When completing an evaluation, evaluators are asked to classify individual program processes according to the Model provided. Once this is done the analysis and evaluation of classified processes has two components:

1. Evaluators are required to answer a set of questions that have been developed for each classification; they must do this for each individual program process within a classification. For example in Accountability (9) a question is: are there adequate mechanisms to have knowledge of beneficiary perception?

2. Evaluators are required to define indicators of efficacy and sufficiency for each individual process used in the program and report on these. Efficacy is defined broadly in the TOR as the extent to which a process accomplishes its goal. Sufficiency is defined by a list of so-called 'Minimum Elements'; basic characteristics based on the Model that a process must have for it to make a valid claim that it is indeed serving a defined function.

Finally, evaluators using the findings above are required to provide a global evaluation of implementation in the program. Specifically evaluators are asked to identify (i) opportunities for improvement in existing normative documents, (ii) bottlenecks, (iii) good practices, and (iv) general recommendations.

Note that in reference to the identification of "opportunities in existing normative documents," Mexico's federal law requires all social development programs to have both *Rules of Operation*, a document that contains basic justification, budgetary and operational information about a program, and a *Matrix of Indicators*, a document articulating a program's causality framework and indicators for monitoring. "Opportunities for improvement to existing normative documents" identified by

evaluators are supposed to be directed primarily at changes in these documents.

PEs are completed using secondary data such as existing norms and completed evaluations. In addition, primary data is gathered from a series of in-depth, semi-structured interviews with program staff at headquarters and field delivery offices.

Bibliography:

CONEVAL. 2009. "Modelo de Términos de Referencia para la Evaluación de Procesos del Programa." Mexico City.

Method

Process and Implementation Analysis of the Welfare-to-Work Grants Program—United States

In 1997 the U.S. government established the Welfare to Work (WtW) program, providing US$3 billion to 700 state and local grantees. These funds were intended to support programs working in high-poverty communities to assist the most disadvantaged welfare recipients and low-income parents make the transition from welfare to work. Grantees were given five years to make use of the funds. Although implementation was very heterogeneous, three general program types can be identified in the context of WtW.

- Enhanced Direct Employment—emphasis on providing participants with individualized pre-employment support, counseling, and case management, along with post-employment services for a year or more.
- Developmental/Transitional Employment—emphasis on skills development, often combined with transitional, subsidized, or community service employment.
- Intensive Post-Employment Skills Development—emphasis on improving both job retention and specific occupational skills, primarily by working with individuals after they start a job.

A process and implementation analysis was completed in the context of the so-called National Evaluation of WtW, mandated by the United States Congress, which included four different components (i) description of WtW, (ii) process and implementation analysis, (iii) cost analysis, and (iv) participant outcomes analysis.

In order to complete the National Evaluation of WtW, data was collected through two rounds of site visits (1999 and 2001) to 11 different implementation sites in different states. During these visits over 900 semistructured interviews were held with staff of grantee agencies and service providers, and focus groups were held with beneficiaries. Data was also collected from management information system data maintained by the programs on participants and services delivered.

The main components of the process and implementation analysis were as follows:

Description: Here the key characteristics of the each of the 11 study sites are described. Extensive analysis of institutional arrangements is provided in the context of large differences between programs. For example, many grantees rely on subcontractors such as community-based organizations to deliver services and employers are key partners.

Enrollment Processes: Here the effectiveness of targeting strategies is analyzed. Planned and actual participation rates for the 11 sites are compared and an analysis of participant characteristics is provided.

Services: Here five of the main services offered by programs are described and assessed. For each service participant rates at the different study sites are reported.

1) *Assessment of participants* (basic skills, professional interests, etc.) to determine appropriate WtW support and for assignment to specific work placements with employer partners

2) *Pre-employment preparation* including job search support through a case manager and job readiness workshops

3) *Education and training* delivered either directly by grantee or via referral to third parties such as universities or high schools (paid for by grantee)

4) *Transitional employment* targeted at individuals with serious problems, such as physical or mental disabilities and low basic education competency

5) *Post-employment services* including ongoing case management support for job retention and advancement and in some case wage supplements

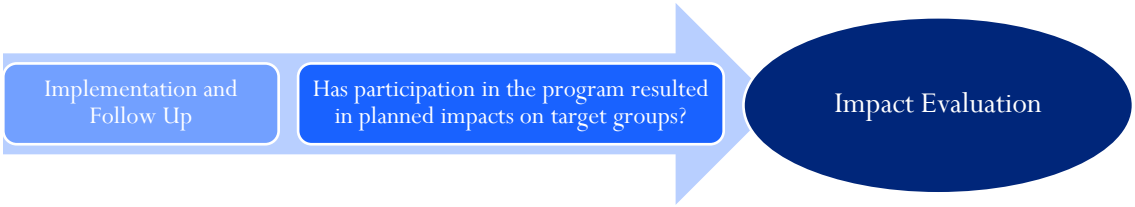
Findings of the process and implementation analysis included that eligibility criterion for participants in WtW sponsored programs were too restrictive. Criteria were amended later but the initiative suffered from this early barrier. Promising strategies highlighted by evaluators where the high collaboration with NGOs for service delivery and the partnerships with employers.

Method

Bibliography:

Nightingale, D., N. Pindus, and J. Trutko. 2002. *The Implementation of the Welfare-to-Work Grants Program.* Washington, DC: Urban Institute.

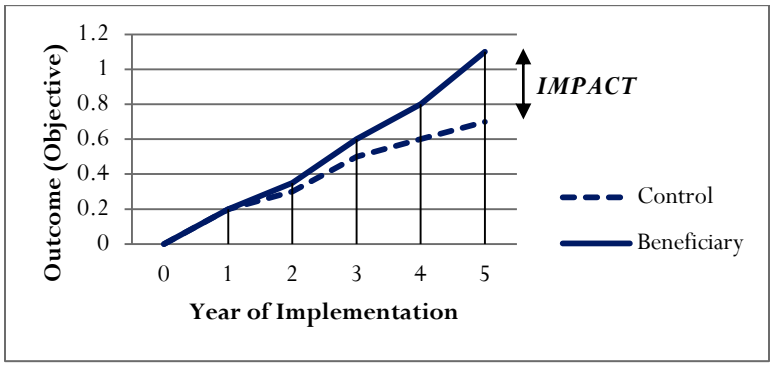
Method



Rationale: Impact evaluations are focused on analyzing if projects, programs, and policies have been successful in fulfilling their medium- and longer-term objectives. Put differently, impact evaluations analyze if the intended impacts have been achieved, estimating both the size of impacts and their distribution among segments of the target group. Impact evaluations have been used to inform a variety of decision-making processes. For example, they can support policy makers' decisions on how to allocate scarce resources in the context of potentially scaling-up an intervention, or in the context of a government austerity reform. Impact evaluations also can provide information to program managers that allow them to make improvements to the design of a program, for example when the evaluation shows that benefits are not reaching all segments of the target population. Finally, impact evaluations are widely considered to provide the most rigorous evidence of all evaluation types. A well-known benefit of this function is the role of impact evaluations in building the evidence-based case for conditional cash transfers as a way to reduce poverty.

Description: Impact evaluations differentiate themselves from other evaluations primarily by depending on control groups to gather strong evidence. Such data is used to ascertain whether or not observable impacts in target groups are attributable to a project, program, or policy as opposed to, for example, favorable external circumstances. Control groups enable evaluators to compare beneficiaries to groups of people with similar characteristics who were not part of the intervention. This allows creation of a 'counterfactual'—that is, what would have happened to beneficiaries if they had not received the intervention. The difference in the variable of interest (for example, the level of well-being if this was the objective of the intervention) between these two groups is the 'impact' of the project, program, or policy.

Figure One: Basic Impact Estimation Graph



Another characteristic of impact evaluations is that they require a large amount of qualitative and quantitative data in order to apply the econometric techniques (for example matching techniques, difference-in techniques, and regression discontinuity) required to provide a reliable analysis of impact. This implies that if relevant data are not available (for example, from a national censuses), then the impact evaluation team may have to design and implement specific surveys or other data collection methods. This in particular can drive up the costs of the evaluation. Furthermore, unlike some evaluative methods and tools, impact evaluations require extensive knowledge and skills in quantitative and qualitative research, as well as knowledge of the sector in which a program is operating. As such impact evaluations are most often completed by trained specialists.

Method

Several types of impact evaluation share the general characteristics discussed above. Evaluation categories reflect the time in the program cycle an impact evaluation is commissioned, prominent aspects of the design of the impact evaluation, and statistical precision and robustness. Table One summarizes two key types of impact evaluations used by organizations today, with a short analysis of the strengths and challenges of each design.

Table One: Two Key Types of Impact Evaluation

Type	Design	Strengths & Challenges
1. Experimental Design	Subjects (families, schools, communities, etc.) are randomly assigned to project and control groups. Questionnaires or other data-collection instruments (anthropometric measures, school performance tests, etc.) are applied to both groups before (creating a high-quality baseline) and after the project intervention. Additional observations may also be made during project implementation.	This type provides the most statistically valid and reliable results, if using large samples. Challenges include that it can be costly and labor-intensive, and its heavy reliance on good-quality primary data. Other challenges include that results may be difficult to interpret for policy and decision makers and results are not disseminated or incorporated in decision-making processes. Finally, randomization of beneficiaries is often considered difficult by governments for ethical and political reasons.
2. Quasi-Experimental Design	In this design beneficiaries of the project are either self-selected or are selected by the project implementing agency. The comparison group is not selected randomly (which is considered the most reliable way), but in a way that the evaluation team feels matches the characteristics of the target group as closely as possible. This "matching" is ideally done using statistical techniques such as propensity score matching. In other cases it may be necessary to rely on "judgmental" matching. For example, sometimes evaluators can construct a comparison group from similar types of communities from which project participants were drawn. Some quasi-experimental designs benefit from baseline data for both project and control group that has been collected before project implementation. However, many others do not and have to employ specific techniques to ensure the highest level of validity of results. Situations of non-perfect baseline data include the following; • baseline data is collected while the project has already been working for a while (for example at midterm) • baseline data is available only for beneficiaries but not control groups	The vast majority of impact evaluations fall into the category of quasi-experimental design and they present a continuum in terms of statistical validity of evaluation results. One of the issues influencing the statistical robustness of any single evaluation is the adequacy of the matching procedure. A second issue is the availability and quality of baseline data. If these issues can be addressed well than quasi-experimental design can be both very statistically robust and cheaper than experimental designs. For example reliance on good secondary data can provide very good but perhaps not excellent estimates of impact. Challenges experienced are the same as in experimental designs—that is, quasi-experimental designs can be costly and hard to disseminate and include in decision-making processes. In addition, quasi-experimental designs face a number of challenges related to the availability and quality of baseline information.

Sources: Bamberg 2009; Boyle et al. 2007.

Method

Bibliography:

Bamberger, M. 2009. "Institutionalizing Impact Evaluation within the Framework of a Monitoring and Evaluation System." World Bank, Washington, DC.

Bamberger, M., and A. Kirk. 2009. "Making Smart Policy: Using Impact Evaluation for Policy Making—Case Studies on Evaluations that Influenced Policy." Doing Impact Evaluation 14. World Bank, Washington, DC.

Boyle, P., K. Lyons, and M. Bamberger. 2007. "Strengthening Results-based Evaluation in Colombia." Social Impact Inc.

Fernald, L., P. Gertler, and L. Neufeld. 2008. "Role of Conditional Cash Transfer Programmes for Childs Health, Growth and Development: An Analysis of Mexico's Oportunidades." *Lancet* 371: 828–37.

Gertler, P., S. Martinez, P. Premand, L. Rawlings, and C. Vermeersch. 2011. "Impact Evaluation in Practice." World Bank, Washington, DC.

Impact Evaluation of Rural Education—Madagascar

Despite significant increases in primary school enrollment in Madagascar following reforms in 2002 and an influx of international financial support to shore up school resources, educational achievement in Madagascar has remained a challenge. Only 63 percent of children in grade 5 pass the primary-cycle exam, an assessment of the minimum level language and math knowledge presumed at this grade (Tan 2005). In response, an impact evaluation was conducted during 2005–07 to learn more about the effects of increased community monitoring (bottom-up approaches) and increased state monitoring (top-down approaches) on increases in educational quality through enhanced accountability of schools. The impact evaluation was completed by the Poverty Action Lab in partnership with the Government of Madagascar's Ministry of Education, Agence Francaise de Developpement, and the World Bank.

Madagascar's 2.7 million children attend 15,000 public primary schools. The impact evaluation included 3,774 primary schools in 30 public school districts. These 30 districts represented all geographic areas in the country, but were focused on schools with the higher rates of grade repetition. For the randomized evaluation three different interventions (programs) relating to increased community participation and accountability of schools were evaluated.

Intervention 1: District administrators received operational tools and training that included forms for supervision visits to schools, and procurement sheets for school supplies and grants. 1,314 schools participated in intervention 1.

Intervention 2: In addition to the district administrator (1), the subdistrict head was also trained and provided with tools to supervise school visits, as well as information on the performance and resource level at each school. 436 schools participated in intervention 2.

Intervention 3: In addition to the accountability support to district (1) and subdistrict heads (2), there was support for parental monitoring. A 'report card' was distributed to schools, which included the previous year's dropout rate, exam pass rate, and repetition rate. Community meetings were then held, and the first meeting resulted in an action plan based on the report card. One example of the goals specified in the action plans was to increase the school exam pass rate by 5 percentage points by the end of the academic year. This meeting was a launch pad for further parental monitoring through, for example, pressuring teachers to complete and communicate with parents student evaluations every few weeks. 303 schools participated in intervention 3.

4. Control Group: 1,721 schools were included in the impact evaluation that did not receive any interventions.

Data was collected by researchers on a variety of teaching practices through a school survey. In addition student attendance data was collected during unannounced visits and information on student test scores was collected from an achievement test administered independently.

The impact evaluation found that interventions based on a top-down approach, targeted at state bureaucrats at the district and subdistrict level, had minimal effects on administrators' behaviors or the schools and students under their responsibility. Although each tool (forms for supervision visits to schools and procurement sheets for school supplies and grants) was used by 90 percent of subdistrict heads and more than 50 percent of district heads, subdistrict heads visited their schools only slightly more often than those in the control group, an insignificant improvement. Teachers in both groups did not plan for lessons more, and no improvement in test scores was seen in the two years following the program.

Interventions based on a bottom-up approach targeted at communities (parents mainly), significantly improved teacher behavior. Teachers were on average 0.26 standard deviations more likely to create daily and weekly lesson plans and to have discussed them with their director. Test scores were 0.1 standard deviations higher than

those in the comparison group two years after the implementation of the program. Additionally, student attendance increased by 4.3 percentage points compared to the control group average of 87 percent, though teacher attendance and communication with parents did not improve.

Bibliography:

Lassibille, G., and T. Nguyen. 2008. "Improving Management in Education: Evidence from a Randomized Experiment in Madagascar."MIT Working paper. www.povertyactionlab.org.

Tan, J.P. 2005. "Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Interventions." Concept Note. World Bank, Washington, DC.

Tool

Impact Evaluation of Small and Medium Enterprises—Mexico

The Mexican national economic census of 1999 showed that small and medium enterprises (SMEs) made up 99 percent of enterprises in the economy, employed 64 percent of the workforce, and accounted for 40 percent of GDP. Despite these facts, many SMEs continue to lack the performance their larger counterparts are showing. This is demonstrated, for example, by very high exit rates of SMEs in Mexico. According to a 2001 study by the Ministry of the Economy only 35 percent of new Mexican SMEs remain in-country after two years. The potential reasons posited for this include constraints arising from poor access to finance and business support services, weak managerial and workforce skills, poor and inconsistent product quality, and imperfect information about market opportunities.

Over the past two decades the Mexican government has invested in different programs in order to support SMEs and alleviate these constraints. Between 2001 and 2006, the Mexican government invested US$13 billion in about 3.7 million SMEs. The quasi-experimental impact evaluation completed in 2007 by the World Bank aimed to establish if these interventions had in fact improved the performance of participating SMEs.

The period of analysis in the impact evaluation was 1994–2005. The evaluative team was able to profit from two sources of high-quality secondary data maintained by Mexico's National Statistics Office (Instituto Nacional de Estadística y Geografía). The first source was a large panel of annual industrial surveys (Encuesta Industrial Annual), which contained annual data on measures of firm performance such as sales, gross value of production, employment, total compensation, and income from exports, as well as some intermediate outputs that the programs may affect, such as technology transfers. The second data source was the National Employment Salary, Training and Technology Survey (Encuesta Nacional de Empleo, Salarios, Capacitación y Tecnología), which includes a module of questions on participation in major government SME support programs, including date of participation, duration, and type of services used. It was necessary to establish a new data set from these two surveys because the later was not annual.

Linking these two surveys enabled the evaluation team to include in their analysis the impacts of approximately 23 programs over a 12-year period. Around 1,500 firms had participated in one or more programs (the potential project group) and 1,100 stated that they had never participated in any program (the potential control group) with similar characteristics.

In their analysis of this dataset the evaluation team used a combination of fixed effects models and propensity score matching. These methods (i) addressed the biases that could be created due to the self-selection of SMEs participating in programs and (ii) ensure that reliable 'matching' took place between the project group and the control group.

The impact evaluation found that participation in certain types of support programs showed positive and statistically significant impacts on indicators of firm performance (value added, gross production, sales, hours worked), ranging from 4 to 17 percent. Some of the most successful programs included tax breaks for firms investing in technological advances. The results also indicated that some outcomes, such as employment and fixed assets, only showed positive effects after the third or fourth year following program participation, and that these effects increased as time went on.

Bibliography:

Allison, P. 2005. Fixed Effects Regression Methods for Longitudinal Data Using SAS. SAS Press.

Lopez-Acevedo, G., and H. Tan. 2007. "Evaluating Mexico's Small and Medium Enterprise Programs." World Bank, Washington, DC.

Lopez-Acevedo, G., and H. Tan. 2011. "Impact Evaluation of Small and Medium Enterprise Programs in Latin America and the Caribbean." World Bank, Washington, DC.

Method

Rationale: Executive evaluations provide policy makers with a 'snapshot' of a program's performance in a number of areas such design, strategic planning, operation, and results, which are considered together to be integral to the success of a program. Executive evaluations are typically completed in a relatively short time span with a limited cost. The focus is on providing an overall view of performance information rather than on detailed analysis. In particular senior policy makers will be interested in the 'overall' picture, which executive evaluations can provide for informing budgetary and planning decisions. As such, international experience shows that executive evaluations are implemented by ministries of finance, ministries of planning, and other agencies at the executive level with cross-ministry mandates. Another value added of executive evaluations is that they can play an important role in accountability processes both with legislative bodies and the public. Impact and process evaluations are often not considered adequate for analysis of accountability because they are considered too long and technical for nonspecialists. Besides aiming to provide a reliable overall picture of performance to policy makers, executive evaluations also have the explicit objective of communicating results to a wide range of stakeholders.

Description: A key characteristic of executive evaluations is that they have a standardized set of core components and questions that are completed for all programs subject to this tool (see Table One). Governments develop specific executive evaluation tools unique to their context with the goal of using these for a large number of programs over a longer period of time to inform decision-making. In this way there are very different from other evaluations, which often are designed specifically for one program and have a defined life time. This also speaks to the role of executive evaluations as communication tools. Governments that have implemented executive evaluations have also desired to create an evaluative tool that public servants and the public are familiar with, understand, and can expect to be completed. This contributes to a wider culture of 'results' in the government and public domain.

Table One: Performance Areas in Executive Evaluations

Type	Tool	Performance Areas
One	Program Rating Assessment Tool (PART), Office of Management and Budget, United States	Program Purpose and Design, Strategic Planning, Program Management, Program Results/Accountability
	Consistency and Results (Consistencia y Resultados), Council of National and Social Development Policy, Mexico	Design, Strategic Planning, Target Population and Coverage, Operation, Beneficiary Perception, Final Results
	Executive Evaluation of Structured Projects (Avaliação Executiva dos Projetos Estruturadores), Minas Gerais State Government, Brazil	Objective and Design, Planning, Program Management, Results
Two	Evaluation of Government Programs (Evaluacion de Programas Gubernamentales), Ministry of Finance, Chile	General Description and Objectives, Design, Organization and Management, Efficacy and Quality, Efficiency and Economy, Financial Health
	Executive Evaluations (Evaluaciones Ejecutivas), Department of Planning, Colombia	Design, Operational Management, Organizational Structure, Financial Management, M&E System, Efficacy, Efficiency

Method

Another key characteristic of executive evaluation is its reliance primarily on secondary information, such as existing evaluations and monitoring reports. Primary data collection is limited to interviews with key stakeholders such a program managers and other staff. On an operational level this also means that executive evaluations are completed in relatively short timeframes (2–3 months) and are cheap compared to other evaluations such as impact evaluations. Indeed one of the key components of executive evaluations is pointing out where reliable information is lacking; in some countries executive evaluations have played an important role in stimulating ministries to invest in other types of evaluation. As such, executive evaluations should be seen as a valuable complement to an evaluation menu that includes other methods such as impact evaluations or process evaluations—on which executive evaluation in large part relies. Similarly, some governments have introduced executive evaluation initiatives at the beginning of wider evidence-based government reforms. They answer the need to gain an overview of the content of, for example, the performance of the social development portfolio, and use this information to design further M&E policy.

Broadly speaking two types of executive evaluation have been used. The focus on providing an overall snapshot of performance is the same in each type and key performance areas for evaluation are very similar. Key differences relate mostly to the flexibility for the evaluator to choose his or her own analytical approach.

Type One executive evaluations are completed using an extensive guide published by the government entity in charge of the initiative. The guide includes predetermined performance areas and a standardized questionnaire for each performance area that evaluators must address. The evaluation is accompanied by detailed guidelines on how each question should be answered. Evaluators are asked to follow these guides in their analysis and assign a number or yes/no indicating their judgment. Performance areas are then assigned weights according to their importance to overall performance

Type Two executive evaluation guides are also published by government entities and like Type One have predetermined areas of performance and specific questions for each area. Type Two evaluations differ in that evaluators have more autonomy in deciding how to evaluate each area, and there is little or no guidance how each question should be answered. Type Two evaluations also can include the practice of assigning a number or yes/no to summarize program performance in an area.

The choice of type when designing an executive evaluation is dependent on many factors. Type Two is perhaps best implemented in contexts where there is good supply of evaluation specialists who have the skills and capacities to apply appropriate methods for each evaluation. In Type One, a government can make a large front-end investment in contracting evaluators to help develop the kind of extensive guidance needed, but after that it is possible for non-specialists within the government to implement the evaluation. This may be appropriate for example in a context where the market for evaluators is tighter. Because of its use of yes/no answers and metrics/numbers which translate in a very simple way findings, Type One may be more useful if it is the intent of the implementing institution is to place a large emphasis on using the executive evaluations for legislative or public accountability. That said, presenting Type One findings can give a skewed view if dissemination is not well thought out. Finally, note that the dichotomy of types presented here is useful for informative purposes, and is an accurate categorization of the so-called first generation of executive evaluations, the PART and EPG established in 2002 and 1997 respectively. However, more recent executive evaluation experiences such as the CYR (2007), E2 (2006), and AEP(2009), can be considered hybrids. They have drawn heavily from both the PART and EPG.

Bibliography:

Website: www.expectmore.gov

Fernando Castro, M., G. Lopez-Acevedo, and G. Busjeet. 2009. "Mexico's M&E System: Scaling Up from the Sectoral to the National Level." ECD Working Paper Series 20. World Bank, Washington, DC.

World Bank and Inter-American Development Bank. 2010. "Challenges in Monitoring and Evaluating: An Opportunity to Institutionalize." Fifth Conference of the Latin America and the Caribbean Monitoring and Evaluation (M&E) Network. Washington, DC.

Zaltsman, A. 2006. "Experience with Institutionalizing Monitoring and Evaluation Systems in Five Latin American Countries: Argentina, Chile, Colombia, Costa Rica and Uruguay." ECD Working Paper Series 16. World Bank, Washington, DC.

Tool

Avaliação Executiva dos Projetos Estruturadores—Minas Gerais, Brazil

At the end of 2009, the state government of Minas Gerais' executive M&E unit—State for Results (Estado Para Resultados, EPR)—commenced designing an executive evaluation, the Avaliação Executiva dos Projetos Estruturadores (AEP). The AEP was introduced as an analysis of programs initiated by the outgoing government. As such the main objectives of the AEP were to (i) gain an idea of the overall performance of programs identified as priorities at the start of the government's term and that had received extra support, (ii) inform strategic planning for the next government term, and (iii) pilot the AEP with the possibility of introducing it as a systematic evaluative exercise within the public service. The AEP is an example of a Type 1 executive evaluation (see method note). In 2010, 56 programs representing the portfolio of priority programs identified by the government in 2005 were evaluated.

The AEP is to be completed in five stages:

1. "Kick-Off" meeting between EPR evaluators and program management staff where the objectives and rules of the AEP are discussed
2. Interviews with relevant stakeholders and collection of secondary data
3. Analysis of findings and development of two reports: a full AEP answering all questions and providing justifications for these, and an executive summary with key findings and recommendations
4. Follow-up meeting by EPR evaluators with program management staff of the program, where draft reports are discussed
5. Finalization of reports based on results of the follow-up meeting

The AEP tool includes four axes of performance. Each of these has 6–9 specific questions (29 total in the AEP) to which the evaluator must assign yes or no with a clear justification of why the final yes or no was chosen. It uses a metric system to summarize performance of the program. All of the performance areas are weighted according to their perceived contribution to overall performance. In this tool weights were set by the EPR. Each question is worth the same amount in its component. This construction allows the AEP to assign numerical values for each performance area and also aggregate these to reach a global performance metric for a program.

Table One: Performance Axes and Weights

Performance Area	*Weight*
Objective and Design	20
Planning	30
Management	30
Results	20
Total	100

The EPR—as of 2011 renamed the Office of Strategic Priorities (Escritorio de Priodades Estrategicas)—beyond sharing detailed findings with program managers in stages 4 and 5 of the AEP process, has published a report documenting the general findings of the 2010 AEP initiative. Findings are presented from the perspective of the portfolio as a whole. For example, in the performance area of management, 45 percent of all priority programs received a yes to the question, "Has the program shown satisfactory financial and budgetary management?" The report also highlights programs who's AEP has shown they can be considered as "best practice" cases in an area of performance.

Because implementation of the AEP in 2010 was also seen as a large pilot, the report also highlights and discusses the experience that EPR encountered when implementing the AEP. Challenges included that not all program managers actively provided comments to AEP drafts. Finally the report reiterates the intent of state government of Minas Gerais to further refine the AEP and institutionalize its use within the public service.

Bibliography:

Governo do Estado de Minas. 2010. *Síntese dos resultados da Avaliação Executiva de Projetos Estruturadores*. Escritório de Prioridades Estratégicas.

Governo do Estado de Minas. 2010. *Manual da Avaliação Executiva dos Projetos Estruturadores*. Escritório de Prioridades Estratégicas.

Method

Evaluación Ejecutiva—Department of Planning Colombia

In 2006 the Directorate for Public Policy Evaluation (DEPP) charged with managing Colombia's M&E system SINERGIA for the Department of National Planning (DNP) started developing an executive evaluation tool, the Evaluación Ejecutiva (E2). Since its creation in 2002 DEPP has played a leading role in establishing a menu of evaluations that can respond appropriately to the information needs of the public sector. Currently three evaluation types are implemented by DEPP: impact evaluation, institutional evaluation, and the latest addition, the E2.

The E2 was piloted in two programs and launched officially in 2008. Between 2007 and 2011, seven E2s were completed for programs operated by various ministries (including the Ministry for the Environment and the Ministry of Social Protection), a municipality, the presidency, and the DNP. The E2 structure falls into the Type Two category of Executive Evaluations (see method note).

The E2 was designed as a rapid evaluation tool that could be completed in a short period of time (3–4 months) for a moderate cost (US$25,000–US$30,000). Its objective was to provide already existing programs with the information necessary for implementing agencies to take decisions regarding adjustments to program design and changes in their operational and financial structure so that programs are more efficient and more likely to achieve intended results. The E2 was seen in particular as an important tool to complement impact evaluation initiatives, which focused on end results only and not the operation of a program, tended to have a long time frame, and were costly.

The criteria for selecting a program to have an E2 is quite broad. Eligible programs include those which have recently begun; have had performance problems; have a large budget and wide beneficiary coverage; are seen as strategically important in their respective sectors; do not justify or need an evaluation of impact; and/or can be subject to program or budgetary adjustment decisions in the short term.

Although until now DEPP has been planning and contracting out the E2, it is envisioned that in the future ministries and other entities will be able to conduct E2s with little or no assistance from DEPP. To that end it has developed for use within the public sector a detailed guide for completion of the E2 and a format for the final evaluation report.

The E2 consists of seven areas of analysis that are detailed by guiding questions. The evaluator must assign a value to each question based on a 0–4 point scale. In addition the tool requires that evaluators complete a selection of matrices and tables.

Table One: Areas of Analysis in the E2

Area	Focus of Questions
Design	Program justification, objectives, components & activities, target group. Evaluator must complete a log frame.
Results	Effective coverage, existing evidence of results achieved
Inputs	Financial resources, budgets, execution rates, expenditures, costs
Management	Selection, production, delivery, quality control mechanisms
Strategic Management	Active use of monitoring and evaluation for program improvement
Organizational Structure	Roles, responsibilities, coordination and communication mechanisms of internal and external partners. Evaluator must complete organigram.
Information	Availability and quality of information for the evaluation and management of the program general. Evaluator must complete an Information Check Box.

The E2 relies on three main sources of information:

1. *Internal program documentation*: obtained from the institution implementing the program and central institutions like the DNP.
2. *Primary data:* obtained through conducting interviews, focus groups, surveys, etc.

Method

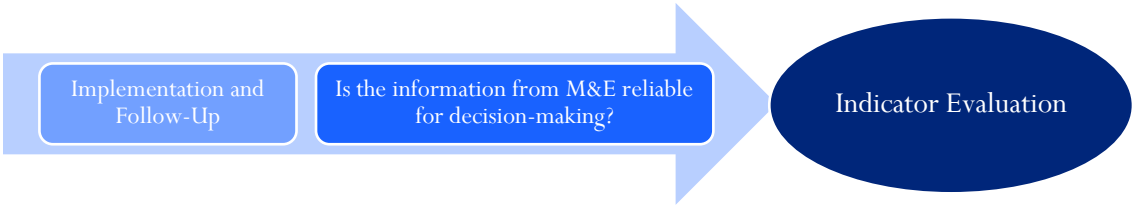
3. *Secondary information*: obtained from data, reports, evaluations, etc. conducted by others.

Bibliography:

República de Colombia. 2009. Evaluación Ejecutiva (E2). De Sinergia Lineamientos Metodológicos, Departamento Nacional de Planeación.

República de Colombia. 2009. Evaluación Ejecutiva (E2). De Sinergia Informe Final, Departamento Nacional de Planeación.

Tool



Rationale: The need for good performance indicators that can provide timely evidence of program performance. Such indicators are crucial for integrating evidence into decision making during program implementation and follow-up. In addition, performance indicators can significantly improve accountability mechanisms to the public and within the public sector if they are used as communication tools. The various roles that indicators play in evidence-based decision making imply that assuring their quality is of great importance. If the indicators are badly designed or do not rely on good data, then the information that is produced will be misleading and decisions might be made that will have negative effects on target groups. Furthermore, the programs that underlie indicators change and evolve. In this context a good development process (using, for example, causality frameworks—see method note) does not eliminate the need to return to indicators and assess their continued validity and quality at a later date. Developing and maintaining good indicators is an iterative process.

Description: The different principles for indicators shown in Figure One—known by their acronyms (SMART, SPICED, CREAM)—represent an informal international consensus by governments, international organizations, and NGOs about the characteristics of good quality indicators.

Figure One: Principles of Quality; Indicators Should Be…

SMART	
S	**Specific:** reflect those things the project intends to change/its objectives
M	**Measurable:** be precisely defined so that their measurement and interpretation is unambiguous
A	**Achievable and Attributable:** be achievable by the project
R	**Relevant:** produce information needed/utilized by stakeholders
T	**Time-bound:** describe by when a certain change is expected
CREAM	
C	**Clear:** precise and unambiguous, understandable to stakeholders
R	**Relevant:** appropriate to the subject at hand
E	**Economic:** available at a reasonable cost
A	**Adequate:** able to provide sufficient basis to asses performance
M	**Monitorable:** amenable to independent validation using quantitative and qualitative data
SPICED	
S	**Subjective:** include insights based on active experience by stakeholders
P	**Participatory:** development should involve a wide range of project stakeholders
I	**Interpreted:** easy to communicate to different audiences
C	**Cross-Checked and Compared:** revised by a range of different stakeholders and compared with other indicators
E	**Empowering:** defined and assessed using a process that allows groups and individuals to reflect critically on their changing situation and feel ownership over that change
D	**Diverse and Disaggregated:** reflect changes experienced by different groups: gender, ethnicity, geography, income level

The above three frameworks expound quite different views of quality but should be seen as complementary. SPICED recognizes the importance of the process of developing indicators, positing that quality is determined in large part by whom defines the indicators. CREAM and SMART focus on the technical aspects of design and more operational issues such as cost. Different institutions choose to emphasize one

framework more than others. SMART is the most widely known and is used frequently in the private sector, whereas SPICED is less well known but has gained a prominent role in the NGO community.[1]

The use of CREAM, SPICED, and SMART quality principles in training models, guides, and checklists is a well-established practice in institutions throughout the world. When these guiding principles of indicator quality are translated into more formal evaluations it has been in the context of (i) evaluations that seek to give an overall picture of program performance and, most recently, in (ii) standalone evaluations that focus solely on the quality of a selection of high-level indicators. These relatively new methodological developments are a promising contribution to the need for quality control of M&E to ensure the sustainability of the Results Agenda.

Figure Two: Evaluation of Indicators

Type	Description
(1) Integrated	Executive evaluations used by governments in Chile, the United States, Mexico, and Brazil are examples of how the assessment of the quality of indicators has been incorporated within a larger evaluation of program performance. They focus primarily on issues relating to indicator design. Specific evaluative questions are commonly placed in strategic planning components (see note on executive evaluations). In Chile and Mexico's executive evaluations, the assessment of the quality of indicators is completed in the context of a wider assessment of the program's causality framework, in these cases the logic framework. The evaluator assesses both horizontal and vertical logic (see note on causality frameworks). In the United States and Brazil's executive evaluations, questions addressing the quality of indicators are not linked to clearly defined causality frameworks and are more generally placed within the context of assessments of a program's performance in the area of strategic planning. Executive evaluations are applied to a single program, and thus the indicators analyzed are program-level indicators. Executive evaluations are often used as important communication tool for results to the public.
(2) Standalone	Standalone indicator evaluations focus on the 'quality' of the indicator in an integral sense. The qualities of design as well as management and dissemination processes are evaluated. Another way to view these is that they assess the entire supply chain of an indicator from the production of data, to the dissemination and use of an indicator within the public sector and the public. Standalone indicator evaluations tend to be quite detailed and technical. The primary audience of standalone indicator evaluation is public sector staff involved in managing M&E activities such as departments of planning. Because indicator evaluations include an integrated assessment of quality, optimal evaluation teams are multidisciplinary, made up of specialists in the area of program design, IT, and the sector the program is operating in. Standalone indicator evaluations are most appropriate for a selection of higher-level indicators, which for example have been developed in the context of a National Development Plan (NDP) or Poverty Reduction Strategy (PRS). These indicators are typically used over longer periods of time to direct policy and have a high profile within the government and the public.

Bibliography:

Kusek, J., and R. Rist. 2004. Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners. Washington, DC: World Bank.

Roche, C. 1999. Impact Assessment for Development Agencies: Learning to Value Change. Oxfam Novib.

Von Schirnding, Y. 2002. Health in Sustainable Development Planning: The Role of Indicators. World Health Organization.

[1] These principles are also frequently employed in assessment of other related concepts such as objectives, goals, and the more generic 'results'.

Avaliação Executiva dos Indicadores—Minas Gerais, Brazil

In the Integrated Development Plan of Minas Gerais (Plano Mineiro de Desenvolvimento Integrado—PMDI 2007–23) the state government of Minas Gerais established a selection of so-called Final Results Indicators (Indicadores Finalisticos—IFS) that would serve as monitoring, evaluation, and accountability tools throughout the government term. The role of the IFS is consolidated in Annual Results Agreements determined through a negotiated process between the executive and sector ministries, where each ministry commits to delivering on a selection of indicators. In 2008, there were 101 IFS.

The management of the IFS to date has been implemented by a specialized team, established within the state government of Minas Gerais' executive M&E unit, State for Results (Estado Para Resultados, EPR).[1] This team, the Unit of Indicators, in 2009 commenced designing an Indicator Evaluation to assess the quality of the IFS in the context of its continuous efforts to strengthen the portfolio. The tool developed by the Unit of Indicators was founded on the CREAM and SMART quality standards (this evaluation falls into the category Type 2 set out in the method note).

At the beginning of the Indicator Evaluation for each IFS, the evaluator is asked to complete a flow chart that documents the details of four basic processes in the 'life' of an indicator: the collection of data, registry of data, consolidation in a database, and production of the indicator. This exercise is supposed to lay a basis for the evaluation and be used to communicate bottlenecks and other findings to stakeholders when the evaluation is completed.

The evaluative component of the Indicator Evaluation is made up of four sections, each of which can have a number of questions that can be answered yes/no. Each section is also given a weight. To answer questions evaluators conduct interviews with program staff at headquarters and field sites and review existing operational protocols, communications plans, IT handbooks, and normative documents such as an NDP or PRS.

Figure One: Indicator Evaluation Structure

# of Questions	Section	Section Weight
13	Data Production	35%
6	Indicator Production	20%
11	Concept and Methodology	30%
4	Use and Communication	15%
34	Total	100%

Questions in each section pertain to assessment of each IFS based on the following elements:

- **Data Production**: Standardization of data collection, registration and consolidation procedures, database organization and security
- **Indicator Production**: Registration and memorization, checking procedures, disaggregation, historicity, accessibility, and timeliness
- **Concept and Methodology**: Measure's relevance, alignment with planning, clarity, methodological adequacy, sufficiency, ambiguity, comprehensibility, and duplicability
- **Use and Communication:** Stakeholders' ownership, dissemination

Once an Indicator Evaluation has been completed a presentation of results to the department involved is given by the Unit of Indicators and a recommendation report is delivered. The intent is that the Indicator Evaluation tool would continue to be used periodically to evaluate the IFS and that, based on recommendation reports, work plans are designed for departments and monitored.

In 2010 the Indicator Evaluation was applied in each of the 101 IFS. Findings included that 37.1 percent of IFSs received a global evaluative number between 80–100 percent and were thus

[1] The state government of Minas Gerais changed at the end of 2010. Currently the Office of Projects (Escritorio de Projectos), which will replace the EPR, is being structured.

deemed 'Sufficient.' 29.5 percent of IFS were considered 'Moderate,' achieving scores of 70–80 percent; 24.8 percent were considered 'Limited,' achieving 60–70 percent scores; and 8.6 percent of the IFS portfolio received a 'Weak' scoring of only 40–60 percent. Note also that the Unit of Indicators also applied the Indicator Evaluation retroactively and found that the portfolio of the IFS significantly strengthened in terms of quality during the period 2007 to 2010.

Bibliography:

Avaliador Governo do Estado de Minas. 2010. *Avaliação Executiva de Indicadores Manual*.

Tool

Evaluación de Programas Gubernamentales—Chile

Since 1997, the Budget Office in the Ministry of the Interior of the Government of Chile has implemented an executive evaluation tool called Evaluation of Government Programs (Evaluación de Programas Gubernamentales—EPG). The EPG aims to provide an overview of program performance that can be used to inform budgetary decisions and program improvements. Between 1997 and 2007, 199 evaluations were completed.

The EPG evaluates program design, organization and management, efficacy and quality, financial resources, and sustainability. The assessment of indicators is included in the component of the EPG assessing program design (this evaluation falls into the category Type 1 set out in the method note) and is grounded in the logic framework methodology. In the area of program design, evaluators are asked to first assess vertical logic (emphasizing the strength of the causality chain) and then horizontal logic (emphasizing indicators) and to enter their findings into a Final Report format provided by the Budget Office.

Focusing here on the evaluation of horizontal logic, evaluators are provided a general framework to guide their assessment. They are asked to analyze if program indicators adequately measure three key types of results during program implementation (processes, products, impacts) along four dimensions of performance (efficacy, efficacy, economy, quality).

Technical Notes published by Budget Office for EPG evaluators outline the key characteristics that define the three key types of results and four dimensions of performance as well as provide examples of these indicators.

Figure One: Results and Performance

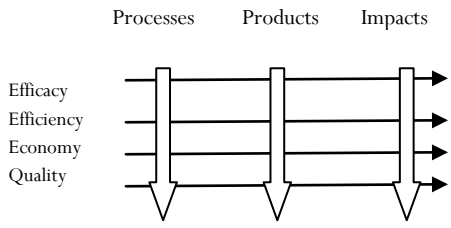


Figure Two: Results and Performance

Results	Description
Processes	Refers to activities linked to the implementation of the program
Products	Refers to the goods and services delivered to beneficiaries
Impacts	Refers to final results or objectives attributable to the program
Performance	**Description**
Efficacy	Indicators that show the extent to which objectives are being met
Efficiency	Indicators that show relationship between inputs (resources) and outputs
Economy	Indicators that show if financial resources are being mobilized to support objectives
Quality	Indicators that demonstrate beneficiary satisfaction

Complementing this framework for assessment, the Technical Notes articulate what are considered the basic requirements for a quality indicator including, relevance, clarity, economy, and comparability. Two key quality issues are highlighted:

- **Timeliness:** if performance indicators are reported in line with the times that different types of results are achieved
- **Data Sources:** if performance indicators come from reliable sources and if periodicity is appropriate

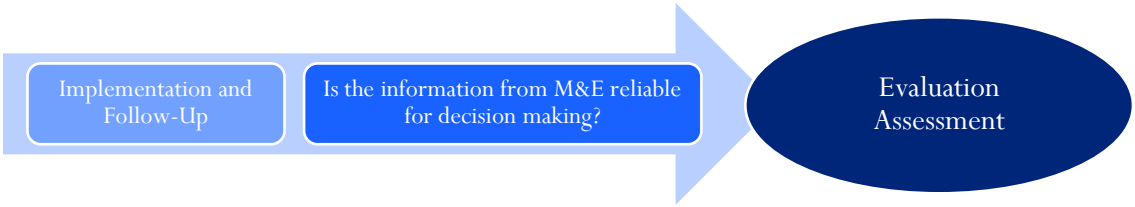
Using the overall framework, guidance on basic requirements, and highlighted quality issues, a program's indicators are assessed and findings reported. If a program's indicators do not cover a specific type of result or dimension the evaluator is asked to provide suggestions.

Bibliography:

Gobierno De Chile. 2008. *Notas Técnicas*. Ministerio De Hacienda, Dirección de Presupuestos.

Gobierno De Chile. 2011. "Términos Técnicos de Referencia: Informe Final." Available at: www.dipres.gob.cl (accessed September 15, 2011).

Method



Rationale: The ability to assess the quality of evaluation can be very important for organizations for different reasons. It allows organizations to have a better sense of how trustworthy evaluations are and subsequently, how much they should influence future policy. Secondly, organizations can assess how well their evaluation capabilities are developing, as they will be able to see if evaluation quality is increasing over time and organizations can get a better sense of where they need to focus their efforts. For example assessments can identify if there is consistent lack of information to feed good-quality evaluations and hence evidence-based decisions; this could lead to initiatives aimed at improving the availability of national statistics. Similarly, if evaluation studies are found to exist and be of high quality but are not integrated into decision-making processes, a government can create mechanisms to support such integration. Finally, evaluation quality assessments can be useful if an organization relies on outside consultants as an exercise to ensure contracting mechanisms focused on the skills of evaluators are working.

Description: Four types of assessment that organizations have implemented to assess quality in their evaluations are shown in Figure One. Though different, these types are not mutually exclusive and often overlap. The types differentiate themselves primarily by their focus and by highlighting the different standpoints through which evaluations are commonly assessed.

Figure One: Four Types of Evaluation Assessment

Type 1: Overall Report
Focus: the 'overall quality' of an evaluation, which is central and is determined by the quality of different components, including evaluation planning, analysis, and utilization for decision making. These assessments seek to establish if an evaluation has provided quality not only in terms of content but also of process. They are founded on standards created by different organizations where evaluation plays a foremost role, such as the Joint Committee on Standards for Educational Evaluation, the French Evaluation Society, and the Organisation of Economic Co-Operation and Development's Development Assistance Committee (DAC). Although these standards share a common approach to quality, they vary according to the focus of the organization that authored them. For example, the DAC guidelines are aimed at assessing evaluations of development assistance projects and the UK Evaluation Society has guidelines that focus on the behavior of an evaluator.
Type 2: Methodology
Focus: determining whether an evaluation has adhered to the accepted principles and quality standards of a certain methodology as defined by experts in this specific method and if, based on this, the evaluation's results/findings can be considered valid. Many methods such as impact evaluations have a large literature to draw from to facilitate the assessment of the application of the method. Often, executive individual evaluation units in organizations, such as the Independent Evaluation Group in the World Bank, assess the methodological rigor of a evaluation produced by sector departments in an organization. In addition some organizations specialized in evaluation such as the Western Michigan University Evaluation Center and the Coalition for Evidence-Based Policy, which have learning or activist mandates, have developed specific tools such as checklists that are available to the public.

Figure One (continued)

Type 3: Validity
Focus: assessing if the results of an evaluation can be considered valid according to seven core threats to validity: 1. **Objectivity (confirmability):** Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias? 2. **Reliability:** Is the process of the study consistent, coherent, and reasonably stable over time and across researchers and methods? 3. **Internal validity (credibility):** Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying? Are there reasons why the assumed causal relationship between two variables may not be valid? 4. **Statistical conclusion validity:** Are there reasons why inferences about statistical association (for example, between treatments and outcome/impact or the differences between project and control group) may not be valid. 5. **Construct validity:** Is there verification of the adequacy and comprehensiveness of the constructs used to define processes, outcomes and impacts, and contextual and intervening variables (moderators and mediators)? 6. **External validity (transferability):** Are there reasons why inferences about how study results would hold over variations in persons, settings, treatments, and outcomes may not be correct? 7. **Utilization:** How useful were the findings to clients, researchers, and the communities studied? This framework was developed and used to create an ex post checklist for assessing individual impact evaluations by Michael Bamberger, evaluation specialist and former Chief Sociologist at the World Bank.
Type 4: Utilization
Focus: assessing if an evaluation, throughout the process of design, planning, implementation, and follow-up, has been tailored to and focused on its utilization by intended users. For example, this model assesses if the organization that implements the program is committed to being evaluated, if it is clear by whom and how evaluation results will be used, and if and how stakeholders are updated throughout the evaluation process. Proponents of this model have been, among others, Michael Quinn Patton, former president of the American Evaluation Association, and the Western Michigan University Evaluation's Center. Based on this framework a checklist has been developed.

Organizations typically complete assessments that either combine aspects of the different types or utilize them at specific times depending on their needs. Tools developed based on these types are both checklists (often used prior to an evaluation to inform evaluation design, planning, and the development of Terms of Reference) and more fully developed assessments used for ex post assessment of evaluations, including some form of grading system and suggested improvements. In terms of which focus or type has played a more prominent role in organizations and governments worldwide, Types 1 and 2 represent widely accepted and mainstream approaches.

Bibliography:

Bamberger, M. 2007 "A Framework for Assessing the Quality, Conclusion Validity and Utility of Evaluations. Experience from International Development and Lessons for Developed Countries." Paper presented at the American Evaluation Association Conference of 2007.

Bamberger, M. 2009. "Checklist for Assessing Threats to the Validity of An Impact Evaluation." www.Bambergerdevelopmentevaluation.org (accessed September 23rd 2011).

French Evaluation Society. 2003. *Charter of Evaluation Guiding Principles for Public Policies and Programmes.*

Organization of Economic Co-Operation and Development (OECD), Development Assistance Committee (DAC). 2010. *Quality Standards for Development Evaluation. OECD, Paris.*,

Patton, M. 2002. "Utilization-Focused Evaluation (U-Fe) Checklist." Western Michigan University Evaluation Center.

Yarbrough, Donald B. et al. 2011. *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users.* Sage Publications.

Coalition for Evidence-Based Policy. 2009. "Which Comparison-Group ("Quasi-Experimental") Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact? A Brief Overview and Sample Review Form." Available at: http://www.coalition4evidence.org/.

Tool

Randomized Control Trials Checklist—Coalition for Evidence-Based Policy

The Coalition for Evidence-Based Policy (Coalition) is a nonprofit advocacy group based in the United States that supports the increased use of rigorous (preferably randomized) evaluations of program effectiveness within the U.S. government. A recent evaluation of the group's work during 2004–09 found that "the Coalition has successfully influenced legislative language, increased funding for evidence-based evaluations and programs, helped shape the Office of Management and Budget's Program Assessment Rating Tool (PART), and raised the level of debate in the policy process regarding standards of evidence" (Herk 2009).

In 2007 and revised in 2010 the Coalition published on their website a Checklist for reviewing to what extent the results of an evaluation using a randomized control trial method could be considered valid. The Checklist and falls into the Type 2 category of evaluation assessments. It is divided into four different components, each including individual questions and detailed guidance regarding how to answer each question. Below is a summary of the Checklist:

1. Overall Study Design
   - Was random assignment conducted at the appropriate level?
   - Does the evaluation have an adequate sample size?
2. Equivalency of Intervention and Control Groups
   - Were the intervention and control groups highly similar in key characteristics prior to the intervention?
   - Did few or no control group members participate in the intervention, or otherwise benefit from it (i.e., there was minimal "cross-over" or "contamination" of controls)?
   - Was outcome data collected in the same way, and at the same time, from intervention and control group members?
   - Was outcome data for a high proportion of the sample members originally randomized obtained (i.e., the study had low sample "attrition")?
   - In estimating the effects of the intervention, were sample members kept in the original group to which they were randomly assigned?
3. Outcome Measures
- Were "valid" outcome measures used—i.e., outcome measures that are highly correlated with the true outcomes that the intervention seeks to affect?
  - Were outcomes that are of policy or practical importance used—not just intermediate outcomes that may or may not predict important outcomes?
  - Where appropriate, were the members of the evaluation team who collected outcome data "blinded"—i.e., kept unaware of who was in the intervention and control groups?
  - Does the evaluation measure whether the intervention's effects lasted long enough to constitute meaningful improvement in participants' lives (e.g., a year, hopefully longer)?
4. Reporting of Intervention's Effects
   - If the evaluation claims that the intervention has an effect on outcomes, does it report (i) the size of the effect, and whether the size is of policy or practical importance; and (ii) tests showing the effect is statistically significant (i.e., unlikely to be due to chance)?
   - Does the evaluation report the intervention's effects on all the outcomes that were measured, not just those for which there is a positive effect?

In addition to the four main components above, the Checklist includes guidance on the following question: "How many randomized controlled trials are needed to produce strong evidence of effectiveness?" In order to assess this, the Checklist recommends finding evidence that shows:

Tool

- The intervention has been demonstrated effective, through well-conducted randomized controlled trials, in more than one site of implementation.
- The trial(s) evaluated the intervention in the real-world community settings and conditions where it would normally be implemented
- There is no strong countervailing evidence, such as well-conducted randomized controlled trials of the intervention, showing an absence of effects.

Bibliography:

Website: http://coalition4evidence.org

Coalition for Evidence-Based Policy. 2010. "Checklist for Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence." Available at: http://coalition4evidence.org.

Herk, M. 2009. "The Coalition For Evidence-Based Policy: Its Role In Advancing Evidence-Based Reform, 2004–2009." Paper for the William T. Grant Foundation. Available at: http://coalition4evidence.org/wordpress/wp-content/uploads/Indep-assessment-of-Coalitions-work-2009.pdf.

Tool

Evaluation Report Standards and Rating Tool—United Nations Children's Fund

In September 2004, the Evaluation Office (EO) of the United Nations Children's Fund (UNICEF), as part of its ongoing efforts to strengthen the role and quality of evaluation, released the Evaluation Report Standards. This document included 22 quality standards with detailed descriptions of each, as well as a rating tool allowing individual evaluations to be rated according to compliance with the standards. This simple rating tool was seen as an important component to establish transparent monitoring and evaluation within UNICEF.

The UNICEF standards and the rating tool fall into the Type 1 category discussed in the method note, and are based on the large body of internationally endorsed quality standards for evaluation that have been published by evaluation associations and international organizations.

Three key mechanisms have been used at UNICEF in order to ensure that the Evaluation Report Standards be actively adopted and used throughout the organization.

1. A requirement has existed since 2004 that every unit commissioning externally or completing internally an evaluation must provide as an addendum to the TOR the Evaluation Report Standards.
2. The rating tool in the Evaluation Report Standards is being used to determine which evaluations should be included in UNICEF's Evaluation and Research Database (ERD) which is used as a learning tool for stakeholders. This role for the Evaluation Report Standards came in the context of the observed need to enhance the quality of the large selection of documents posted on the ERD.
3. The EO carries out an annual quality review of evaluation reports submitted from all levels (HQ, region, country) whose methodology is based on the rating tool in the Evaluation Report Standards. The EO has also used these the Annual Reviews to analyze the trends in evaluation quality in UNICEF over time.

The Evaluation Report Standards include very detailed requirements of what an evaluation report should include for it to be useful in UNICEF, such as the completeness of the title page. They also include wider requirements such as the adherence of the evaluation to the OECD's general evaluation criteria.

Summary of UNICEF Standards
1.Completeness of title page and opening pages
2. Assessment of executive summary
3. Clear description of logic of program design
4.Clear description of role of UNICEF and contributors in evaluation
5. Justification for completing the evaluation at this time
6. Use of OECD/DAC evaluation standards
7. Clearly defined scope of for the evaluation
8. Inclusion of human rights–based approach
9. Assessment of results-based management in program evaluated
10. Transparent description of methodology of evaluation
11. Clear and appropriate choice of evaluation methodology
12. Description of stakeholder participation in the evaluation
13.Use of information from beneficiaries and non-beneficiaries
14. Use of ethical safeguards where appropriate
15. Inclusion of measurement of inputs, outputs, outcomes, and were possible impacts
16. Inclusion of cost-benefit analysis were possible
17. Clear discussion of relative contribution of stakeholders to the results of the program
18. Clear discussion of accomplishments, difficulties, and constraints for the program included
19.Evaluation conclusions that are based on data findings and that offer insights and solutions to identified challenges
20. Recommendations that are based on evidence
21. Lessons learned that pertain not only to the program evaluated but have wider relevance to other interventions
22.Completeness of annexes

Evaluations are rated by scoring each standard on a five-point scale (1–1.99 Poor, 2–2.99 Satisfactory, 3–3.99 Very Good, 4–5 Excellent).

Tool

A weighted average of these ratings is computed for an overall rating for the evaluation.

Bibliography:

UNICEF. 2004. *Evaluation Report Standards*. Evaluation Office, UNICEF.

UNICEF. 2004. *UNICEF Evaluation Report Quality Review 2006*. Evaluation Office, UNICEF.