

What's involved in "rigorous impact evaluation"?

IOCE proposes more holistic perspectives

Presented by Jim Rugh to NONIE Conference in Paris 28 March 2011
(Original PowerPoint presentation here pasted to Word)¹

Introduction: In this presentation I invite you to join me in a review the basics of:

1. Evaluation Design
2. Logic models
3. Counterfactuals
4. Context (simple-complicated-complex)
5. Evaluation Implementation

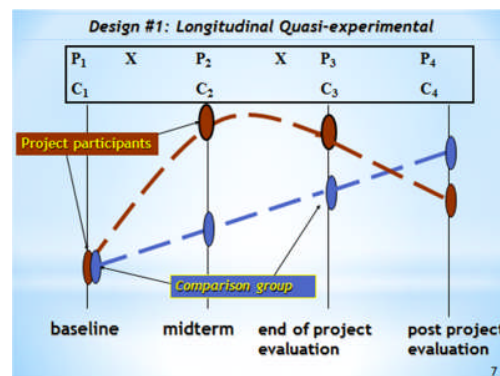
1. Evaluation Design

Here's a very short refresher course on **Evaluation (research) Design**.²

Let's assume a project's Final Goal has a quantifiable indicator. (Think of an example. The enrollment rate of girls in schools might be one.) Higher is better.

First of all: the key to the traditional symbols:

- X** = Intervention (treatment), I.e. what the project does in a community
- O** = Observation event (e.g. baseline, mid-term evaluation, end-of-project evaluation)
- P** (top row): Project participants
- C** (bottom row): Comparison (control) group

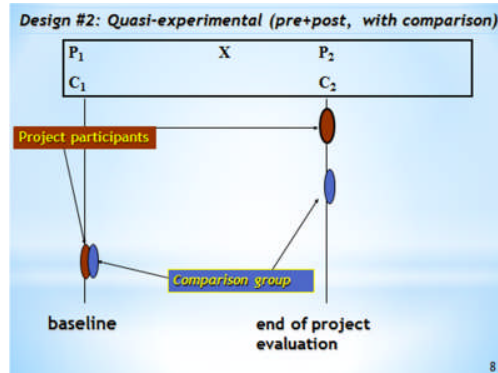


Design #1 is the most rigorous evaluation design or framework – at least in terms of how many observations are made before, during and after the project's interventions; also collecting longitudinal data on comparison group. This is the most comprehensive framework, revealing not only before-and-after and with-and-without, but also trends during the life of the intervention, plus an ex-post

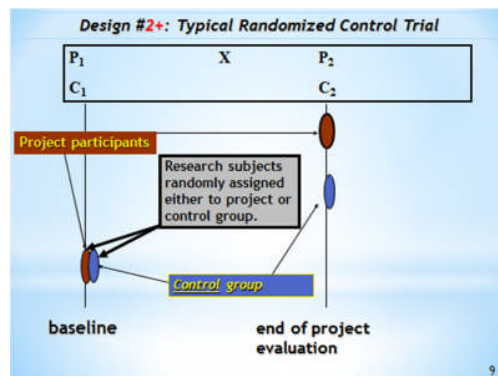
¹ Original PowerPoint presentation is accessible at http://realworldevaluation.org/Additional_RWE_materials.html.

² See the *RealWorld Evaluation* book (Sage 2006), or even the summary chapter available at www.RealWorldEvaluation.org for more detailed descriptions of these designs, and what needs to be done to fill in the missing data.

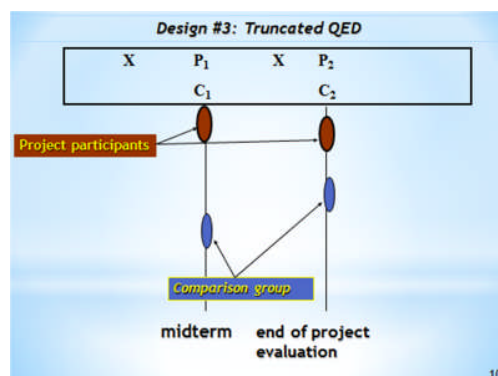
evaluation to measure sustainability of the impact. Comprehensive; but it is expensive to collect that much data that often.



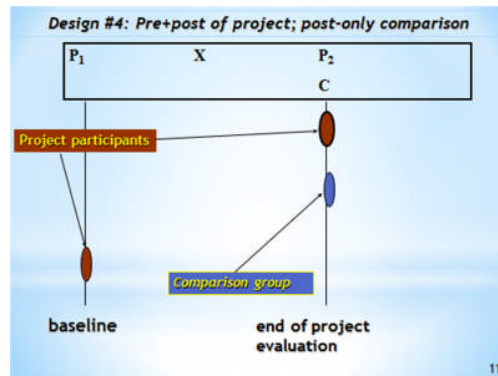
Design #2: Notice how much less information is available as various components of evaluation design framework are removed. In this case there is no trend data collected during the life of the project (not even a mid-term assessment) nor an ex-post evaluation. Yet this is the typical Quasi-Experimental Design (QED).



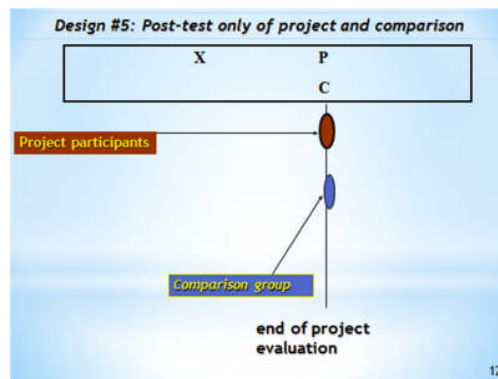
Design #2+: Even if there is an initial random selection of who should participate in the project and who should be held as "control" (the essential yet controversial element of Randomized Control Trials), notice how much information is missing without longitudinal data (including information about the quality of the intervention process) and ex-post evaluation.



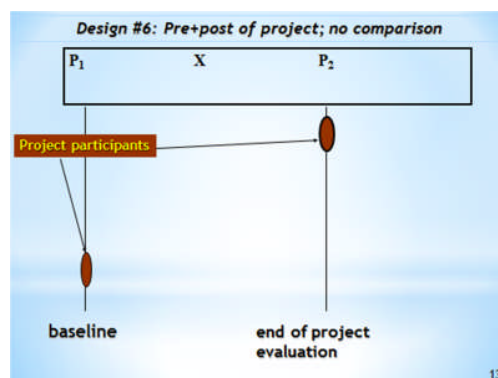
Design #3: In this scenario there was no baseline conducted at the beginning of the project, but there was a midterm survey (e.g. year 2.5 of a 5-year project) when relevant data was measured. This helps, but the remaining time is so short it will be hard to create much change.



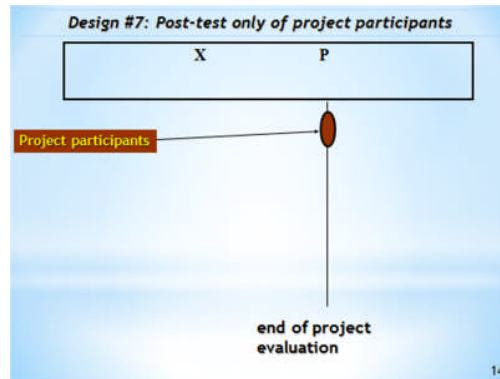
Design #4: In this design no baseline data was collected on the comparison group, though data was collected on a comparison group at the endline. Requires seeking adequate information to indicate what the conditions of the comparison group were at the time the project started. Look for relevant secondary data, key informants, use of recall methods, etc.



Design #5: No baseline. Begs the question on what the conditions were for both the project's participants and the comparison group at the time the project started. Look for relevant secondary data, key informants, recall, etc. to fill in the missing information.



Design #6: Pre-test and post-test of the project participants measures change in the indicator, but this design provides no information on the *counterfactual*-- what would have happened without the project. Needs to be supplemented by information from other sources on what changes occurred in the general population or in comparable communities during the life of the project.



Design #7: The status of the high-level outcome indicator only measured at end of project, only of project beneficiaries. Obviously the weakest evaluation design – yet by far the most common scenario in the real world (at least in international development projects). Even if the indicator in question is measured very precisely (e.g. with a very rigorous survey, or exhaustive qualitative methods) there was no direct measurement of what change occurred during the life of the project, nor any form of counterfactual. Very important to use complementary methods to obtain other information. In other words, we need to fill in missing data through other means:

- What change occurred during the life of the project?
- What would have happened without the project (counterfactual)?
- How sustainable is that change likely to be?

Table 1 A summary of Seven Basic Impact Evaluation Design Frameworks					
Key: T = Time period P = Project participants C = Control/comparison Group (Note) P ₁ , P ₂ , C ₁ , C ₂ = First and second and any subsequent observations X = Project intervention	Start of project [baseline / pretest]	Project Intervention (continues through life of project)	Mid-term evaluation	End of project evaluation [endline]	Post-project evaluation (some time after intervention ended) [ex-post]
	T ₁		T ₂	T ₃	T ₄
1. Longitudinal design with pretest (baseline), mid-term, posttest (endline) and ex-post observations of both project and comparison groups.	P ₁ C ₁	X	P ₂ C ₂	P ₃ C ₃	P ₄ C ₄
2. Pretest + posttest project and comparison group design i.e. before-and-after plus with-and-without comparisons. (Typical quasi-experimental design – ‘experimental’ if random selection of Participants and Control groups.)	P ₁ C ₁	X		P ₂ C ₂	
3. Truncated pretest + posttest of project and comparison groups where the initial study is not conducted until the project has been underway for some time (most commonly at the mid-term evaluation)		X	P ₁ C ₁	P ₂ C ₂	
4. Pretest + posttest comparison of project group combined with posttest (only) of comparison group	P ₁	X		P ₂ C ₁	
5. Posttest (only) comparison of project and comparison groups		X		P ₁ C ₁	
6. Pretest + posttest of project group (no counterfactual comparison group)	P ₁	X		P ₂	
7. Posttest (only) analysis of project group (no baseline or statistical comparison group)		X		P ₁	
Note: Technically a <i>control group</i> is only used in an experimental design (as randomization supposedly ensures there is no systematic difference in the distribution of subject characteristics between the two groups, i.e. selection <i>controls</i> for differences) and a <i>comparison group</i> is used in quasi-experimental designs where different selection procedures are used for the non-treatment group (sometimes called a “non-equivalent control group”).					
Source: Bamberger, Rugh, Mabry. <i>RealWorld Evaluation</i> , Sage 2006					

What kinds of evaluation designs are actually used in the real world (of international development)? Findings from meta-evaluations of 336 evaluation reports of an INGO.

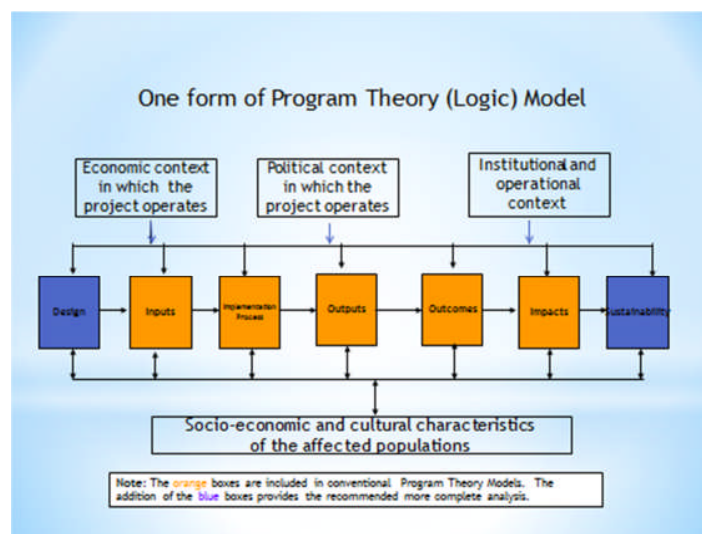
Post-test only	59%
Before-and-after	25%
With-and-without	15%
Other counterfactual	1%

The data in this table was summarized from four bi-annual meta-evaluations of evaluation reports from CARE projects in many countries. Colleagues familiar with other agencies report proportions of evaluations with no pre-test + post-test, nor counterfactual, are typically higher than the percentages reported here.

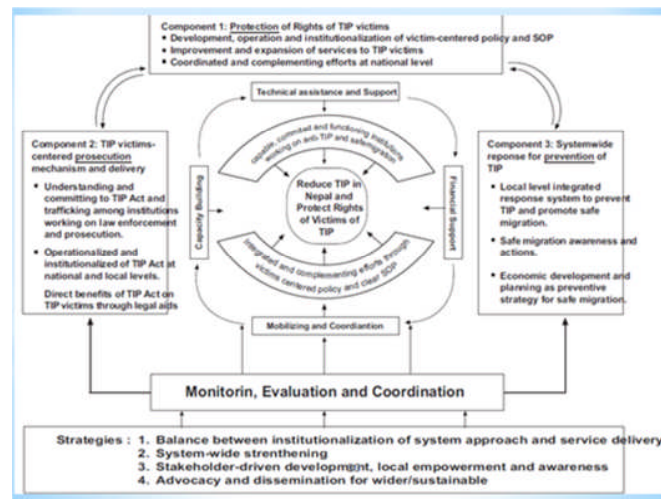
Even proponents of RCTs have acknowledged that RTCs are only appropriate for perhaps 5% of development interventions. An empirical study by Forss and Bandstein, examining evaluations in the OECD/DAC DReC database by bilateral and multilateral organisations found only 5% used even a counterfactual design. (Personal communication from Burt Perrin.)

While we recognize that experimental and quasi experimental designs have a place in the toolkit for impact evaluations, we think that **more attention needs to be paid to the roughly 95% of situations where these designs would not be possible or appropriate.**

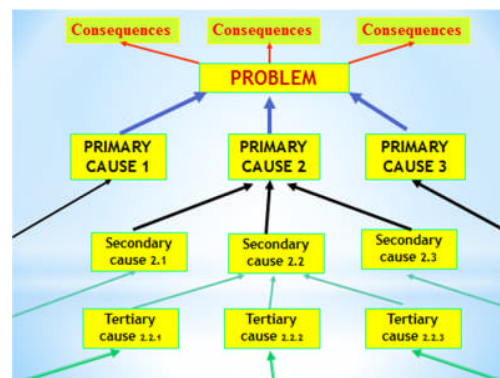
2. Logic Models



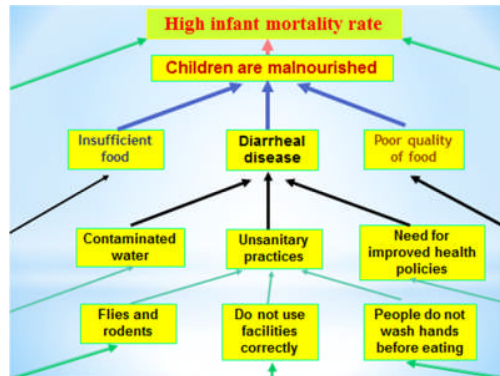
Typically project logic models are linear. This one above is an illustration of a somewhat more holistic model, where key external conditions are considered, since the success of the project is obviously dependent on those conditions. And if those external conditions change (negatively or positively) during the life of the project, the project’s own plans may need to be flexible and change accordingly. Note that the scope of the model could be expanded to include analysis of how the project was designed, and also evidence of sustainability of whatever impact might have been measured at the end of project implementation.



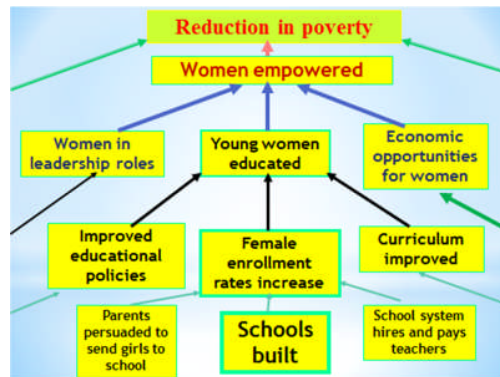
There are many creative ways to depict logic models. Here is one developed by the Asia Foundation for a Trafficking In People (TIP) program in Nepal. It includes three Components (sub-projects), with a variety of types of interventions. The challenge is to make it compressible enough to show the “big picture logic” of the program without becoming so complex that it becomes confusing. Another challenge is to know what aspects of a logic model like this should be tested by an evaluation.



Ideally project design should begin by key stakeholders (including intended beneficiaries) going through a process of identifying the main problem they want to address – the change they want to bring about. Then identifying primary causes of that problem, then secondary causes, tertiary causes, etc., (here illustrated with three at each level, though there could be fewer or more). There can, of course, be higher-level consequences; but the project should identify a major problem it will address that can reasonably be expected to be achieved during the project’s lifetime.



Here is an illustration of a “Solution Tree” using the example of a project with a goal of reducing diarrheal disease. Though that could make a plausible contribution towards the reduction of childhood malnutrition (and, at an even higher level, reduction in the infant mortality rate), it needs to be recognized that the quantity and quality of food eaten by these children are also vital for reduction of malnutrition.



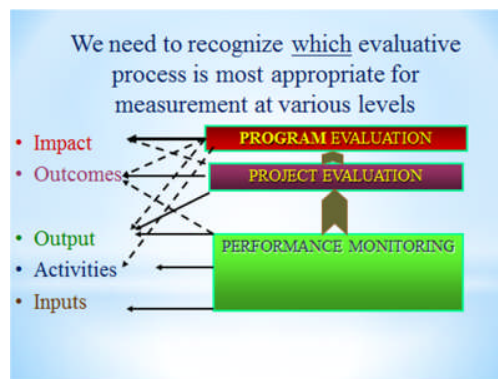
Here we illustrate this form of logic model with an education project – more specifically one that is focused on building schools. (Typically beginning at the bottom of the logic model and projecting hoped-for effects.) The designers need to recognize that more than classrooms are needed in order to achieve the outcome of increased enrollment of girls. And there need to be even other external assumptions fulfilled if higher outcomes and impact are to be achieved, such as girls completing quality education, leading to their long-term empowerment, and even (a higher desired consequence) that this will all lead to a reduction of poverty in these households and communities.



Here is an illustration of what a Program looks like: Three (there could be more) projects working in the same area, addressing the needs and rights of the same target population, collaborate together to provide the synergy needed to achieve higher level (program) impact.

As in this example, each project has an Outcome Objective at a level that can be achieved and measured during the life of the project. Though our own agency may choose to directly implement one or more of the projects, it may be more appropriate to assist one or more partners to implement complementary projects. If one assumes that someone else will take care of some aspect (e.g. Government Ministry of Education improving educational policies), it is important that we monitor that assumption. For the hypothesis developed from the problem analysis (in this example) is that all three of these causes must be addressed in order for the Program Impact to be achieved (young women completing quality education.)

Note that each project can be held *accountable* for achieving its Outcome Objective, but collectively they make *plausible contributions* to the achievement of the higher-level impact.



During the recent past there was a trend by USAID and some other agencies to rely more and more on Performance Monitoring and away from regular Project Evaluation, much less occasional Impact Evaluation. This has its limitations, as illustrated on this slide.

It is reasonable to expect a Performance Monitoring system to collect, aggregate and report quantitative data at the input, activities and output levels on a fairly routine basis. But to determine whether or not outcomes (intermediary effects) and impacts have been achieved requires more than an analysis of monitoring data -- they require a special study, a greater "stepping back" and broader, more holistic perspectives -- that's the unique role of evaluations.

The “Rosetta Stone of Logical Frameworks”

COMPARISONS AMONG DIFFERENT INTERNATIONAL AGENCIES’ TERMINOLOGIES for RESULTS/LOGICAL FRAMEWORKS³

	Ultimate Impact	End Outcomes	Intermediate Outcomes	Outputs	Interventions	
Needs-based	Higher Consequence	Specific Problem	Cause	Solution	Process	Inputs
American Red Cross	Program Goal	Project Impact	Outcomes	Outputs	Activities	Inputs
AusAID ⁴	Scheme Goal		Major Development Objectives	Outputs		Activities
CARE logframe	Program Goal	Project Final Goal	Intermediate Objectives	Outputs	Activities	Inputs
CARE terminology ⁵	Program Impact	Project Impact	Effects	Outputs	Activities	Inputs
CIDA ⁶ + GTZ ⁷	Overall goal		Project purpose	Results/Outputs		Activities
CRS Proframe,	Goal	Strategic Objective	Intermediate Results	Outputs	Activities	Inputs
DANIDA + Dfid ⁸	Goal		Purpose	Outputs		Activities
EIDHR ⁹		Overall Objectives	Specific Objective	Expected Results		Activities
European Union ¹⁰	Overall Objective	Project Purpose	Results	Activities		

³ Originally compiled by Jim Rugh in 1996 for CARE and subsequently for InterAction’s Evaluation Interest Group. Updated here to include additional agencies.

⁴ AusAID NGO Package of Information, 1998

⁵ CARE Impact Guidelines, October 1999.

⁶ Guide for the use of the Logical Framework Approach in the Management and Evaluation of CIDA’s International Projects. Evaluation Division.

⁷ ZOPP in Steps. 1989.

⁸ A Guide to Appraisal, Design, Monitoring, Management and Impact Assessment of Health & Population Projects, ODA [now DFID], October 1995

⁹ EU's Initiative for Democracy & Human Rights (EIDHR), 2008

FAO ¹¹ + UNDP ¹² + NORAD ¹³	Development Objective		Immediate Objectives	Outputs		Activities	Inputs
GTZ ¹⁴	Indirect Impact	Direct Impact	Use of Output	Outputs	Activities	Inputs	

¹⁰ Project Cycle Management: Integrated Approach and Logical Framework, Commission of the European Communities Evaluation Unit Methods and Instruments for Project Cycle Management, No. 1, February 1993

¹¹ Project Appraisal and the Use of Project Document Formats for FAO Technical Cooperation Projects. Pre-Course Activity: Revision of Project Formulation and Assigned Reading. Staff Development Group, Personnel Division, August 1992

¹² UNDP Policy and Program Manual

¹³ The Logical Framework Approach (LFA). Handbook for Objectives-oriented Project Planning.

¹⁴ GTZ Corporate Development Department, Results-based Monitoring: Guidelines for Technical Cooperation.2008

3. Alternative Counterfactuals

How do we know if the observed changes in the project participants or communities (e.g. income, health, attitudes, school attendance, etc.) are due to the implementation of the project (e.g. credit, water supply, transport vouchers, school construction, etc.) or to other unrelated factors (e.g. changes in the economy, demographic movements, other development programs, etc.)?

A counterfactual can be defined as answering the following question: What change would have occurred in the relevant condition of the target population if there had been no intervention by this project?

- Control group = randomized allocation of subjects to project and non-treatment group
- Comparison group = separate procedure for sampling project and non-treatment groups that are as similar as possible in all aspects except the treatment (intervention)



In recent years there has been an increasing emphasis on impact evaluation in public policy and development. There has been a focus on evidence-based policy and practice, drawing on approaches developed for evidence-based medicine. In both areas there have been debates about what constitutes 'rigorous evidence' and 'scientific approaches' to impact evaluation. Some people and organizations have argued exclusively for increased use of specific research designs – in particular Randomized Controlled Trials (RCTs). Others have argued that these designs are not always appropriate or feasible, and that we need a greater variety of approaches to doing more holistic, systemic, rigorous and useful impact evaluation in the real world.

So, are Randomized Control Trials (RCTs) are the Gold Standard and should they be used in most if not all program impact evaluations?

- Yes or no? Why or why not?
- If so, under what circumstances should they be used?
- If not, under what circumstances would they not be appropriate?

Evidence-based policy for simple interventions (or simple aspects): when RCTs may be appropriate:¹⁵

- Question needed for evidence-based policy → What works?
- What interventions look like → Discrete, standardized intervention
- How interventions work → Pretty much the same everywhere
- Process needed for evidence uptake → Knowledge transfer

When might rigorous evaluations of higher-level “*impact*” indicators *require much more than a simple RCT?*

- Complicated, complex programs where there are multiple interventions by multiple actors
- Projects working in evolving contexts (e.g. countries in transition, conflicts, natural disasters)
- Projects with multiple layered logic models, or unclear cause-effect relationships between outputs and higher level “vision statements” (as is often the case in the real world of international development projects)

There are other methods for assessing the counterfactual:

- Reliable secondary data that depicts relevant trends in the population
- Longitudinal monitoring data (if it includes non-reached population)
- Qualitative methods to obtain perspectives of key informants, participants, neighbors, etc.

There are situations in which a statistical counterfactual is not appropriate – even when budget and time are not constraints. *A conventional statistical counterfactual (with random selection into treatment and control groups) is often not possible/appropriate:*

- When conducting the evaluation of complex interventions
- When the project involves a number of interventions which may be used in different combinations in different locations
- When each project location is affected by a different set of contextual factors
- When it is not possible to use standard implementation procedures for all project locations
- When many outcomes involve complex behavioral changes
- When many outcomes are multidimensional or difficult to measure through standardized quantitative indicators.

Some of the *alternative* approaches for constructing a counterfactual:

A: Theory based approaches

1. Program theory / logic models
2. Realistic evaluation
3. Process tracing
4. Venn diagrams and many other PRA methods
5. Historical methods
6. Forensic detective work
7. Compilation of a list of plausible alternative causes
8. ...

B: Quantitatively oriented approaches

1. Pipeline design
2. Natural variations
3. Creative uses of secondary data

¹⁵ Based by work by Patricia Rogers

4. Creative creation of comparison groups
 5. Comparison with other programs
 6. Comparing different types of interventions
 7. Cohort analysis
 8. ...
- C: Qualitatively oriented approaches
1. Concept mapping
 2. Creative use of secondary data
 3. Many PRA techniques
 4. Process tracing
 5. Compiling a book of possible causes
 6. Comparisons between different projects
 7. Comparisons among project locations with different combinations and levels of treatment
- (For more details see www.RealWorldEvaluation.org)

4. Context

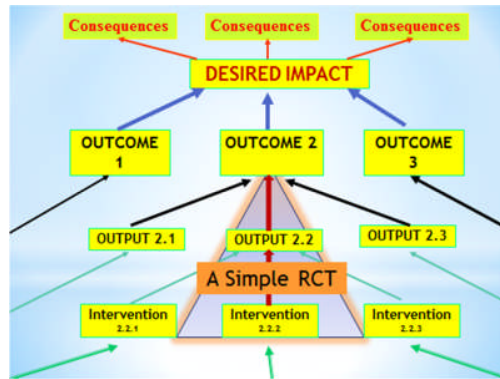
Different lenses are needed for different situations in the real world:

Simple	Complicated	Complex
Following a recipe	Sending a rocket to the moon	Raising a child
Recipes are tested to assure easy replication	Sending one rocket to the moon increases assurance that the next will also be a success	Raising one child provides experience but is no guarantee of success with the next
The best recipes give good results every time	There is a high degree of certainty of outcome	Uncertainty of outcome remains

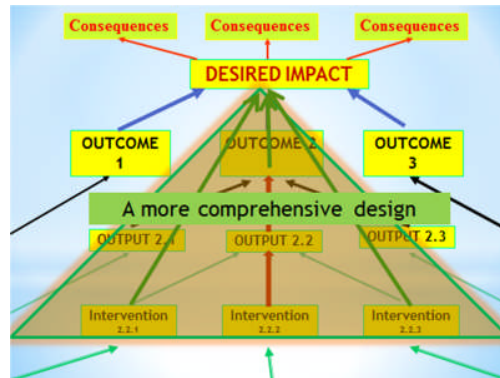
Sources: Westley et al (2006) and Stacey (2007), cited in Patton 2008; also presented by Patricia Rodgers at Cairo impact conference 2009.



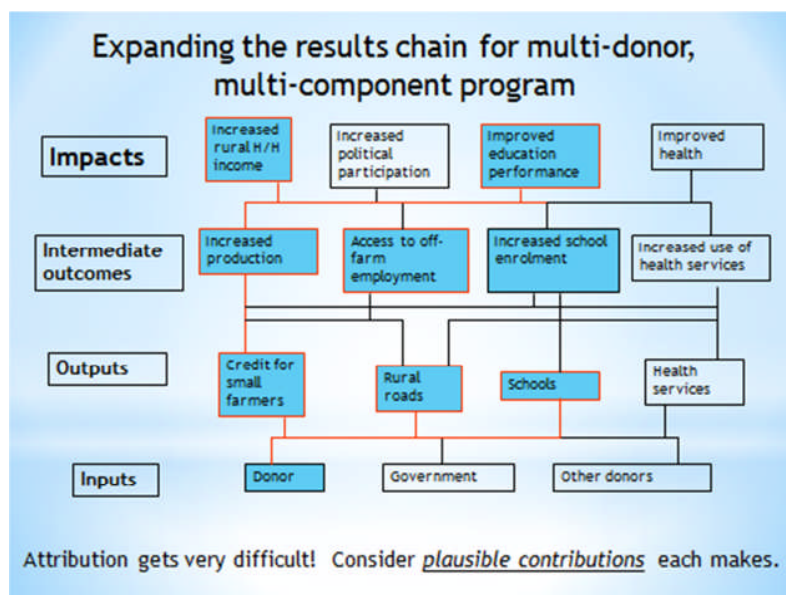
What's a conscientious evaluator to do when facing such a complex world?



Unfortunately, too often simplistic RCTs only conduct research on one intervention, and judge its “impact” by attributable change in a fairly near-term outcome, without adequate consideration of other causal chains.



A more comprehensive design should factor in an adequate number of causal streams to determine not only *which* but *what combination* of interventions by one agency or others, or necessary pre-conditions, need to be in place to achieve higher level impact.



Many programs are much more complex than the simple cause-effect hierarchy depicted in typical project logframes. Although a more comprehensive picture (such as the one on the above slide) is more realistic, especially at broader geographical levels where multiple agencies are involved, assessing *attribution* of each gets rather challenging. Ideally we can identify *plausible contributions* each actor makes to the achievement of higher level outcomes and ultimate impact.

5. Evaluation Implementation

OECD-DAC (2002: 24) defines impact as “*the positive and negative, primary and secondary long-term effects produced by a development intervention, **directly or indirectly**, intended or unintended. These effects can be economic, sociocultural, institutional, environmental, technological or of other types*”.

Is that definition limited to direct attribution? Or does it point to the need for counterfactuals or Randomized Control Trials (RCTs)?

So what should be included in a “rigorous impact evaluation”?

1. Direct cause-effect relationship between one output (or a very limited number of outputs) and an outcome that can be measured by the end of the research project? → Pretty clear *attribution*.
... OR ...
2. Changes in higher-level indicators of sustainable improvement in the quality of life of people, e.g. the MDGs (Millennium Development Goals)? → More significant. But assessing *plausible contribution* is more feasible than assessing unique direct *attribution*.

Rigorous impact evaluation should include (but is not limited to):

- 1) thorough consultation with and involvement by a variety of stakeholders;
- 2) articulating a comprehensive logic model that includes relevant external influences,;
- 3) getting agreement on desirable ‘impact level’ goals and indicators;
- 4) adapting evaluation design as well as data collection and analysis methodologies to respond to the questions being asked;
- 5) adequately monitoring and documenting the process throughout the life of the program being evaluated;
- 6) using an appropriate combination of methods to triangulate evidence being collected;
- 7) being sufficiently flexible to account for evolving contexts;
- 8) using a variety of ways to determine the counterfactual;
- 9) estimating the potential sustainability of whatever changes have been observed;
- 10) communicating the findings to different audiences in useful ways;
- 11) etc. ...

The point is that the list of what’s required for ‘rigorous’ impact evaluation goes way beyond initial randomization into treatment and ‘control’ groups.

Here, in a nutshell, is one of the main points of this presentation:

To attempt to conduct an impact evaluation of a program using only one pre-determined tool is to suffer from myopia, which is unfortunate.

On the other hand, to prescribe to donors and senior managers of major agencies that there is a single preferred design and method for conducting all impact evaluations can and has had

unfortunate consequences for all of those who are involved in the design, implementation and evaluation of international development programs.

We must be careful that in using the “Gold Standard”
we do not violate the “Golden Rule”

In other words:

“Evaluate others as you would have them evaluate you.”

Or “Judge not that you not be judged!”



Caution: Too often what is called Impact Evaluation is based on a “we will examine and judge you” paradigm. When we want our own programs evaluated we prefer a more holistic approach.

To use the language of the OECD/DAC, let’s be sure our evaluations are consistent with these criteria:

RELEVANCE: The extent to which the aid activity is suited to the priorities and policies of the target group, recipient and donor.

EFFECTIVENESS: The extent to which an aid activity attains its objectives.

EFFICIENCY: Efficiency measures the outputs – qualitative and quantitative – in relation to the inputs.

IMPACT: The positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended.

SUSTAINABILITY is concerned with measuring whether the benefits of an activity are likely to continue after donor funding has been withdrawn. Projects need to be environmentally as well as financially sustainable.

The bottom line is defined by this question:

Are our programs making *plausible contributions* towards positive impact on the quality of life of our intended beneficiaries?

Let’s not forget them!

