International Initiative for Impact Evaluation

WORKING PAPER 4

Designing impact evaluations: different perspectives

Robert Chambers, Dean Karlan, Martin Ravallion, and Patricia Rogers

July 2009

About 3ie

The International Initiative for Impact Evaluation (3ie) works to improve the lives of people in the developing world by supporting the production and use of evidence on what works, when, why and for how much. 3ie is a new initiative that responds to demands for better evidence, and will enhance development effectiveness by promoting better informed policies. 3ie finances high-quality impact evaluations and campaign to inform better program and policy design in developing countries.

3ie Working Paper series covers both conceptual issues related to impact evaluation and findings from specific studies or synthetic reviews.

This Working Paper was edited by Dr. Howard White, 3ie Executive Director.

© 3ie, 2009

Contacts

International Initiative for Impact Evaluation
c/o Global Development Network
Post Box No. 7510
Vasant Kunj P.O.
New Delhi – 110070, India
Tel: +91-11-2613-9494/6885
www.3ieimpact.org

Table of Contents

Preface

Debates on approaches to impact evaluation design appear to have reached an impasse in recent years. An objective of the international conference, Perspectives on Impact Evaluation, March 29th to April 2nd, Cairo, organized by 3ie, NONIE, AfrEA and UNICEF, was to bring together different voices and so work toward a consensus. A key session in this approach was a plenary in which experts from different perspectives were asked how they would approach the evaluation of three interventions: a conditional cash transfer, an infrastructure project and an anti-corruption program. The motivation for the session was that debates get stuck when they remain at the conceptual level, but that a greater degree of consensus can be achieved once we move to the specifics of the design of a particular evaluation. I am very pleased that the four presenters agreed to write up their views so they can be more widely disseminated.

Thanks are due to Hugh Waddington and Rizwana Siddiqui for assistance in the preparation of this collection.

Howard White
Executive Director, 3ie

So that the Poor Count More: Using Participatory Methods for Impact Evaluation

Robert Chambers, Institute of Development Studies, University of Sussex[*]

Abstract

The starting point for an evaluation is to ask why it is being conducted, who will benefit, and what impact the evaluation will itself have, and how. Participatory approaches and methods fit in a paradigm that is pluralist, evolutionary and iterative. They include stakeholder analysis, individual story-telling, participatory social mapping, causal-linkage and trend and change diagramming, scoring, and brainstorming on program strengths and weaknesses. Well designed and facilitated, participatory methods are rigorous, and besides offering qualitative insights can count the uncountable, and generate statistics for relevant dimensions that would otherwise be overlooked or regarded as purely qualitative. They open studies to the voices of those most affected by a project in a ways not possible using more conventional methods and can make the realities and experiences of poor people count more.

Introductory remarks

Though flattered to be invited, comparative incompetence made me reluctant to be on this panel. A majority of you in this conference have designed evaluations [we did a show of hands and it was more than half]. I have only designed one: that was in 1970 and was a disaster. I was reluctant too because we are living in a time of explosive innovation with participatory methodologies, including for monitoring and evaluation, and the three programmes chosen lent themselves less to these than would have others concerned with, say, community development, agriculture or natural resource management. I also feel embarrassed, as other speakers have been, to be yet another person from the North. An African in one of the smaller sessions said: "anything that comes from the North is truth'

Let us recognise that much of the 'truth' - the creativity and innovation - with participatory methodologies has come and is coming from the South, from Asia, from Latin America, and, notably, from Africa.

I was encouraged, though, by Sulley Gariba's stress, in his opening remarks, on 'empowering communities', and Erma Manoncourt's appeal to 'open the door for people's participation and empowerment'.

In all three cases considered here – a CCT, infrastructure development, and an anti-corruption commission - I would argue for an approach that was pluralist, evolutionary and iterative. Mixed methods would be used. The starting point would be to ask about the political economy of the evaluation: who would gain? Who might lose? And how? And, especially, how was it intended and anticipated that the findings would make a difference. This might well require a brainstorming workshop with staff from the funding agency. If they were unwilling or unable to find the time for this, or to shed light on these questions, I hope I would have the guts and resources to turn down the assignment. I would negotiate the MOU to include other steps. One would be a stakeholder analysis and negotiations to involve relevant stakeholders in the process.

I will limit my suggestions largely to participatory methods that could be used, either instead of or complementing others. For these, a priority would be a search for good facilitator innovators. I would negotiate for time (probably longer than the funders had

[*] E-mail: robertc@ids.ac.uk

anticipated) for workshops and fieldwork to evolve and pilot test a participatory approach, avoiding premature closure, and to borrow the title of Irene Guijt's book, 'seeking surprise'. Above all, and throughout the process, there would be the 'so what?' question about pathways to impact for the impact assessment itself and its cost-effectiveness and value added.

Case 1: Conditional Cash Transfer (CCT)

Application of the following participatory methods and approaches would be explored:

- Participatory census mapping in representative communities. Careful selection of communities would be important. The mapping would cover all the people in each community and so avoid sampling issues. It would be facilitated to identify females of school age, in school, not in school before the programme, and in school and not in school now. This would generate statistics to be checked against official records which would include changes in girls' performance.

- Follow up with individual cases, including outliers and angry people, if any

- Focus groups and semi-structured interviews as appropriate with girls, teachers, parents, administrators…

- Inviting girls and others to devise and perform drama of their lives and experiences before and after

- Facilitating causal-linkage diagramming

- Collection of stories

- Brainstorming on strengths and weaknesses of the programme and how to improve it.

Facilitator/researchers might stay a few days and nights in communities and would meet to compare notes. New issues and questions would be expected, adding to the agenda.

Case 2: Ex-Post Evaluation of a Transport Sector Program in a South Asian Country

Application of the following participatory methods and approaches would be explored:

- Focus groups and semi-structured interviews with key stakeholders and affected people, for example: Small business people from the informal and formal sectors, transport contractors, drivers, trade unionists in the port, and large employers

- For the impact of rural feeder roads, selection of a range of conditions and communities, informally identifying gainers and losers, and then semi-structured interviews and/or focus groups of affected people, facilitating: further identification of gainers and losers, time lines, trend and change diagramming, before and after matrix scoring, causal-linkage diagramming of changes, and scoring of linkages.

- Numerical estimates of gains (and losses) and, where appropriate, 'interviewing the diagram' for deeper insight.

Producing numerical estimates goes against the view that participatory approaches can only produce qualitative data. However, the last decade has seen a growth in 'parti-numbers', that is participatory approaches that generate quantitative data (see Chambers, 2007).

Case 3: Evaluating donor support to an anti-corruption commission in an African country

In a sensitive area such as corruption, the evaluation would more than usually itself be an intervention and treated as an opportunity. This would be reflected on throughout to enhance benefits and minimise damage. An early step would be to identify forms and levels (high-level, low-level) of corruption and hypothetical causal linkages between the donor support and changes that might have occurred. Advocacy and human rights NGOs, and journalists, would be key informants.

With those who went on the study tours, and members of the commission:

- Reflective discussion in focus groups (but see individual meetings below)

With low-level corruption, if there was any plausible connection, search for varied sources of evidence and insight:

- Citizen scorecards or the equivalent if available
- Focus groups of angry people
- Casual discussions in tea shops etc.

With high-level corruption a major part would be

- Informal one-to-one private discussions without taking notes. To support this, in the best tradition of innocent but Macchiavellian facipulation (facilitation but with the intention to manipulate), evaluators would be provided with a bar allowance

Final remarks

A question has been asked about counting the uncountable. Participatory methods have a largely unrecognised ability to generate numbers which can also be commensurable and treated like any other statistics. Through judgement, estimation, and expressing values, people quantify the qualitative. The potential of these methods is overdue for recognition. As always that there are ethical issues. Well facilitated, participatory methods can be win-win – empowering people as well as providing credible and reliable insights for policy-makers.

Power and the political economy of methodologies have been raised as issues in this conference. I have heard a concern expressed that the choices of intervention might be influenced by their amenability to impact evaluation by methods to which certain forms of rigour are attributed. This is not something I can judge. But if, for example, this approach led to a bias favouring programmes with simple standard quick acting inputs over others that were more complex, long-term, pluralist, participatory, and empowering, the opportunity costs to poor people could be high.

Finally, it is striking how rarely in this conference the primacy and capabilities of poor people have been mentioned. The purpose of impact assessment is learning and change that makes life better for them. To achieve this, we need mixed methods and pluralism. Many approaches and tools can be, and should be, used for impact assessment.

Whatever they are, they must always recognise that it is those who live in poverty, those who are vulnerable, those who are marginalised, who are the best judges and the prime authorities on their lives and livelihoods and how they have been affected. We now know, as we did not two decades ago, that they have far greater analytical capabilities than we supposed. We know that ‘They can do it’. To facilitate their own empowering

analysis we now have a wealth of participatory methodologies. We need to make more and better use of them. Again and again, the injunction bears repeating: Ask them!

Reference

Chambers, Robert (200?) 'Who Counts? The Quiet Revolution of Participation and Numbers' *IDS Working Paper 296*. Falmer: Institute of Development Studies.

Thoughts on Randomized Trials for Evaluation of Development: Presentation to the Cairo Evaluation Clinic

Dean Karlan[*]
Yale University
Innovations for Poverty Action
Jameel Poverty Action Lab

Abstract

We were asked to discuss specific methodological approaches to evaluating three hypothetical interventions. This article uses this forum to discuss three misperceptions about randomized trials. First, nobody argues that randomized trials are appropriate in all settings, and for all questions. Everyone agrees that asking the right question is the highest priority. Second, the decision about what to measure and how to measure it, that is through qualitative or participatory methods versus quantitative survey or administrative data methods, is independent of the decision about whether to conduct a randomized trial. Third, randomized trials can be used to evaluate complex and dynamic processes, not just simple and static interventions. Evaluators should aim to answer the most important questions for future decisions, and to do so as reliably as possible. Reliability is improved with randomized trials, when feasible, and with attention to underlying theory and tests of why interventions work or fail so that lessons can be transferred as best as possible to other settings.

1. Introduction

Why do we evaluate? Three reasons stand out: to know where to spend limited resources, to know how to improve programs, and to motivate those with money to give or invest more.

I would like to begin with a thought experiment from the utilitarian philosopher Peter Singer. Would you save a child drowning in a lake if it would cost you $100 in ruined clothing or a missed appointment? Most people answer yes to this question. But would you also send $100 right now to an NGO in a poor country to save a child? Many people say no, arguing that no one really knows if their $100 can save a child or will just get wasted. This is a common excuse for inaction. Evaluation rebuts this excuse.

There has been much discussion about the use of randomized control trials (RCTs) versus other methods of evaluating programs. But, in many cases, this hard split between experimental and other approaches is manufactured and masks the overlap between experimental and qualitative methods that can characterize good evaluation. In this note, I begin by outlining some common misunderstandings of the measurement method, attribution and feasibility of randomized control trials (RCTs). Then, I will describe three examples of common development programs – conditional cash transfers, infrastructure and anti-corruption measures -- and the circumstances in which RCTs should or should not be employed as part of the evaluation strategy.

2. Misperceptions of RCTs

A common misperception is that one much choose either to do a qualitative evaluation or an RCT. Underlying this is an erroneous spectrum of "attribution" rigor, with RCTs on one end and qualitative methods on the other. In reality, qualitative methodologies are not the opposite of RCTs. For one, a good RCT evaluation often involves a thorough

[*] Email: dean.karlan@yale.edu.

assessment of how the program functions, its initial design, theory of change, beneficiary participation, etc.

To clarify the discussions on evaluation methods, it is imperative to separate our conversations about collecting data and measuring outcomes—what to measure, how to measure, and who to include in the process—from how to establish causality between the outcomes and intervention. RCTs establish causality by providing a measure of the counterfactual: what would have happened had the program or policy not existed. Just as is standard practice in medical trials, they achieve this by randomly assigning people to treatment and control groups, so that, except for the random program or offer, the groups are alike on observable and unobservable characteristics if the sample size is sufficiently large.

2.1. Establishing Causality

The random assignment is helpful because of selection bias, or in other words because program participants are often different from non-participants. If instead we were to compare those who could participate in a program but choose not to, we would end up comparing two potentially very different sets of people. It is easy to see how these groups might differ in important but hard to measure ways. Those who join the program might be more driven to improve their situation, or more empowered, or better educated. They might have more free time. Researchers often try to control for these differences, but inevitably there are omitted variables, or others, like motivation, that can be problematic to measure. These differences mean that estimates of the impact of the intervention are biased, since differences in outcomes in the treatment and control groups may result from these unobserved characteristics, rather than being caused by the intervention.

2.2. Data and Measurement

Quantitative outcome measures are useful for evaluations because they allow researchers to establish statistical significance for program impact. But RCTs do not specify any one method for data collection. Both quantitative and qualitative data can be used within the RCT framework, often in combination within the same evaluation. Methods from economics, sociology and psychology or other disciplines can be used, as well as participatory processes involving local voices (e.g., see Chattopadhyay and Duflo 2004 which found that women in West Bengal were more likely to participate in the policy making process if the leader of their village council was a woman), among others, and even "outliers" as Chambers discusses in this forum (Karlan and Zinman 2009).

A common misperception directed at advocates of RCTs is that we suggest they can and should be conducted on every program. RCTs are an important research tool because the causality they establish provides rigorous measure of program impact, and thus helps to know whether to replicate elsewhere, as well as how to improve. However, RCTs are not always feasible. Where RCTs are appropriate depends partly on the situation, and also on the question being asked. And as Ravallion's (2009) article in this forum discusses, one should never start first with the methodology and then figure out what to ask. The evaluators must first establish the questions that need answering, and then examine the most appropriate tool to answer them. When feasible, RCTs provide the most unbiased estimate of program impact, but merely being feasible by no means suggests they should be done just for the sake of doing one. Where no convenient identification strategies exist RCTs are without doubt the most practical means of creating a credible research setup.

2.3. Creative Approaches in RCTs

While we emphasize that RCTs cannot work everywhere, many settings which seem infeasible are in fact feasible with a little creativity. For example, interventions can often take advantage of implementation limits and randomize at the community or other geographical level rather than randomly selecting individuals into control and treatment groups. There are several evaluations measuring the impact of microfinance that use this approach. In other cases, differences in the intensity of marketing a program to different areas (encouragement design) can be exploited in an RCT. The key criterion for RCTs is sample size, in separable enough units such that spillovers and general equilibrium effects can be measured. If planned properly and if the effects are not overly aggregated, (e.g., at a country level), then careful RCT designs can measure both the direct impacts of the intervention as well as the positive and negative spillovers onto groups outside of the direct beneficiaries. These are in fact some of the most exciting RCTs to read about, because they help us understand not just whether an idea works on a particular individual, but how it will play out on a larger scale with direct and indirect effects.

2.4. Static vs. Dynamic Implementation Approaches

Another common misperception of RCTs is that the intervention must be homogenous and static. Indeed, 'emergent, complex' or 'comp lex' interventions, such as those discussed by Rogers (2009) in this forum, are not more difficult for an RCT to handle than for a non-RCT. Arguments that suggest complexity and a dynamic process wreak havoc with an RCT are failing to recognize what exactly an RCT gets us. An RCT simply helps to generate an objective comparison group against which to compare changes. The intervention itself of course can be static and simple, or complex and changing. If the latter, then the evaluation is of course described as such: one is evaluating a process, an opportunity coupled with some resources, a dynamic and fluid intervention that was led in a certain way, etc. The key here is that it is the process, not the individual activities that make up the program implementation, is thus being evaluated. If the project were to work, then what needs to be replicated is the process of putting resources into place, facilitating the use of them, etc. This is much akin to many community development interventions in which resources such as training and customized technical assistance are provided to communities and facilitation exercises are put in place to help communities grow and prosper.

We are conducting just such an evaluation using an RCT approach, complete with qualitative and quantitative tools, of The Hunger Project in Ghana, and of a community-driven development program in Sierra Leone. It is important of course to understand that what is being evaluated here is a collaborative process rather than a clearly defined intervention. It is not possible to know up front what inputs the particular actors will select, nor to expect that the same process elsewhere would yield the same choices. Thus the lessons from such an evaluation are about the changes one can expect from just such a process—not from the specific choices and investments the actors choose to make, but from the process of facilitating and/or financing the villages as they develop the program themselves. That said, if program officials or managers were interested in measuring the individual impacts of the activities that make up the intervention, an RCT could be designed to deliver discrete results from complex interventions. This would require randomly varying the components of the intervention into multiple treatment groups. The likely comparison would be impact of a base set of services, with or without the interaction of one or more add-on components.

RCTs have an important advantage over other methods here, because they can address selection biases inherent in many social programs, and in addressing the direct impacts of different activities in a multiple treatment design. For instance, if one conducted an evaluation of business training and found an increase in profits, especially among those who were found to engage in better recordkeeping, does this suggest training in recordkeeping should be promoted? Maybe recordkeeping is a key component of the

training, or potentially the better entrepreneurs naturally engaged in recordkeeping. RCTs can disentangle these issues by assigning participants to receive training with or without a special recordkeeping module.

Another common misperception of RCTs is that they measure impacts of an intervention only on the population average, ignoring the differential impacts on different segments of the population. In fact, given sufficient sample size and a sampling plan that includes a variety of people that might be eligible for the broader program, an RCT can help identify groups for which the program has the largest impact and groups for which the impact is insignificant or even negative. For example, one surprising result from an RCT measuring the impact of a micro enterprise business training program in Peru, was that businesses who expressed no interest in additional training actually benefited somewhat more from the program (Karlan and Valdivia 2008).

3. Three examples

Ravallion's article in this forum provides an excellent overview of the types of questions one must ask in the beginning of the evaluation process in order to define the aim and scope of the evaluation, and thus the key research questions. As he discusses, depending on the unit of assignment, randomization will or will not be feasible. These three examples provide an excellent spectrum of just that point. I will discuss here both broad plans for each on how one could evaluate them, and then specific ideas within each on how subsidiary questions about specific implementation questions could be answered through randomized trials, even if the core intervention is employing other methods to assess its overall impact. These ideas are not in lieu of the overall non-experimental evaluation, but can provide useful methods of generating precise and objective data to help with important future implementation questions.

Turning to the first example, conditional cash transfers (CCTs), the recommended method for impact evaluation is the randomized control trial, involving quantitative and qualitative data collection. These have been conducted in several countries. Where governments have had limited resources to scale up CCT programs randomization is an especially fair and transparent way to distribute benefits in a staged manner. Recent research has shown that, designed appropriately, CCTs can be an effective means to achieve important public policy goals. However, the question of how best to implement these programs is far from settled. For example, implementation questions include how frequent to make the payments, whether to consider adding savings services, and whether to coincide payment with education expenditures.[1]

For the second example, infrastructure, there are several options for designs that are technically feasible, but that require a varying degree of commitment on the part of government officials managing the programs. I will discuss port rehabilitation, trunk roads and rural feeder roads. Unfortunately, evaluators are too often asked to evaluate only after it is too late. Regardless of the method employed, it is far preferable to set up the evaluation in advance, have clear objectives and be inclusive about what and how to measure the results.

The evaluation of port rehabilitation and trunk roads could involve a heavy focus on process evaluation methodologies. The first step will be to establish a log frame, with targets for example for the number of days wait time and the number of days to transport; the cost of shipping and transporting over land; the value of goods being shipped; the quantity of goods shipped; and the number of ships, trucks and cars entering or leaving. There is potential for econometric tools to be used, depending on differential effects on industries, and tariffs, for example. This is simply program monitoring, and it is important both for implementation management and accountability for results.

[1] For a good discussion of how to design choice environments that help people choose ethically, see *Nudge* by Richard Thaler and Cass Sunstein.

In these cases, RCTs can be employed to help answer critical aspects of the theory of change for the program, but are not likely to involve the entire intervention. For example, a key question for a port rehabilitation in a developing country might be 'Will lower transport costs lead to more growth of industry in rural areas?' In this case, one could consider an RCT which randomly subsidizes transport costs in some areas, to examine the change in economic activity as a result.

Answering key policy questions on the impact of feeder roads programs via an RCT approach can be technically feasible, but is also likely to require a great deal of commitment on the part of policy makers. This example is one that has huge benefits in terms of policy lessons for other countries, but is also one that we recognize might be difficult to accomplish politically. If there are enough roads, and geography and construction costs permit, there is potential for a randomized phase-in of the road construction. Imagine a ten-year plan to improve or build rural feeder roads.

Randomizing the order is both (a) fair, and (b) easily evaluable. This could be implemented incorporating road prioritization within the ten-year plan, if some roads are more important for economic and geographic (or political) reasons. Enterprising policymakers might recognize that one advantage of an RCT in this context is that it avoids political favouritism to decide ordering. (That is, roads would be selected for wave 1 or wave 2 in a transparent deliberative process, then the order of road construction within each wave would be randomly drawn, hence fair.) In this case, the RCT is one facet of the evaluation design. It could also involve the use of econometric methods, including difference-in-difference approaches, before versus after, or a cross-sectional comparison of built versus not built (for example towns 5 miles from the repaired or built road, versus 5 miles from un repaired or un built road).

The final example, anti-corruption measures, is not usually the place to look for attribution-style evaluation, although there can certainly be some measurable process outcomes such as arrests made, or politicians thrown out of office. But inside the box of how and why public officials resort to unlawful tactics, much can be learned. And furthermore, this is one area where transparency in the evaluation approach really matters! For example, Brazil's municipal audits were televised. Work by Olken (2007) in Indonesia is another good example, enabling us to learn the relative effectiveness of competing anti-corruption methods, through a mixture of qualitative (perception of corruption and participatory methods from village meetings) and quantitative (quality of actual roads) data collection.

4. Conclusion

One final advantage of the RCT approach is the independence that it allows, in that one can establish clear statistical tests ex-ante, and then let the data speak as to whether something worked or not. Ultimately, the goal for evaluation should be to help decide what to do in the future. This is both for donors who need to know where to put their money, for sceptics who want to see that programs can work, and for implementers who need to know how best to design their programs. Some of the most exciting work uses mixed methods by incorporating qualitative methods into randomized trials, and by using randomized methods for evaluating dynamic and complex processes, such as community development programs.

In this paper, I have focused on a couple key issues in the debate surrounding impact evaluation methods: the parsing of what to measure versus what to compare. Looking at these distinct questions we can see RCTs focus on the latter, and are flexible to including many participatory, qualitative, and quantitative methods for the former. I have also tried to dispel some common misperceptions about the extremes of the debate. Even proponents of RCTs do not advocate that they be conducted everywhere and for every program. If I were to hazard a guess, it would be that less than 1 percent of evaluation

budgets are used for RCTs. I think they should be done more, but not 100 percent (or 99 percent) either.

References

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. Women's leadership and policy decisions: evidence from a nationwide randomized experiment in India. Econometrica, 72(5), 1409–1443

Karlan, Dean, and Martin Valdivia. 2008. Teaching Entrepreneurship: Impact of Business Training on Microfinance Institutions and Clients. Yale University Economic Growth Center working paper.

Karlan, Dean, and Jonathan Zinman. 2009. Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. Review of Financial Studies.

Olken, Benjamin. 2007. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. Journal of Political Economy 115: 200-249.

Ravallion, 2009. Evaluating Three Stylized Interventions, Journal of Development Effectiveness, forthcoming.

Rogers, Patricia J., 2009, Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation, Journal of Development Effectiveness, forthcoming.

Thaler, Richard, and Cass Sunstein , 2008, Nudge: Improving Decisions About Health, Wealth, and Happiness, New Haven, CT: Yale University Press.

Evaluating Three Stylized Interventions

Martin Ravallion[*]
Development Research Group
World Bank, Washington DC, 20433, USA

Abstract

Along with the other panellists in a session of this conference, I was asked to discuss evaluation designs for three stylized interventions: conditional cash transfers, a transport sector program and an anti-corruption commission. This paper records my responses, and elaborates a little on some points, including references to the literature. I begin with some general suggestions on the issues to think about at the outset of any evaluation. I then try to illustrate these points with reference to the three stylized interventions.

1. Introduction

The participants in this session were asked to discuss how to evaluate three interventions:

1. '*A conditional cash transfer in a Central American country, in which households receive a monthly payment if females of school age remain in school and meet specified attendance and performance requirements.'*

2. *'A transport sector program in a South Asian country that includes port rehabilitations, trunk road rehabilitation, and new investments in rural feeder roads.*

3. *An anti-corruption commission (ACC) in an African country. The program includes helping develop guidelines, infrastructure upgrading and study tours. Similar programs are being implemented in six countries.*

These are three very different interventions in many respects, but the most important difference in an evaluation context is in the degree to which they can be treated as assigned programs, meaning that some observational units (households, firms, villages, areas) receive the program but some do not. A related difference is the extent of spill over effects, whereby non-assigned units are affected (positively or negatively) by the program. The fact that a program is assigned does of course not mean that the non-assigned units are unaffected. Spill over effects can be a serious source of bias in classic evaluation methods.[2]

At the two extremes, a conditional cash transfer (CCT) program is an assigned program, targeted to specific households with probably modest spill over effects, but an ACC is generally economy-wide, although it may have some assigned aspects, such as when it 'targets' specific parts of government or specific firms. Between these extremes, the transport program is a hybrid; most components are assigned, though spill over effects can be large.
Different evaluation tools are needed for assigned versus non-assigned programs, although the essential principles of evaluation—including the need to assess impact

[*] These are the views of the author and should not be attributed to the World Bank or any affiliated organization. Address: mravallion@worldbank.org. The author is grateful to participants at the conference for their comments and to Phil Keefer, Norbert Schady, Dominique van de Walle and Howard White for helpful discussions and comments.
[2] See, for an example, Chen *et al.* (2009) in the context of local public spending responses to poor-area development programs in China and Oduor *et al*. (2009) on spill over effects from social marketing of malaria treatment in Kenya.

against an explicit counterfactual—are the same. An assigned program facilitates observational comparisons, whereby some selected sample of non-assigned units is used to try to infer the counterfactual under certain identifying assumptions.[3]

In each case I will begin with some key questions about the context of the intervention, before discussing evaluation issues for these interventions. Naturally that makes it hard to be very specific without knowing more about the setting, but I will try.
First some general comments relevant to all three.

2. Generic questions

The key questions I like to ask at the outset of any evaluation are the following:

- *Why this intervention?* Understanding the rationale for the specific intervention is important to designing a useful evaluation, but it also matters to good policy making more broadly (which is after all the ultimate objective of evaluation). Probing into the rationale for the intervention might even lead to a different intervention.

- *What do we currently know about this type of intervention and what are the most important knowledge gaps?* There is almost always some relevant past experience. At the outset, a good review of past evidence can be revealing, and may well influence program design and implementation as well as the issues that the evaluation chooses to focus on.

- *What is the relevant counterfactual?* The classic counterfactual is the absence of the program, but this need not be the counterfactual of greatest interest to policy makers who will often spend the same resources on some other program. A specific program may appear to perform well against the option of doing nothing, but poorly against some feasible alternative. Formally, the evaluation problem is essentially no different if some alternative program is the counterfactual; in principle we can repeat the analysis relative to the 'do nothing counterfactual' for each possible alternative and compare them. This is not often done in evaluating development projects but is more common in health care and medical trials (where the control group gets the existing intervention and the treatment group gets the new one).

- *What are the desired outcomes? Over what period? And what are the potential undesired outcomes?* It is clearly very important to know the objectives and how they can be translated into specific measurable (quantitative or qualitative) outcomes. It is no less important to know over what time period the (positive and negative) outcomes are expected.[4]

- *What are the relevant parameters to estimate?* Classic evaluations focus on just two parameters, namely the average impact of an intervention on the units that are given the opportunity to take it up (the 'intent-to-treat' parameter) and the average impact on those who receive it. However, policy makers typically do not just care about these two parameters. Other questions of interest include: Does the intervention work the way it was intended? What types of people gain, and what types lose? What proportion of the participants benefit? What happens when the program is scaled up? How might it be designed differently to enhance impact?

- *What are the expected transmission mechanisms?* The classic impact evaluation can be a 'black box' that tells us very little about how a program does or does not have impact. To design an evaluation that can throw light into this black box one needs to understand the theoretical rationale for a program—the precise ways in which the intervention is expected to improve peoples' lives. (I discuss such 'theory-based'

[3] For further discussion of the assumptions and methods used for assigned programs and references to the (extensive) literature see Ravallion (2008).
[4] See King and Behrman (2009) for a useful discussion of this point.

evaluation later.) Policy makers can often be rather vague about those mechanisms, and probing by evaluators may be helpful in exposing the rationale for the intervention, and sometimes even revealing that the rationale is dubious.

- *What are the feasible methodological options in this setting?* There will be relevant constraints (technical, economic, political and ethical) on the evaluation and these should ideally be identified at the outset. Some otherwise desirable evaluation methods might be infeasible in the specific setting. Amongst the feasible options, the choice should depend on the answers to all the preceding questions, rather than the methodological preferences of the evaluator. This may seem obvious, but it is not in fact common practice. Too often the evaluator brings his or her own favourite set of tools to the job, and chooses questions that can be addressed with those specific tools (rather than the other way round). Sometimes the admissible toolkit is remarkably sparse, and the evaluator starts with a single preferred method (such as randomization, some non-experimental econometric method or a favourite qualitative tool) and looks for questions that can be addressed with that method. This does not, as a rule, make for the most useful evaluations.

The answers to these questions are often key to the design of an evaluation but will be specific to each program and its setting. There is no single right way appropriate to any given intervention independently of the setting. In this light, let me offer some thoughts on the evaluation issues raised by these three, rather different, interventions.

3. A conditional cash transfer program

This is the easiest of the three for two reasons. First, a CCT is an assigned program, and it is probably reasonable to assume that spill over effects are minimal. Second (and partly because of the first reason), there has been a lot of evaluative research on CCTs, so we know quite a bit about the issues related to these programs and how best to evaluate them. However, knowledge gaps remain.

The essential idea of a CCT is that the recipient family must demonstrate adequate school attendance and (in some examples) health care; the transfer payment is only made if these conditions (sometimes called 'co-responsibilities' are verified).[5] Early influential examples were the *Food-for-Education Program* in Bangladesh, Mexico's *PROGRESA* (Programa de Educacion , Salud y Alimentacion) program (now called *Oportunidades*) and *Bolsa Escola* in Brazil. There is evidence from impact evaluations that such CCT programs bring non-negligible benefits to poor households, in terms of both current incomes and future incomes, through higher investments in child schooling and health care; for a recent review of the evidence from past evaluations see Fiszbein and Schady (2009).

The rationale for a CCT is not as obvious as one might think. A longstanding question is why the conditions are imposed, rather than making unconditional transfers. Credit market failures, whereby poor households cannot borrow to finance their kids' schooling, are often cited as the reason for a CCT, but there may be better policies to address that problem, including unconditional transfers (Das *et al.* 2005). Issues about intra-household inequality and political economy often dominate other rationales including credit market failures.

One of the key knowledge gaps about these programs is how well they work in low-income countries (most of the programs and evaluations come from middle-income countries, notably Latin America.) This knowledge gap is in the process of being filled,

[5] The term "conditional cash transfers" is a misnomer since most almost all transfer and social protection programs impose conditions on recipients, but the term has been applied solely to this specific type of program.

based on CCTs currently being evaluated (notably in Africa), with many new results likely in the next few years.

We need to know more about the composition of the CCT package of instruments, to try to figure out how to set the precise composition of transfers (how much. and whether it is in cash or kind) and incentives for behavioural change (which behaviours are to be encouraged). There is some evidence that a budget-neutral switch of the enrolment subsidy in *PROGRESA* from primary to secondary school would have delivered higher school attainments, by increasing the proportion of children who continue onto secondary school (de Janvry and Sadoulet, 2006). This is an example of my earlier point that evaluators need to better understand precisely how programs have their impacts, so we can better advise how to improve their design.

While there has been some good evaluative research on behavioural responses it has focused more on the behaviours intended by the program's designers (namely compliance with the program's co-responsibilities). We need to know more about other responses, including labor supply and savings decisions of parents.

Past research has also left holes in knowledge about the welfare impacts of the induced behavioural changes. The outcomes should include both current and future poverty. Current poverty is easier to measure (though still not easy to measure). It can be useful to focus on 'intermediate outcomes' such as child labour and school attendance, but we know less about final outcomes. Sure, more kids from poor families go to elementary school (say) when the price of schooling is lowered through the conditions. But do they learn anything useful? Or is it too late to affect their learning abilities? What about impacts on post-school labour earnings?

The importance of contextual factors related to the supply side looms large in CCTs. Disappointing outcomes in terms of child learning and nutritional status revealed by some past evaluations can be linked to supply side factors—poor quality of schools and health clinics—that vary from one place to another. The behavioural changes induced by a CCT may be insufficient to attain desired welfare outcomes without supply side improvements (Fiszbein and Schady, 2009).

Important knowledge gaps relate to the longer-term impacts of CCTs. This will naturally take time. But the investments in data need to be made now. Alas, the externalities involved and the consequent difficulties in financing and sustaining studies of long-term impacts are severe.

A further gap in our knowledge about CCTs concerns their flexibility in adjusting eligibility to changes in need. This is important if these programs are to serve a safety net function, such as during the present global financial crisis. One expects that it is easier to implement temporary top-up payments to current beneficiaries than to temporarily expand the number of beneficiaries in a crisis. But more research is needed on this issue, and how CCT programs might be made more flexible in practice.

In thinking about the options for evaluation, it is important to look into the phasing-in plans for the CCT program. Even if the program is not being implemented on a trial basis, it will not always be feasible (on budgetary and technical grounds) to introduce the program in one go nationally. Some geographic areas may get the program before others. Understanding the selection of targeted areas is key. Look for options for creating a control group out of the observationally similar non-participants areas. Or if there is a formula for selecting which areas go first, one can also consider a discontinuity design, which identifies the impact in a neighbourhood of the eligibility cut off point. As the program expands, the comparison areas will enter, but before then there may be options for identifying impacts. Here it can be important to know whether the comparison areas know that they will be joining the program and when this is expected, since this may generate contamination effects, whereby the comparison units are affected by the anticipation of joining the program.

In many respects the original *PROGRESA* evaluation is a model for other CCTs. The key design feature is that a 'randomized out' group acted as the controls in the program's phasing-in period. The longevity of this program (surviving changes of government) and its influence in the development community clearly stem in part from the substantial, and public, effort that went into its evaluation. One third of the sampled communities deemed eligible for the program were chosen randomly to form a control group that did not get the program for an initial period during which the other two-thirds received the program. Public access to the evaluation data has facilitated a number of valuable studies. A comprehensive overview of the design, implementation and results of the *PROGRESA* evaluation can be found in Skoufias (2005).

There is scope for improving on *PROGRESA*'s design. A more complete accounting of outcome variables would be desirable, as discussed above. There are concerns about possible contamination of the control group, notably through anticipation effects (especially when, as in *PROGRESA*, the program goes national over the course of the evaluation period, and naturally has a high public profile). The evaluation design could also have done more to inform program design issues, such as whether the subsidy should be at the primary or secondary-school level. Supply-side factors have also emerged as important in subsequent research, and better use of linked facility surveys of schools and clinics might have helped.

4. Transport-sector program

There are a number of important contextual issues to think about when evaluating transport-sector programs. Logically the first step is to understand the need for the intervention, which will relate to (inter alia) economic history, geography and political economy. Were the specific ports and transport links selectively neglected, leading to the need for their rehabilitation, and if so why? Questions related to public finance and spill over effects are also likely to loom large. How is the program to be funded, and are there likely to be fungibility or flypaper effects (where, despite partial fungibility, aid remains 'stuck' to the sector in which it is intended)? Are we evaluating the right intervention?

Impacts can be expected in both the short term and longer term. There will be temporary effects on labour earnings (for both the workers on the projects and other workers who may gain from the tightening of labour markets), which may be very important in some contexts, such as when the transport project is part of a crisis response. There will also be various costs in the construction phase, beyond the costs of construction, such as social and environmental impacts (including displaced families).

But the bulk of the impacts will be expected after construction is finished and the evaluation efforts must plan accordingly in its time horizon. And those impacts can be wide-ranging and hard to quantify. Transport cost savings (both out-of-pocket costs and the value of travel time savings for people as well as freight) will no doubt be of first-order importance. For large projects there will probably also be broader (general equilibrium) effects on economic activity, including its geography (which I return to).

The evaluation will need to consider the functioning of, and implications for, markets. We will want to know how well the existing markets for geographically traded goods work, which raises the possibility for identifying impacts through prices. (One of the main ways that a transport project affects welfare is via changes in prices.) We should also explore how well existing land markets work in the specific setting, which may provide scope for identifying impacts through capitalization in land prices. (The classic Von Thunen model predicts that land rents will decline with distance to the city centre, reflecting transport costs; for an example of how this type of model can be exploited in evaluating transport projects see Jacoby, 2000.)

We will probably also like to know how the intervention affects the geography of economic activity, including local market and institutional development. Does the improvement in rural roads, for example, attract markets, institutions and new economic

activity to lagging poor areas, or encourage further geographic concentration of these activities? How do large transport projects change the geography of economic activity?

Given the obvious importance of transport costs to location decisions, a large transport project can be expected to change the landscape of economic activity, benefiting some areas and activities but possibly leading to the decline of others. The new economic geography predicts that lowering transport costs for (farm and non-farm) products through transport infrastructure investments will increase the geographic concentration of other (non-farm) activities in urban areas to exploit agglomeration economies (Fujita *et al*. 2001, Chapter 7). One evaluation of a World Bank supported rural roads project in Vietnam found evidence that the project stimulated local market development in poor areas—partially de-concentrating economic activity; see Mu and van de Walle (2008). Local rural impacts on the geographic concentration of activity may well differ from regional impacts, including urban areas.

One question that is bound to arise is whether the evaluation should focus solely on the outcomes for aggregate economic efficiency. It is sometimes argued that development policy making should not be concerned about equity aspects of the spatial allocation of activity, and not interfere with the market-determined allocation, while using other ('spatially-blind') policy instruments to address concerns about equity, including concerns that have a geographic dimension, such as lagging poor areas; see World Bank (2009) for an argument along these lines. Applying this view to transport programs we would be solely concerned with their efficiency, as measured by their impact on average income.

However, I doubt that many policy makers would accept this view, and for good reasons. They know that they do not have the full set of policy instruments that are needed to deal with equity considerations in a spatially-blind way.[6] Indeed, in poor countries, location is one of the most widely used dimensions for identifying poor people for redistributive purposes, in lieu of better information on individual welfare levels. For some time in developing countries, one expects that policy-relevant evaluations of transport projects will need to consider equity aspects, and in 'non-income' dimensions as well (including, for example, impacts on child health through better maternal access to health care due to transport improvements).

The evaluation design for a transport program will need to distinguish the components that can be treated as assigned programs (such a local rural roads) from those that are likely to have much more spatially dispersed impacts (trunk roads for example), and so will call for a sector- or economy-wide approach.

Evaluations of assigned and unassigned components require different designs. For assigned components, such as local rural road improvements, we will want to identify the relevant 'catchment area' that safely encompasses likely impacts (van de Walle, 2009). We also need to understand how geographic placement was determined, and find observationally similar areas that do not receive the new rural roads or rehabilitation spending. Longitudinal observations will almost certainly be required (to allow a 'matched difference-in-difference' estimator; see Ravallion, 2008, for further discussion and examples). Evaluators will also have to assess the likely sources of selection bias. If selection is largely based on observables, it should be possible to collect the right data to adequately correct for this bias. If selection is on unobservable, it will be necessary to consider the scope for statistical methods such as instrumental variables estimation. (Randomization will rarely be feasible for obvious reasons.)

[6] It has long been recognized in economics that quite strong conditions need to hold for a strict separation of equity and efficiency instruments in attaining overall social welfare objectives. It is widely acknowledged that those conditions (notably the feasibility of non-distortionary lump-sum transfers) do not hold in practice.

If the transport project is large then so too will be its catchment area. Then it may be hard to find good comparison areas, unaffected by the program. We will need to turn to rather different tools.[7] At one extreme (in terms of aggregation), cross-country growth regressions have been used to try to identify the impacts of infrastructure, including transport; see, for example, Calderon and Servén (2008). Sub-national geographic data linked to household and firm data can provide a finer lens; see, for example, Jalan and Ravallion (2002) who find that rural roads in China impact positively on the micro growth process. Spatial computable general equilibrium models, grounded in the New Economic Geography (NEG), may well play an important role in the future (Fujita *et al*. 2001).

These models are conceptually well-suited to the problem of evaluating the impacts of large transport projects, though they are also complex models that are demanding in terms of their data and calibration requirements. The European Union has developed a spatial CGE model for evaluating transport improvement; for an application in the context of evaluating a large transport project (a railways project in Holland) see Knaap and Oosterhaven (2000).

It must be acknowledged that past NEG models have made some rather implausible assumptions in how they model transport costs, and these assumptions may well play a role in the lessons drawn for the impacts of transport improvements.[8] Further advances in building operational models on more realistic assumptions can be expected in the future.

5. An anti-corruption commission

This is an independent body with unusual authority to investigate and prosecute corruption, typically reporting to the head of state or parliament.[9] Two main contextual questions will need to be considered in the evaluation of an ACC. First, an assessment is required of whether the intervention addresses the known causes of corruption. Do the root causes of corruption lie in enforcement mechanisms or in the incentives facing political actors to oversee public officials? Should we be talking instead about reforms of fiduciary institutions, especially public sector management? Secondly, the evaluators will need to be clear about what is motivating this intervention. Is it merely donors' desires for action to punish corrupt officials, or a political leader keen to repress the opposition, or has the intervention come about from a deeper analysis of the problem?

Evaluations can be designed for assessing the impact on corruption of the sorts of tools used by an ACC. An innovative example is found in Olken (2007) study of corruption on roads projects in Indonesia using a randomized design. Olken found that increasing governmental audits reduced the extent of 'missing expenditures' as measured by the difference between officially recorded project costs and independent estimates by engineers. (By contrast, Olken (2007) found that participatory approaches to monitoring had little impact.)

While such evaluative studies could provide useful inputs, they are not identifying the impact of the ACC as such. Olken's study tells us that monitoring by government auditors can help fight corruption in Indonesia; however, by at least one (possibly dated) assessment, Indonesia's own ACC-type efforts do not appear to have been effective in facilitating such monitoring and punishment beyond some well-publicized arrests of notoriously corrupt individuals associated with the prior political regime (Sherlock, 2002).

---

[7] A useful overview of the methodological options can be found in Oosterhaven and Knaap (2003).

[8] I refer here to the "iceberg transport-cost function" which implies that the delivered prices of goods increase exponentially with distance shipped, which is inconsistent with the evidence; for further discussion see McCann (2005).

[9] A useful overview of the history and record on ACCs can be found in Rose-Ackerman (1999, Chapter 9).

The typical ACC is not an assigned program like the CCT or some components of the transport program discussed above. There may be some scope for phasing in geographic and/or ministerial coverage, by assigning the ACC powers to some local governments or ministries and not others. If phasing-in is feasible, and anticipation effects are not likely to be too severe, comparing outcomes of objective corruption tests (such as used by Olken, 2007) between observationally similar government agencies (in different local government areas), with the difference being that some are under ACC scrutiny but some not, might help in assessing the impact of the ACC. However, even aside from feasibility (the government may well prefer to give the ACC a free rein from the start), there are serious concerns about anticipation effects and selection biases. (The sign of the selection bias is ambiguous without knowing more about the setting; a President committed to fighting corruption would presumably pick the places or sectors where the problem is considered greatest; a President using the ACC as a political tool would pick differently, and may even avoid the centres of corruption.)

A more promising (possibly complementary) approach is to consider the likely channels of impacts of the ACC, built on a more-or-less explicit theoretical model of why corruption exists. The evaluation effort could then focus on what seem to be the key links in the expected causal chain, similarly to the type of 'theory-based evaluation' discussed in Weiss (2001), Rogers (2009) and White (2009). For example, Klitgaard (1988) has argued that corruption is the outcome of three factors: a monopoly over some resource, discretionary power of officials in allowing access to that resource, and the absence or failure of mechanisms for making those officials accountable. One might then start by assessing how the ACC addresses each of these elements, by changing the incentives facing officials in the specific institutional environment. This will require assessments of what incentives face individual officials and how that has changed with the ACC. That is never going to be easy, since there may also be few incentives to reveal the truth to an evaluator. But at least by building the evaluation effort on an understanding of why we see corruption in the first place (going far deeper than appeals to the 'immorality' of officials) one can have some chance of determining whether the ACC is having any real impact on the problem.

ACCs are fond of citing the number of prosecuted officials as a measure of success. This could be deceptive, given the possibility that the ACC involves selective targeting of political opponents. Recognizing this, the evaluation might usefully focus directly on political support or opposition for the ACC. Data might be collected on the political affiliation of ACC targets, together with staff surveys on contacts with politicians, and reasons for departure of departed staff. Follow-up surveys (after the baseline survey) will then be needed, asking households, firms and government officials about the level of politicization of the ACC compared to other institutions such as the Central Bank, Finance Ministry and Prosecutor's Office.

6. Conclusion

The art of good evaluation is to ask the right questions at the outset, motivated by existing knowledge gaps and to tailor the data and analysis to answering those questions in the specific context. One cannot anticipate all the most important questions for the specific evaluation in a paper such as this, or anticipate all the options for evaluation methods, both of which will naturally depend on the specific context. However, this paper has hopefully provided some useful starting points for thinking.

References

Calderon, Cesar and Luis Servén, 2008, Infrastructure and Economic Development in Sub- Saharan Africa, Policy Research Paper 4712, World Bank, Washington DC.

Chen, Shaohua, Ren Mu and Martin Ravallion, 2009, Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China, *Journal of Public Economics* 93: 512-528.

Das, Jishnu, Quy-Toan Do and Berk Ozler, 2005, A Welfare Analysis of Conditional Cash Transfer Schemes, *World Bank Research Observer*, 20(1): 57-80.

De Janvry, Alain and Elisabeth Sadoulet, 2006, Making Conditional Cash Transfer Programs More Efficient: Designing for Maximum Effect of the Conditionality, *World Bank Economic Review* 20(1): 1-29.

Fiszbein, Ariel and Norbert Schady, 2009, *Conditional Cash Transfers for Attacking Present and Future Poverty,* World Bank Policy Research Report, World Bank, 2009.

Fujita, Masahisa, Paul Krugman and Anthony Venables, 2001, *The Spatial Economy*, Cambridge, Mass.: MIT Press.

Jacoby, Hanan, 2000, Access to Markets and the Benefits of Rural Roads, *Economic Journal* 110: 713-737

Jalan, Jyotsna and Martin Ravallion, 2002, Geographic Poverty Traps? A Micro Model of Consumption Growth in Rural China, *Journal of Applied Econometrics* 17(4): 329-346.

King, Elizabeth M. and Jere R. Behrman, 2009, Timing and Duration of Exposure in Evaluation of Social Programs, *World Bank Research Observer* 24(1): 55-82.

Klitgaard, Robert, 1988, *Controlling Corruption*, University of California Press, Berkeley CA.

Knaap, Thijs and Jan Oosterhaven, 2000, The Welfare Effects of New Infrastructure: An Economic Geography Approach to Evaluating New Dutch Railway Links, mimeo Erasmus University Rotterdam.

McCann, Philip, 2005, Transport Costs and the New Economic Geography, *Journal of Economic Geography* 5: 305-318.

Mu, Ren and Dominique van de Walle, 2008, Rural Roads and Local Market Development in Vietnam Policy Research Working Paper, World Bank.

Oduor, Jacob, Anne Kamau, and Evan Mathenge, 2009, Evaluating the impact of micro-franchising the distribution of anti-malarial drugs in Kenya on malaria mortality and morbidity , *Journal of Development Effectiveness*, this volume.

Olken, Benjamin A., 2007, Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy* 115(2): 200-249.

Oosterhaven Jan and Thijs Knaap, 2003, Spatial Economic Impacts of Transport Infrastructure Investments, in *Transport projects, programmes, and policies (edited by* A. D. Pearman, Peter J. Mackie, John Nellthorp) Interdisciplinary Centre for Comparative Research in the Social Sciences, Ashgate Publishers.

Ravallion, Martin, 2008, Evaluating Anti-Poverty Programs. In Paul Schultz and John Strauss. eds., *Handbook of Development Economics Volume 4*, Amsterdam: North-Holland.

_____, 2009, Evaluation in the Practice of Development, *World Bank Research Observer* 24(1): 29-54.

Rogers, Patricia J., 2009, Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation, *Journal of Development Effectiveness*, forthcoming.

Rose-Ackerman, Susan, 1999, *Corruption and Government: Causes, Consequences and Reform*, Cambridge University Press, Cambridge.

Sherlock, Stephen, 2002, Combating Corruption In Indonesia? The Ombudsman And The Assets Auditing Commission**,** *Bulletin of Indonesian Economic Studies,* 38(3): 367–83.

Skoufias, Emmanuel, 2005, *PROGRESA and Its Impact on the Welfare of Rural Households in Mexico*, Research Report 139, International Food Research Institute, Washington DC.

van de Walle, Dominique, 2009, Impact Evaluation of Rural Road Projects, *Journal of Development Effectiveness*, 1(1): 15-36.

Weiss, Carol, 2001, Theory-Based Evaluation: Theories of Change for Poverty Reduction Programs. in Osvaldo Feinstein and Robert Piccioto, eds, *Evaluation and Poverty Reduction*, New Brunswick, NJ: Transaction Publications.

White, Howard, 2009, Theory-based Impact Evaluation: Principles and Practice, *Journal of Development Effectiveness*, this volume.

World Bank, 2009, *Reshaping Economic Geography*, World Bank, Washington DC.

Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation

Patricia Rogers, Professor of Public Sector Evaluation at CIRCLE (Collaboration for Interdisciplinary Research, Consulting and Learning in Evaluation) at Royal Melbourne Institute of Technology, Melbourne, Australia. [**]

Abstract

Appropriate impact evaluation design requires situational responsiveness - matching the design to the needs, constraints and opportunities of the particular case. The design needs to reflect the nature of the intervention and the purposes of the impact evaluation. In particular, impact evaluation needs to address simple, complicated and complex aspects of the intervention. Simple aspects can be tightly specified and standardized; complicated aspects work as part of a causal package; complex aspects are appropriately dynamic and adaptive. Different designs are recommended for each case, including RCT, regression discontinuity, unstructured community interviews, Participatory Performance Story Reporting, and developmental evaluation.

The situational responsiveness approach to impact evaluation

This conference session was intended to demonstrate the application of different approaches to impact evaluation design. While the presentations focused particularly on measurement and causal analysis, it is worth remembering that there are other tasks that an impact evaluation must address (Rogers, 2008a). A comprehensive design for impact evaluation sets out how the evaluation will perform the full range of tasks involved in impact evaluation, which are "comprehensive identification of important impacts; systematic and defensible data collection and analysis of evidence of these impacts; sound inferences about the contribution of the intervention to achieving these impacts; and effective management of the evaluation, including transparent reporting of methodology and, where appropriate, formal meta-evaluation" (NONIE Subgroup 2, 2008). For each of these tasks, there are a range of options, and an impact evaluation design needs to choose the most appropriate method or combination of methods for each task.

Some approaches to impact evaluation focus on the use of a particular design or method for data collection and analysis or form of governance. My approach to impact evaluation design can best be described as 'situational responsiveness' (Patton, 2008a).

There is increasing recognition in development and in other areas of evaluation, that different evaluation situations will be best addressed by drawing appropriately from a range of methods and techniques. For example, NONIE[1] in its statement on impact evaluation in 2008 stated that

> "NONIE advocates an eclectic and open approach to finding the best methods for the task of impact evaluation - drawing on the wide range of techniques available from different disciplines." (NONIE, 2008)

In an invited address on 'The State of the Art in Measuring Development Effectiveness' to the World Bank Independent Evaluation Group conference 'Measuring Development Effectiveness: Progress and Constraints', Michael Quinn Patton summed this up as a

[*] Contact details: Patricia.Rogers@rmit.edu.au 124 Latrobe Street Melbourne VIC 3000 Australia.

reframing of what we understand to be the 'gold standard' for impact evaluation "The methodological gold standard here is *appropriateness*, not any one particular method" (Patton, 2008b).

Situational responsiveness involves matching the design to the needs, constraints and opportunities of the particular situation. The two key questions that need to be answered before developing an impact evaluation design are therefore 'What is the nature of the intervention?' and 'Why is an impact evaluation being done?'

WHAT is the nature of the intervention?

The type of intervention and its scale are important determinants of appropriate impact evaluation design. Is it a small project being piloted for possible replication and scale-up? Is it an ongoing program which is likely to continue in some form? Is it a uniform intervention or a collection of disparate initiatives? Is the intervention tightly specified and standardised or does it vary in different locations in response to local conditions, needs and opportunities? These questions have implications for the type of impact evaluation that will be needed, the likely availability of resources for the impact evaluation, and the options in terms of research design.

What is the nature of the impacts that are sought? Are they produced directly by the intervention (like a splash) or indirectly (like a ripple)? Are they short-term impacts that will be evident during the life of a project and an evaluation (such as children's school performance) or long-term impacts that will be evident only many years later (such as post-school employment, or soil recovery from salinity)? Are they transformational impacts, which once achieved are unlikely to be reversed (such as learning how to read or ride a bicycle), or fragile impacts that can be easily undone (such as adequate nutrition or female attendance at school)? Are the impacts likely to be the result of a 'silver bullet" intervention, that achieves results irrespective of context, or a 'ducks-lined-up' intervention', that achieves results only in conjunction with favourable circumstances, including perhaps other interventions.

These different characteristics can be summarised in terms of a three-part typology - simple, complicated or complex (Stacey 1992; Glouberman, 2001; Glouberman and Zimmerman, 2002, Kurtz and Snowden, 2003). This has been shown to be useful for planning and analysing evaluations (Guijt, 2008, Patton, 2008a; Rogers, 2008b). The typology is particularly useful when it is used to classify aspects of interventions rather than the whole intervention.

In this typology, the term 'complex' has a specific and important meaning, which it does not always have in common use. In evaluation the term 'complex' is sometimes used as a synonym for 'complicated', sometimes used to refer to anything which is difficult, and sometimes used as an excuse for inadequate planning. In this typology, 'complex' refers to appropriately dynamic and emergent aspects of interventions, which are adaptive and responsive to emerging needs and opportunities. Simple aspects of interventions can be tightly specified and are standardized - for example, a specific product, technique or process. Complicated aspects of interventions have multiple components, are part of a larger multi-component intervention, or work differently as part of a larger causal package, for example in particular implementation environments, for particular types of participants, or in conjunction with another intervention.

These different aspects of interventions have significant implications for how interventions operate, how we can understand them, and how we can use this understanding, as indicated in Table 1.

Table 1 Implications of simple, complicated and complex aspects of interventions

Aspects	Implications for:		
	Implementation of the intervention	**Causal processes**	**Reporting and use of impact evaluation findings**
Simple	Single organisation	Single causal strand needed to produce the impacts	Single message – what works
Complicated (multiple components)	Multiple organisations in contractual relationship with clearly defined roles	Multiple causal strands needed to produce the impacts:: Multiple sequential interventions or Multiple simultaneous interventions or Multiple levels of intervention or Different causal mechanisms operating in different contexts	Contingent message – what works for whom in what situations
Complex (dynamic and emergent)	Multiple organisations in developing partnership relationship	Causality is recursive, with feedback loops Emergent outcomes – the whole is more than the sum of the parts	Dynamic, emergent message – what is working

WHY is an impact evaluation being done?

The purpose of an impact evaluation also needs to be considered when developing an impact evaluation design. Who are the intended users of the evaluation? What will they consider credible evidence in terms of the impacts to be included, the measures to be used and the approach to causal analysis? Who needs to be involved in deciding the parameters of the evaluation?

Whose values will be used in the evaluation? What will be considered significant impacts, either positive or negative? What will be considered desirable distributions of costs and benefits? Will the focus be on average effect, or the effect on the most disadvantaged? What are the intended uses of the evaluation? Is it being done to retrospectively justify expenditure, in which case credible estimates of net benefit will be sufficient? Or is it being done to inform possible scaling up of a pilot, in which case good information will be needed on how it works? Is it intended to inform incremental change or significant reworking of a program or policy?

Finally, before designing an impact evaluation, logistical issues need to be addressed. By when is a report needed? What evidence is already available about this intervention and about similar interventions? What additional resources are available to do the impact evaluation?

Taking a situational responsiveness approach, it is only when we have addressed all of these questions that we can address the question 'How should it be done?' Clearly, in a real evaluation, all these questions would be answered before, or during the process of, developing a design. In this design clinic, we filled in the gaps in the descriptions of the cases to produce more specific scenarios.

1. Conditional cash transfers

This case was described as follows:

> *A conditional cash transfer in a Central American country, in which households receive a monthly payment if females of school age remain in school and meet specified attendance and performance requirements.*

Conditional cash transfer (CCT) programs have some important aspects that are best characterised as simple – that is, they are discrete, standardised interventions that are intended to be implemented in the same way in different locations. However, evaluations of CCTs have found considerable heterogeneity of outcomes (for example, Soares et al, 2009). CCT may need a package of other interventions to achieve the intended outcomes. For example, CCT may be successful in achieving the initial objective of school attendance but improvements in longer-term impacts such as student learning, graduation and employment outcomes will be dependent on also having effective schools in place. It would therefore be important to consider complicated aspects of the intervention, and include measurement and analysis of other elements (such as particular features of the implementation environment and participant characteristics) needed to achieve the intended impacts.

In this case, assuming that the purpose of this impact evaluation is to decide whether to scale up a pilot program, a Randomised Control Trial (RCT) might be a suitable design for causal analysis, particularly if it was combined with systematic collection of data about other factors, such as the quality of schooling, that might plausibly be needed to achieve the intended impacts, and analysis, such as disaggregation, to identify their contribution. Developing and testing a program theory which included additional factors, differential effects, as well as intermediate outcomes, would improve the quality of the analysis. If CCTs are effective in improving school performance, but only in combination with effective schools, then they will not be a quick-fix by themselves. Alternatively, since eligibility for CCTs is set at a certain level of income, a regression discontinuity approach might provide compelling evidence of causal attribution at less expense than an RCT, although it would also need to be disaggregated by implementation environment.

In addition to measuring intended impacts, it is important that impact evaluation includes other significant impacts, following the DAC (Development Assistance Committee) definition of impact as 'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'.

Some potential unintended impacts can be anticipated and included in targeted data collection, using a 'negative program theory' (Weiss, 1997). Unanticipated impacts, particularly negative impacts, are most likely to be observed through participatory methods (Chambers, 2009). then case studies, based on iterations of interviews, observations and document review, might be needed to uncover unintended impacts (positive or negative), followed by large-scale surveys or review of administrative statistics aimed at producing estimates of the frequency of these impacts.

2. Transport infrastructure program

This case was described as follows:

> *An ex-post evaluation of a transport sector program in a South Asian country that includes port rehabilitations, trunk road rehabilitation, and new investments in rural feeder roads.*

For this case, we assumed that the purpose of the impact evaluation was to understand the overall impact of the investment, primarily for reporting to funders, but also to the public, and that no evaluation had been designed at the beginning of the program. This shortfall, combined with the nature of the intervention (which is multi-faceted, diverse, and affecting an entire region) presents considerable challenges for causal analysis, as classic experimental and quasi-experimental approaches are likely to be impossible to implement effectively (Ravaillon, 2009).

Impact evaluation of this case needs a way to gather together evidence about a large number of diverse components, and to address the issue of attribution without a comparison or control group. For this case, it might therefore be appropriate to use a new approach, Participatory Performance Story Reporting (PPSR), showcased at this conference, (Dart, 2009) which can perform these functions. Since, for some people, the term 'participatory' is understood as implying a less rigorous approach to collecting and analysing data, it is important to understand the very strong empirical base that underlies this approach.

PPSR is a development and systematization of Multiple Lines and Levels of Evidence (MLLE), an approach to causal analysis designed for high-stakes situations where construction of a control group or comparison group is not possible, but where considerable high quality evidence is available, and where systematic causal analysis is required.

> "A line of evidence is a type of evidence, such as an ecosystem attribute (e.g. fish abundance, macroinvertebrate species richness, macrophyte biomass), that is investigated in relation to a stressor or intervention; and
> A level of evidence is a strength-of-evidence value used to determine the case for inferring that a given human activity causes a given ecological or geomorphological change." (Cottingham et al. 2005)

While this approach does not appear in most guides to impact evaluation, its use goes back to the strategies used by in the 1960s analyses of evidence about the link between smoking and lung cancer. Faced with considerable scientific data, but the absence of evidence from RCTs, it was nevertheless important to be able to draw conclusions about the impact of smoking on health. In more recent years MLLE has been used in human and ecological risk assessments and natural resource management (for example, Keough et al, 2002, Boyes 2006).

MLLE involves systematically investigating the strength of the causal argument linking an intervention or a cause and its effects, by analysing an observed association in terms of particular causal criteria and by identifying and ruling out possible alternative explanations. For the association between smoking and lung cancer, Hill (1965) considered its strength, consistency, specificity, temporality, coherence with other accepted evidence, plausibility, and analogy with similar interventions. For investigating ecological responses to management intervention, Norris et al (2005) considered biological plausibility, biological response, dose and consistency of association. Given the specialist and often cross-disciplinary nature of the scientific evidence, this investigation is undertaken by a panel of credible experts spanning a range of relevant disciplines who are asked to judge the credibility of the evidence and the causal analysis. Participatory Performance Story Reporting extends MLLE to also include primary data collection, and a meeting of stakeholders where the existing data, the additional primary

data, and the expert panel analysis are reviewed to produce an agreed impact evaluation report, which includes an auditable trail of the evidence (For example Clear Horizon et al, 2008).

Applying PPSR to the transport infrastructure case would begin by developing a results chain, or program theory, identifying the intended impacts and a chain of intermediate results linking the program with these impacts. Secondly, available evidence about the implementation of the program, the achievement of intermediate results and final impacts, and the contribution of the program to these, would be located and the quality of this evidence assessed. The evidence would be assessed in terms of its quality relative to the type of information not in terms of a hierarchy of evidence based on the type of research design used. For example, interviews with transport users would be assessed in terms of the sampling methods used, the questions asked, the processes used to reduce bias, and the auditability of the data. Thirdly, additional evidence would be gathered to address specific gaps in the available data. For example, if the available data were mostly in the form of official statistics such as amount and type of cargo shipped, number of vehicles using roads and gross economic activity, additional data collection might usefully gather vignettes of the effect of the transport program on individual people, households and villages. This additional data collection would also be intended to capture information from a range of perspectives about important unintended effects, whether positive or negative. For example, improvement of local roads can increase accessibility to central medical facilities, improving outcomes from childbirth complications. Examples of impacts arising from a small number of vignettes could be followed up by examination of official health statistics.

Next, a panel of experts in a range of relevant areas would be engaged to review the data in terms of the credibility of its claims both about describing impacts and about attributing them (either solely or in combination with other factors) to the program. While the PPSR approach does not specifically link the expert review with the additional data collection, it would be desirable for the additional data to be included in this review, and for the panel to recommend further data collection and analysis to address gaps in the evidence base. Finally a meeting including representatives of all major stakeholder groups would be convened to review the impact evaluation report. The final report would be published in different versions – a summary report, with major findings and vignettes providing examples of impact, and a detailed report with links to all the evidence on which the conclusions were based. PPSR therefore would provide a more comprehensive accounting to the various stakeholders, including the community, of the different impacts of the transport program than is likely using other approaches.

3. Anti-corruption program

This case was described as follows:

> *Donor support to an anti-corruption commission in an African country. The program includes helping develop guidelines, infrastructure upgrading and study tours. Similar programs are being implemented in six countries.*

It is not a standardised intervention, nor does it appear to be one that is tightly prescribed in advance. Instead the specific objectives of the program, and the means of achieving these, are likely to emerge as the program proceeds, and a better understanding of the priorities and possibilities is developed. For this reason, I would suggest use of 'developmental' evaluation (Patton, 1994, 2008), which is not intended to provide a report at the end of implementing a standardised and fixed intervention, but to provide information during implementation of a continually changing intervention with important complex aspects.

> "Developmental evaluation refers to long term, partnering relationships between evaluators and those engaged in innovative initiatives and development.

> Developmental evaluation processes include asking evaluative questions and gathering information to provide feedback and support developmental decision-making and course corrections along the emergent path. The evaluator is part of a team whose members collaborate to conceptualize, design and test new approaches in a long-term, ongoing process of continuous improvement, adaptation and intentional change. The evaluator's primary function in the team is to elucidate team discussions with evaluative questions, data and logic, and to facilitate data-based assessments of where things are, how are things unfolding, what directions hold promise, what directions ought to be abandoned, what new experiments should be tried – in other words, data-based decision-making in the unfolding and developmental processes of innovation. (Westley et al, 2006).

In this case, the impact evaluation would involve working with the implementers of the program and the participants to develop a program theory representing what they understand to be the major problems in terms of corruption, and how they intended to address them. This program theory would be developed, and revised over time for each specific initiative that was developed as part of the program, along with appropriate measures and ways of testing causal attribution. For example, if an initiative focused on a hotline for the public to anonymously report cases of corruption, the impact evaluation should follow up these cases (which could involve all cases, or a stratified random sample, depending on the numbers and the available resources) to see what happened in terms of where they were referred, what investigations were made, whether the claim was substantiated, and, if so, what the consequences were for the perpetrator.

By developing program theories for each initiative it would be possible to identify different theories of change that might be evident – for example, were the interventions intended to work through improving officials' understanding of probity requirements, through identifying, punishing and removing corrupt officials, through deterring corruption by increasing the risk of detection by formal audits or by making it more possible for the public to report corruption? Did particular theories of change apply to different types of corrupt practice or corruption in different types of programs?

If a primary purpose of the evaluation was to identify good practice and to translate these to other settings, a 'positive deviance' approach might be effective. This involves identifying sites or cases where extraordinary results are being achieved, verify this, and analysing what is producing the results. The significant feature of 'positive deviance' is that this investigation is done by those seeking to learn from the good practice. It is not done by an evaluation team and then disseminated later to information users. Positive deviance has been used in public health, nutrition, female genital cutting, education, and agricultural development, with a reported example in the area of corruption, specifically public functionary extortion demands (Horowitz, 2006).

One of the challenges of evaluating anti-corruption programs is the effect of corruption itself on the ability to gather accurate evidence and to take appropriate action as a consequence of findings. In this case, specific attention would need to be given to the management of the evaluation to ensure its independence and the safety of the investigators. This might include a direct line of report to a trusted level of government, to avoid findings being buried or changed, and/or the involvement of citizen advocacy organizations to gather or investigate claims of corruption.

4. Conclusion

Interestingly, each of these cases would include some component of theory-based evaluation, although program theory would be used in very different ways. In the first case it is used to identify intermediate outcomes that could indicate the achievement of longer-term impacts, and to identify contextual factors that should be investigated in analysis. In the second case, it is used as a conceptual framework to bring together diverse evidence about a diverse set of components. In the third case, it is used as a

conceptual framework to guide an evolving design to collect and analyse data to inform ongoing change.

The analysis of the cases, and their implications for impact evaluation, are summarised below in Table 2. While this table focuses on key aspects of the interventions, it is likely that all would have some degree of other aspects (for example, all cases would be likely to have some simple aspects).

Table 2 Analysis of cases in terms of characteristics and purposes

	Key aspects of intervention	Purposes of evaluation
Case 1 - CCT	Simple aspects – standardised intervention Complicated aspects – works in conjunction with other factors and programs	Learn if it works – and in what contexts it works – to inform ongoing policy
Case 2 – Transport infrastructure	Complicated – diverse multiple components that need to work together effectively	Assess the overall impacts of a completed program
Case 3 Anti-corruption	Complex – adaptive and emergent intervention, responsive to needs, problems and opportunities	Understand and improve an ongoing and changing program

The designs developed for these three cases have shown that situational responsiveness requires knowledge of a wide range of methods and techniques. Even if the actual conduct of the impact evaluation were to be contracted out to an external evaluator with relevant expertise and experience, the commissioning agency would need sufficient understanding of the method to be able to develop appropriate terms of reference, select an appropriate consultant, and effectively manage the contract, including assess the quality of the work. The implications of this for building evaluation capacity are that we need both depth (specialists in particular methods) and breadth (understanding that a range of methods exist, and when these might be most appropriate).

Acknowledgements
Thanks to Michael Patton, Kaye Stevens, Howard White, and Bob Williams for helpful comments on an earlier version of this paper.

References

Boyes, B. (2006). Determining and managing environmental flows for the Shoalhaven River, Report 1 - Environmental Flows Knowledge Review. NSW Department of Natural Resources, May 2006. Retrieved 14 May 2009 from www.dwe.nsw.gov.au/water/pdf/monitor_sholahaven_sh003.pdf

Chambers, R. (2009) Participatory methods. *Journal of Development Effectiveness,* this issue.

Clear Horizon and O'Connor NRM (2008) Performance Story Report: A study of the Mount Lofty Ranges Southern Emu-Wren and Fleurieu Peninsula Swamps Recovery Program and how it contributed to biodiversity outcomes in the Adelaide and Mount Lofty Natural Resources Management region. Canberra: Commonwealth of Australia. Retrieved 16 May 2009 from http://www.nrm.gov.au/publications/books/pubs/psr-mount-lofty.pdf

Dart, J. (2008) 'Report on outcomes and get everyone involved: The Participatory Performance Story Reporting Technique'. Paper presented at the 2008 Australasian

Evaluation Society conference, Perth. Retrieved 16 May 2009 from http://www.aes.asn.au/conferences/2008/papers/p100.pdf

Dart, J. (2009) 'Participatory Performance Story Reporting Technique'. Paper presented at the 2009 Impact Evaluation conference, Cairo.

Downes, B. J. L. A. Barmuta, P. G. Fairweather, D. P. Faith, M. J., Keough, P. S. Lake, B. D. Mapstone and G. P. Quinn (2002) Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters Cambridge: Cambridge University Press.

Glouberman, S. (2001) 'Towards a New Perspective on Health Policy', CPRN Study No. H/03, Canadian Policy Research Networks Inc., Ottawa.

Glouberman, S. and B. Zimmerman (2002) Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like? Commission on the Future of Health Care in Canada, Discussion Paper 8. Retrieved 14 May 2009 from http://www.hc-sc.gc.ca/english/pdf/romanow/pdfs/8_Glouberman_E.pdf

Guijt, I (2008) 'Seeking surprise : rethinking monitoring for collective learning in rural resource management'. PhD thesis. Wageningen, the Netherlands.

Hill, A.B. (1965) 'The Environment and Disease: Association or Causation', Proceedings of the Royal Society of Medicine 1965 May;58:295-300.

Horowitz, B. (2006) Bridge Ogres, Little Fishes and Positive Deviants: One-on-one deterrence of Public Functionary Extortion Demands. Retrieved 14 May 2009 from http://www.positivedeviance.org/projects/law/Bridges_final.doc

Kurtz C, Snowden D. (2003) The new dynamics of strategy: Sense-making in a complex and complicated world. IBM Systems Journal 2003;42(3):462-483.

Land and Water Australia ( ) Improving the Natural Resource Management System for Regions. Canberra: Australian Government Publishing Service. Retrieved 14 May from http://www.rkrk.net.au/images/3/34/PR061220.pdf

NONIE Subgroup 2 (Network of Networks on Impact Evaluation) NONIE Impact Evaluation Guidance. Retrieved 14 May 2009 from http://www.worldbank.org/ieg/nonie/docs/NONIE_SG2.pdf

NONIE (Network of Networks on Impact Evaluation) Statement on Impact Evaluation. Retrieved 14 May 2009 from http://www.worldbank.org/ieg/nonie/docs/IE_statement.doc

Norris, R.; Liston, P.; Mugodo, J.; Nichols, S. (2005) Multiple Lines and Levels of Evidence for Detecting Ecological Responses to Management Intervention. Paper presented at the American Geophysical Union, Spring Meeting 2005.
Patton, M. Q. (1994) 'Developmental Evaluation', Evaluation Practice, Vol 15, No. 3: 311-319.

Patton, M.Q. (2008a) Utilization Focused Evaluation, 4th ed. Text. Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (2008b). State of the Art in Measuring Development Assistance. Address to the World Bank Independent Evaluation Group. Conference, 10 April 2008, Washington, DC. Retrieved 14 May 2009 from http://www.worldbank.org/ieg/conference/results/patton.pdf

Ravaillon, M. (2009) Evaluating Three Stylized Interventions. *Journal of Development Effectiveness*, this issue.

Rogers, P.J. (2008a) Four key tasks in impact assessment of complex interventions, Keynote address. Workshop on Rethinking Impact. Understanding the Complexity of Poverty and Change, Consultative Group on International Agricultural Research (CGIAR) Cali-Colombia. Cali, Colombia. Retrieved 14 May 2009 from http://www.prgaprogram.org/riw/files/papers/Rogers%20material%20for%20workshop.ppt

Rogers, P.J. (2008b) 'Using programme theory for complicated and complex programmes' Evaluation: the international journal of theory, research and practice. 14 (1): 29-48.

Soares, F.V, R. Perez, G.I. Hirata (2009) 'Achievements and Shortfalls of Conditional Cash Transfers: Impact Evaluation of Paraguay's Tekoporã Programme'. Paper presented at the 2009 Impact Evaluation Conference, Cairo.

Stacey, R. (1992). Managing the Unknowable. San Francisco: Jossey-Bass.

Weiss, C. (1997). Evaluation: Methods for Studying Programs and Policies. (2nd ed.) Upper Saddle River, NJ: Prentice Hall.

Westley, F., B. Zimmerman, M.Q. Patton (2006) Getting to Maybe: How the World Is Changed. Random House Canada. Extract retrieved 14 May 2009 from http://innovationlabs.com/r3p_public/rtr3/pre/pre-read/Patton.Developmental%20Evaluation.pdf

NOTE

[1] NONIE is a Network of Networks for Impact Evaluation comprised of the Organisation for Economic Co-operation and Development's Development Assistance Committee (OECD/DAC) Evaluation Network, the United Nations Evaluation Group (UNEG), the Evaluation Cooperation Group (ECG), and the International Organization for Cooperation in Evaluation (IOCE)-a network drawn from the regional evaluation associations.