It's All About MeE:
Using Structured Experiential Learning ("e") to Crawl the Design Space

Lant Pritchett

Salimah Samji

and

Jeffrey Hammer

November 21, 2012

Abstract: There is an inherent tension between *implementing organizations*—which have specific objectives and narrow missions and mandates—and *executive organizations* --which provide resources to multiple implementing organizations. Ministries of Finance/Planning/Budgeting allocate across ministries and projects/programs within ministries, development organizations allocate across sectors (and countries), foundations or philanthropies allocate across programs/grantees. Implementing organizations typically try to do the best they can with the funds they have and attract more resources, while executive organizations have to decide *what* and *who* to fund. Monitoring and Evaluation (M&E) has always been an element of the accountability of implementing organizations to their funders. There has been a recent trend towards much greater rigor in evaluations to isolate causal impacts of projects and programs and more "evidence based" approaches to accountability and budget allocations. Here we extend the basic idea of rigorous impact evaluation—the use of a valid counter-factual to make judgments about causality—to emphasize that the techniques of impact evaluation can be directly useful to implementing organizations (as opposed to impact evaluation being seen by implementing organizations as only an external threat to their funding). We introduce structured *experiential learning* (which we add to M&E to get MeE) which allows implementing agencies to actively and rigorously search across alternative project designs using the monitoring data that provides real time performance information with direct feedback into the decision loops of project design and implementation. Our argument is that within-project variations in design can serve as their own counter-factual and this dramatically reduces the incremental cost of evaluation and increases the direct usefulness of evaluation to implementing agencies. The right combination of M, e, and E provides the right space for innovation and organizational capability building while at the same time providing accountability and an evidence base for funding agencies.

Introduction[1]

Any effective development project must ultimately be based on an adequate "theory of change"-- a complete, coherent, and correct causal model from funding to inputs and activities to outputs to outcomes and impacts. Any theory of change has to answer two "why" questions.

- Why will the agents of the implementing organization translate funding into inputs and inputs into activities that will create useful outputs?
- Why will the outputs produced by the project/program increase the well-being of the intended beneficiaries?

Answers to these "why" questions require positive behavioral models of how people (implementers and intended beneficiaries) respond to the opportunities created by the project. Projects can fail if either funding doesn't lead implementing agencies to produce outputs or if those outputs don't lead to better outcomes. An irrigation project can fail either because it doesn't actually produce a better water supply for the farmers or because water wasn't a key constraint to farmer output. An education project can fail either because funding doesn't expand the supply of educational opportunity or because supply wasn't the key constraint to expanding education. Micro-finance projects to promote new micro-enterprises could fail either because the project didn't provide greater availability of credit to potential borrowers or because credit was not a key constraint to business formation.

The key question is *how* and *when* these needed theories of change are discovered. One view is that projects are planned in detail in advance on the basis of a known theory of change for which there is rigorous evidence and implementation is just following the plan. Another view is that, while planning is useful, rapid feedback loops and learning in response to ongoing challenges is essential. Both views are important in any field of endeavor from military[2] to business[3] and development is also quintessentially about human beings and human systems and hence intrinsically complex.

The traditional approach to monitoring and evaluation (M&E) of development projects and its contribution to effective theories of change has been under attack on two fronts.

[1] We would like to thank Finn Tarp for comments on the penultimate draft, Sanjeev Khagram for his ideas and motivation about evaluation, and Matt Andrews, Ricardo Hausmann, Dani Rodrik, Michael Woolcock for on-going interactions, and several cohorts of HKS students for their reactions. UNU-WIDER provided financial support under the ReCom project.

[2] Military strategists have always known that, while planning is essential, the "fog of war" precludes detailed planning from generating certainty, as summarized in the adage "No battle plan survives the first contact with the enemy." Hence Napoleon's famous approach: "Engage with the enemy and see what happens"—which, it must be said, served him alternatively well and badly.

[3] Business theorists distinguish between "deliberate" strategy and "emergent" strategy (Mintzberg and Waters 1985) and emphasize that too slavish an adherence to a deliberate strategy can lead to massive business disasters. Bhide (2000) argues that 93 percent of successful start-ups did not follow their original strategy.

First, that traditional "M" is too focused on input utilization and process compliance and does not actually contribute useful information to project management. This leads to a vicious circle in which up to date and reliable monitoring data is not a priority for project management (as it does not feed into decision-making and management) and therefore monitoring data is not reliable or timely.

Second, that evaluation practice was based, at best, on crude "before and after" comparisons. Evaluations of impact on outcomes typically lacked any coherent counter-factual for evaluating the causal impact of project outputs on the outcomes for intended beneficiaries. This critique has led to a massive rise in the use of Rigorous Impact Evaluation (RIE) techniques, including Randomized Control Trials (RCTs)[4], and increased pressure on funding organizations that their activities be based on "rigorous" evidence about "what works."

In this paper, we extend the ideas behind RIE (and RCTs) by introducing *structured experiential learning* (little "e") for *implementing* organizations. Structured experiential learning builds learning objectives into the cycle of project design, implementation, completion, and evaluation. "e" helps implementers first articulate the "design space" of available project/program/policy alternatives and then dynamically "crawl the design space" by simultaneously trying out design alternatives and then adapting the project sequentially based on the results.

The use of an integrated approach to MeE releases the tension between implementing agencies and funders by balancing the space for implementers to innovate using experiential learning with the need for rigorous evidence of effectiveness from impact evaluations.

MeE is an integral part of a different strategic approach to development—that emphasizes the power of bottom-up driven innovation in building capability as well as success. Andrews, Pritchett, and Woolcock (2012) describe one variant of this approach to development called Problem Driven Iterative Adaptation (PDIA). This strategy emphasizes the role of development projects - not as scaling up known solutions using implementation by edict[5] - but rather as instruments for "experimenters" (in the broad sense of Rodrik 2008) or "searchers" (Easterly 2006) or to learn about what works to address specific, locally nominated problems in a particular context, for creating organizational capability and for mobilizing the commitment of implementing agents.

This paper is organized as follows: section 1 defines terms and discusses first generation M&E; section 2 discusses the second generation of M&E — the increased use of RIE and RCTs; section 3 highlights the need to move to the next generation of M&E, from

[4] It is important to note that not all rigorous evaluations use RCTs nor are all RCTs actually "evaluations" of actual projects. That is, many of the current RCTs are "field experiments" that are designed and implemented by researchers for the purposes of research on techniques rather than evaluations of actual development projects.

[5] The ideal of top down "planners" who attempt to reduce development to a series of logistical challenges.

experiments to experimentation; section 4 introduces structural experiential learning and provides a 7 step dynamic approach of how "e" can be used; section 5 discusses how MeE can be used as an organizational learning strategy for both implementers and funders of development projects. A conclusion, perhaps surprisingly, concludes.

1. First Generation M&E[6]

We follow standard practice as articulated in project planning or logical framework approaches and define a "development project" as *inputs* (financial and other resources), which are translated by an *implementing agency* into specified *activities* to produce useful *outputs*. These *outputs* have the goal of *outcomes* and *impacts* of higher well-being for the intended beneficiaries. A development *funding organization* provides resources to promote development. Development funding organizations range in structure from large multilateral organizations like the World Bank or the regional development banks (IADB, AfDB, ADB), UN agencies (UNDP, UNICEF), bilateral agencies (USAID, MCC, DFID), to purely private foundations (Bill and Melinda Gates, William and Flora Hewlett). Governments themselves often act as funding organizations by structuring expenditures into discrete projects and programs. *Funding organizations* typically structure their support into discrete *projects* carried out by *implementing agencies*. Implementing agencies also take a variety of forms and can be agencies of government (often units within a government responsible for implementing a particular project), private contractors, or NGOs that take on implementation responsibilities[7]. All of these development projects have the goal of improving the well-being of some target population, the *intended beneficiaries*[8].

Our definitions are both standard and are intended to include everything people consider a development project—and more. Building physical infrastructure or facilities (e.g. highways, schools, ports, health clinics, power plants) are development projects. Training is a development project. Social programs (e.g. conditional cash transfers, micro-lending) are development projects. Policy advocacy is a development project. Empowerment is a development project. Research is a development project.

It is worth pointing out that evaluation of a development project is itself a development project. Evaluation uses funds to finance inputs and *activities* (collection of data, analysis

[6] It is worth noting that we are not focusing on development projects in the belief that the success or failure of individual development projects is the major determinant of development outcomes. Many analyses attribute the vast majority of differentials in the improvement of human well-being to "institutions" or "policies" that promote broad based economic growth which leads to rising prosperity (Acemoglu et al, Easterly, Pritchett). For instance, recent rapid progress in poverty reduction in China or India or Vietnam, as well as the prior progress in East Asia (e.g. Korea, Taiwan, Indonesia) had little to do with "projects" as we define them, but does have to do with capable public sector organizations or, at least, the policies and projects they generate.

[7] Some development organizations do both fundraising and implementation, Save the Children, Oxfam-UK, often utilizing both their own raised funds and receiving funding from funding organizations.

[8] This definition is flexible enough to include *any* dimension of well-being (not just "economic") and includes as development projects activities that protect human rights or expand democracy or raise awareness about the natural environment.

of data) that produce *outputs* (reports, research papers, policy advocacy) by an *implementing agency* (in this case an evaluation organization) with the ultimate intention of producing better developmental outcomes for intended beneficiaries. Table 1 illustrates our delineation of the stages of a development project with a highly schematic summary of an array of development project examples.

Table 1: Examples of the wide range of development projects

	Inputs **(what is made available to the project)**	***Activities*** **(what the project does)**	***Outputs*** **(achievements that will lead to outcomes)**	***Outcomes*** **(changes external to the project)**	***Impacts*** **(long-run impact on well-being)**
Construction of a road	Financial and human resources and public authorization	Procurement of equipment, asphalt, labor, construction	A new road	Lowered transport costs	Higher incomes/lower prices
Promotion of better health practices (e.g. breastfeeding, HIV prevention)		Hire and train health workers, train existing workers with new messages	Trained health workers, communication materials developed	Changed behavior, better individual health outcomes	Improved population health and well-being
One-stop shop for Small Medium Enterprises (SMEs)		Create public officials/offices to facilitate SME regulatory compliance	One-stop shops created, easier regulatory compliance	Individuals and enterprises using one-stop shop	Higher productivity firms in compliance, higher incomes, more opportunity
Micro-credit		Hire workers equipped to make loans available	Loans made	Incomes increased, people empowered	Better livelihoods
Governance, Policy advice		Revise laws, procedures for civil service, train government workers	Laws changed, civil servants trained, analysis and policy recommendations	Government agencies working more effectively, policy advice being used	Reduced corruption, better services, greater citizen satisfaction with government
Advocacy for climate change		Design materials for campaign	Materials (print, audio, video, reports) created and disseminated	Changed beliefs of general public, key decision makers	Reduced damage from climate change
Impact Evaluation		Design evaluation, data collection and entry, analysis and findings	Report or paper with analysis and key findings of research	Use of research findings	Change in policy or behavior or beliefs

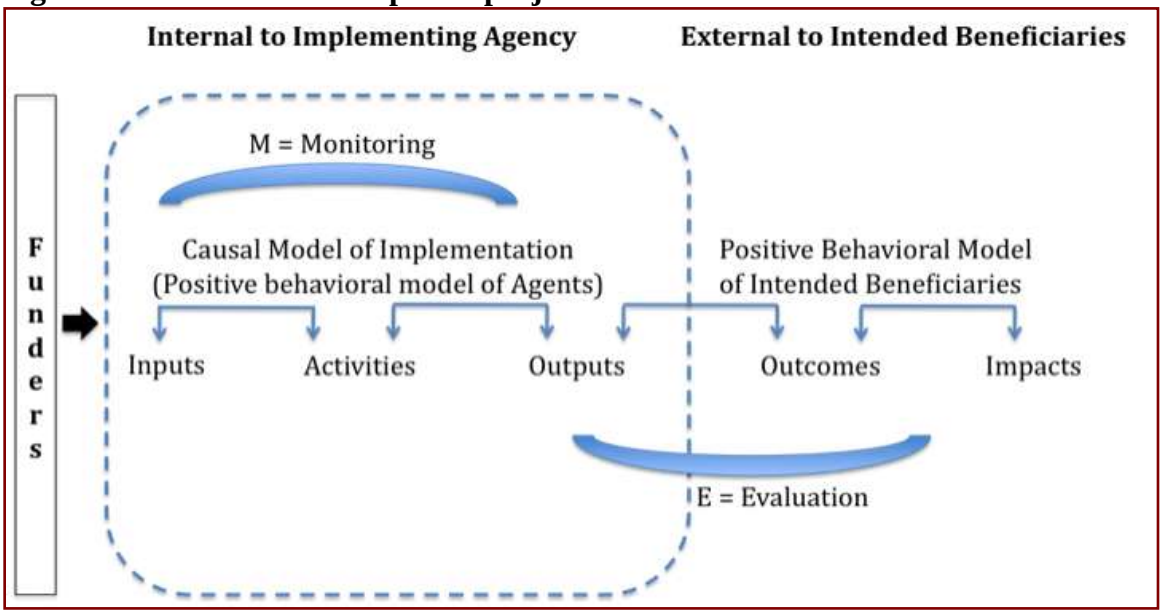
1.1 Traditional Learning from Development Projects: M&E

Monitoring and Evaluation (M&E) are routine, and nearly universal, components of externally funded development projects.

Monitoring is the regular collection of information to track implementation progress and is an integral part of reporting and accountability. Monitoring generates data *internal* to the development project and is focused on compliance, both in process and progress, with the project plans. Are *inputs* being used (e.g. is the project disbursing according to plans)? Are inputs being used according to acceptable processes (e.g. are procurement rules being followed)? Are the inputs used translating into the planned *activities*? Are those activities producing the expected *outputs?*

Monitoring is used by the implementing agency to manage the project and by funding agencies for accountability. Implementing agencies use monitoring data to track progress, identify bottlenecks and keep the project moving forward. Funding agencies use monitoring data for accountability, making sure that inputs, financial and otherwise, are used only for the agreed activities and follow the agreed upon processes[9].

Figure 1: M&E for a development project



[9] None of this is unique to development, or even in the public sector. In the private sector, this use of routinely collected data on process and progress in the utilization of funds is called "auditing."

Evaluation: While monitoring asks: "is the project doing things right?" evaluation asks: "is the project doing the right things?"— is the project an effective use of resources for its intended purposes? In practice however, "project evaluation" has been used to mean three completely different things, for which we propose three distinct terms: *project valuation, implementation evaluation* and *impact evaluation.* Three equations clarify the distinctions.

A *project production function* maps inputs into activities and outputs.

$$\boldsymbol{Project\ production\ function}\text{: } Outputs^P = f(Activities^P(Inputs^P))$$

A *beneficiary uptake* equation relates outputs of the project (P) to outcomes for beneficiaries (*k*).

$$\boldsymbol{Beneficiary\ Uptake}\text{: } Outcomes^k = g(Outputs^P, EE^k)$$

The *valuation* equation places a *value* to the well-being of beneficiaries of the outcomes *and* aggregates those values across beneficiaries.

$$(\boldsymbol{Social})\ \boldsymbol{Valuation\ of\ Outcomes}\text{: } \Delta SW = SW(\Delta WB^k(Outcomes(with) - Outcomes(without))$$

Valuation Evaluation. Historically, the first use of "project evaluation" was as a tool for *ex ante* analysis of development projects in a planning exercise to decide which projects to fund from a limited available budget (Little and Mirrlees 1969, Dasgupta, Marglin, Sen 1969). The main intellectual problem was the *valuation* of the outcomes of the project into a cost-benefit calculus (Dreze and Stern, 1987). That the inputs (costs) would produce the *outcomes* was simply assumed (that is, the project production function and beneficiary uptake equation were treated as known). The valuation evaluation question was whether the *value* of the outcomes as aggregated social benefits was worth the costs, compared to other available projects. When projects produce marketed goods at non-distorted prices and distributional issues are ignored, this reduces to the types of financial programming and cost-benefit analysis that private firms use (and in precisely those situations there is no rationale for public sector engagement)[10]. The difficulty that valuation evaluation addressed was allocating limited public sector funds across projects

[10] Perhaps the most egregious problem with the practice of *ex ante* project valuation in its heyday was that, as highlighted in Devarajan et al (1997) and Hammer (1997), the outputs valued were often purely private goods. Believe it or not, several training manuals on project evaluation for World Bank economists used the construction of a tomato canning factory as an example for project evaluation—almost certainly a private good. By ignoring the ultimate concern – how to improve well-being of society over and above what the private sector can do on its own, that is, how a project could correct or mitigate a market failure – measuring outputs and not outcomes (social well-being in this case) could lead to governments doing exactly the wrong things. This lack of measurement of *social* rather than *private* returns continues to undermine evaluation methods both old and the new impact evaluation often ignores these issues entirely. A major exception being the measurement of externalities of deworming children in Kremer and Miguel.

when market prices for inputs and outputs were distorted and hence "shadow prices" were needed for valuation, when markets were non-existent (e.g. environmental public goods), when valuing non-marketed outcomes (e.g. health), and when addressing distributional concerns (Squire and van der Tak 1976). *Valuation* evaluation of projects was about the financial, economic, and social (each of which could be very different) valuation of the project stream of inputs (as costs) and outputs (as benefits).

Implementation Evaluation. The second use of evaluation is *ex post* evaluation to certify that the project was implemented as designed. Did the project spend the money? Did the activities happen as planned? Were outputs produced? These evaluations might exist mainly for accountability purposes, both the accountability of the implementers to the funders (e.g. an agency to the funder/NGO that funded them) and the funders to the sources of funds (e.g. taxpayers). Nearly all funding organizations have implementation evaluation as a required part of the project cycle. Sometimes the rhetoric that there has been "too little" *rigorous* evaluation is confused with a claim there is too little evaluation—which is not true.

Impact Evaluation. The currently popular use of evaluation is in assessing the impact of the project on *outcomes* for the intended beneficiaries. This requires *ex post* measurement not only of internally generated data about inputs, activities, or outputs but also of *outcomes* which are external to the project. Impact evaluation requires a *counter-factual*—to know the causal impact of a project one has to know not just the factual (what did happen) but also what would have happened without the project. This includes people's behavioral responses to the project itself.

Table 2 outlines the types of evaluation with illustrations from different types of projects and the types of project "failure" the various types of evaluation can detect.

Table 2: Three distinct uses of "Project Evaluation" in development			
	Valuation Evaluation *Do the benefits (properly weighted and discounted) exceed the costs?*	**Implementation Evaluation** *Did the implementing agency succeed in doing what they said they would do in processes, activities and outputs?*	**Impact Evaluation:** *Did the project lead to the desired outcomes and impacts on the intended beneficiaries?*
Construction of a road	Does the predicted road volume justify the costs of reducing travel time by constructing the road?	Why didn't the inputs produce the outputs (i.e. why no quality roads)? - Corruption - Delays in procurement - Unanticipated weather - Poor engineering so roads washed away	Road was constructed but projected traffic volume did not materialize – why? - Economy wide recession - Monopoly of truckers
Promotion of better practices to improve nutrition	Are the costs of personalized promotion too high versus other ways of producing same nutrition gains (cost effectiveness)?	Why didn't the inputs produce the outputs? - Retention/staff turnover - Trained health workers don't think this is priority and don't change their behavior	Nutritional outcomes did not improve – why? - Beneficiaries, having received messages do not change practices - Messages were wrong
Micro-credit	Do the costs of providing credit at micro level have higher net returns than other uses of capital?	Why weren't loans made? - Loan officers do not generate lending activity - Low repayment rates decreases the total possible lending	Why did incomes not increase? - Little demand for borrowing - Borrowed money displaces other lending with little net impact - Borrowed money used in low return activities so net income small
Impact Evaluation	Is the scope of the findings sufficient to justify time and cost of evaluation?	Evaluation not completed even after baseline is done. - Project is not carried out - Contaminated experimental design - Poor quality data collection	Evaluation results have no impact on beliefs or behaviors of key actors.

2. **Second Generation: The rise of the *randomistas*[11]**

> *Often, what passes for evaluation follows a two-two-two model. Two contractors spend two weeks abroad conducting two dozen interviews. For about $30,000, they produce a report that no one needs and no one reads. And the results they claim often have little grounding in fact. ... Today, I'm announcing a new evaluation policy that I believe will set a new standard in our field. By aggressively measuring and learning from our results, we will extend the impact of our ideas and of knowledge we helped generate. Every major project will require a performance evaluation conducted by independent third parties, not by the implementing partners themselves. Instead of simply reporting our results like nearly all aid agencies do, we will collect baseline data and employ study designs that explain what would have happened without our interventions so we can know for sure the impact of our programs.*
>
> *Raj Shah, USAID Administrator, January 2011*

In the last ten years there has been an accelerating rise in the criticism of traditional M&E and a corresponding rise in the prominence given to the use of rigorous techniques for project evaluation. The criticisms of M&E have not been that there is not "enough" M&E—in most mainstream development funding organizations M&E is built into every single project[12]. The criticism is of M&E practice that has two key elements:

- evaluation was too *ex ante* and needed to be more *ex post*,
- evaluation should be more focused on the impact on *outcomes* not just inputs, and based on a rigorous *counter-factual.*

Demise of ex-ante project valuation as a decision tool. For reasons both good and bad *ex ante* project valuation for decision making has more or less disappeared. Even in agencies that once used and promoted the technique and insisted on cost-benefit analysis as part of project preparation, like the World Bank, its use dwindled (see Warner 2010). Students in development economics today routinely complete their studies with no exposure even to the theory, much less the practice, of project *valuation evaluation*. This demise has had serious deleterious contexts as even if one can specify the entire logical framework or project inputs, outputs and outcomes without some idea of valuation these alone cannot be decision tools. Part of our MeE motivation is to bring valuation back into design by at least asking how large outputs and outcomes would need to be for a project to be an attractive development activity.

---

[11] This term can be attributed to Angus Deaton (2009) and expresses the view that randomization has been promoted with a remarkable degree of intensity.

[12] For instance, in the World Bank *every* project has an *ex post* evaluation conducted to assess the impact of the project by the unit responsible for project implementation. These *ex post* evaluations were reviewed by a part of the World Bank—once called the Operations Evaluation Department (OED), now called IEG—that was autonomous from management and answered directly to the World Bank's Board, who were representatives of the shareholders. On selected projects this group also carried out an independent evaluation. (And OED/IEG would periodically carry out "thematic" evaluations of all, say, "directed credit" or "integrated rural development" projects). Every other assistance agency we know of also had policies of evaluating its projects. There has *never* been any debate that development projects should be evaluated.

Rigorous counter-factual. By far the most influential critique was that funding agencies (including governments) relied on *implementation evaluation* which, when it contained estimates of project impact at all (as opposed to reporting only on project compliance with use of inputs and production of activities and outputs) rarely had any *counter-factual.* Implementation evaluations' estimates of impact used simple "before and after" comparisons, or compared project area outcomes to non-project area outcomes after the project. There are two methodological issues with "before and after" and "project and non-project" as estimates of the "with and without" impact of a project.

First, "before and after" assumes the counter-factual to the project was no change in outputs or outcomes. One might think this point obvious beyond belaboring, but the temptation to claim project success if outcomes have improved is powerful. A recent "evaluation" of the Millennium Villages in Kenya compared cell phone use in the project villages before and after the project and claimed this increase as a project impact, ignoring the obvious point that technological and market factors have independently led to increased cell phone ownership all across rural Kenya[13]. Another example is that India's program for expanding primary education enrollments has been widely declared a success because *total* enrollments in India increased. However in some states of India *public sector enrollment* (the only type of schooling supported by the project) went *down* in absolute numbers.

The second problem with using either "before and after" or "project and non-project" area comparisons is that the purposive selection of project areas or the self-selection of individual beneficiaries into participation in project areas. For example, suppose the effectiveness of a weight loss program was demonstrated by comparing the weight loss program joiners versus non-joiners. Joiners could easily be more motivated to lose weight than non-joiners and this motivation itself explain observed weight loss, independently of any causal impact of the program. Selection problems also potentially affect project placement. If, after a school construction project an evaluation compares enrollments in project and non-project areas this may overstate or understate the impact of the project depending on how school construction was located subject to the intended benefits[14]. Even "differences in differences" impact estimates (comparing "before and after" across "project and non-project" areas) are suspect unless the trajectory of the non-project areas reliably estimates the "without the project" counter-factual for the project area. It is important to note however that endogenous placement can be a good thing and an essential feature of project design. For example, project locations might be chosen precisely because those are the places the project is likely to work. Extrapolating the

[13] See http://blogs.worldbank.org/africacan/millennium-villages-project-continues-to-systematically-overstate-its-effects.

[14] If the project selected areas for school construction based on estimates of pent-up demand, then enrollments were likely to have grown in project areas even without the project and standard project versus non-project area comparisons would overstate project impact. If, on the other hand, the schools were placed where the Ministry felt they were most needed, that is, where enrollments were low because education was not valued by parents a fact unknown to the Ministry, the estimator could *understate* the potential impact of the program since the schools were built in the most difficult circumstances.

effect to other places or the average place would be seriously overstated. On the other hand, if the project isn't expected to work in those places, why do it there?

Table 3: Estimating change in the average outcome (Y) due to a project: "before and after" versus "with and without"

	Before	After	Difference over time
Project	$\bar{Y}^{Project,T}$	$\bar{Y}^{Project,T+N}$	$\bar{Y}^{Project,T+N} - \bar{Y}^{Project\,T}$
Non-Project	$\bar{Y}^{No,T}$	$\bar{Y}^{No,T+N}$	$\bar{Y}^{No,T+N} - \bar{Y}^{No,T}$
Difference non-project and project exposure	$\bar{Y}^{Project,T} - \bar{Y}^{No,T}$	$\bar{Y}^{Project,T+N} - \bar{Y}^{No,T+N}$	$(\bar{Y}^{Project,T+N} - \bar{Y}^{Project\,T})$- $(\bar{Y}^{No,T+N} - \bar{Y}^{No,T})$ (Differences in differences)
Difference in outcome with and without the project		$(\bar{Y}^{Project,T+N} - \bar{Y}^{Without\ Project,T+N})$ (the latter is unobservable)	

The problems with inferring *causal* impact from observational, non-experimental data have been well known for decades in many fields, from public health to agronomy to psychology to economics. There are many statistical methods for recovering an estimate of the "treatment effect" (from project to outcomes) of a project—propensity matching, regression discontinuity, instrumental variables—even without an *ex ante* experimental design (Angrist 2010). We use the term "Rigorous Impact Evaluation" (RIE) to mean any of the variety of methods of estimating causal impact which take into account identification issues (Ravallion 2011). Many consider the "gold standard" of RIE to be a prospectively designed Randomized Control Trial (RCT). A well designed RCT produces an estimate of the causal impact called the Local Average Treatment Effect (LATE)—because the estimate is "local" to the range tested and it is the "average" over potentially heterogeneous impacts across the treated units. RCT LATE estimates are internally valid, that is, rigorous evidence of impacts when applied to exactly the same program in exactly the same conditions.

In part as a response to the critiques of the weaknesses of previous approaches to M&E there has been a massive shift towards RIE in development projects and a concomitant rise in RCTs[15]. Since the PROGRESA evaluation there has been a veritable explosion in the number of RCTs being done in the developing world by academics, foundations, and development organizations[16]. J-PAL (as of June 2011) had 116 studies completed or

[15] The use of randomization and *ex ante* control trials in social projects and programs was not itself an innovation as these have been widely, if not routinely, used in the USA at least since the 1970s (e.g. the "negative income tax" experiments (1968-79), the Rand Health Insurance experiment (1974-82), housing, evaluation of the Job Training and Partnership Act (JTPA), community policing (1979)).

[16] The influential breakthrough in development projects was the use of an independent team of academics to do an impact evaluation of a conditional cash transfer scheme, PROGRESA (since renamed

underway and IPA had over 500 staff working around the world[17].

In 2004 the Center for Global Development (CGD), with support from the Gates and Hewlett foundations, launched the "Evaluation Gap Working Group" headed by Nancy Birdsall, Ruth Levine, and William Savedoff at CGD to examine what could be done to improve evaluation in development projects. This group produced a report in 2006—"When Will We Ever Learn?" that made recommendations for improving the support for impact evaluations. This resulted in a new organization 3ie (International Initiative for Impact Evaluation), which: "*funds quality studies that will have a real policy impact and affect many lives. In terms of standards, this means only studies that are built around a credible counterfactual with an evaluation design based on the underlying programme theory to learn what works and why, and also at what cost."*

Most development organizations have responded to the critiques of, particularly, impact evaluation within their overall evaluation approach and have been promoting greater use of impact evaluation, many for a decade or more[18].

3. **Next Generation: From Experiments to Experimentation**

The use of more RCTs and more RIE in development funding organizations is an important advance[19]. However, while M&RIE is an improvement on traditional M&E, it is insufficient as a learning strategy for development funders and implementing agencies. In much of its current practice RIE is still a tactic which is often still embedded in top-down *strategies* for implementation and learning in development projects but, as we emphasize RCT and RIE is also be a valuable tactic in alternative project learning strategies.

There are three fundamental reasons why M&RIE needs to be supplemented by structured experiential learning ("e").

- A rugged and contextual fitness function over a high dimensional and complex design space implies that learning "what works" has to be flexible and dynamic.

Oportunidades), in Mexico. This was influential as it was a rigorous evaluation of an ongoing government program carried out at scale in a developing country.

[17] It is worth noting that many of the ongoing RCTs are not evaluations of an ongoing project (with its necessary bureaucratic or other constraints) being funded by a development agency and implemented. Rather, they are "field experiments" in which the "intervention" evaluated is not of an ongoing activity of an existing funding or implementation agency but rather of an activity undertaken as an experiment and often in effect implemented by the research oriented organization.

[18] For instance, the World Bank's research group has been promoting building RCTs into Bank operations since at least the mid-1990s.

[19] Keeping in mind that RCTs run the spectrum from "project evaluation" of activities already being implemented at scale (e.g. the evaluation of PROGRESA) to "field experiments" in which academics essentially implement their own (or work with an NGO) to do a small project to do a study so there are many more RCTs than RCT project evaluations.

- Many development problems are problems of implementation—moving from inputs to outputs for which an *impact* evaluation that measures outputs to beneficiaries is not yet needed.
- The use of RIE is not yet typically embedded in a realistic positive model of how organizations and systems actually learn.

This is not a critique of the fundamental idea behind the use of RIE or RCTs but the opposite, what we propose is an *extension* of that idea. But rather than thinking of RCTs as only about impact evaluation of *outcomes* we propose the more active use of the principles and practices of RCTs: specification of alternatives, rigorous counter-factuals, and increased real-time measurement to learn about project efficacy to learn about causal models *inside* the implementing agencies and organizations[20].

3.1 Learning with a high dimensional design space and rugged and contextual fitness function

Imagine you run the experiment of drilling for water at spot X on the surface of the earth on September 1st 2012. Suppose you find water at exactly 10 feet deep. What have you learned? What if you drill a hundred feet northwest? A month later? Without a theory of hydrology, and contextual factual information such as seasonal rainfall patterns and run-off and knowledge of the surface and underground topology, your experiment taught you nothing useful. *Every useful statement is about the future*—what *will* be the outcomes I care about if I do Y versus doing Z--and *experiments can only make rigorous statements about the past*.

3.1.1 High dimensional design spaces

Try and answer the question: "Does the ingestion of chemical compounds improve human health?" It is obvious that the question is ridiculously under-specified as some chemical compounds are poison, some are aspirin or penicillin and huge numbers have no impact at all. With chemical compounds one has to specify a particular compound and the particular conditions under which it is expected to help.

Names of development projects are labels for *classes* and any specific project is an *instance* of a class of that type of project. A *micro-credit* project, a *nutrition* project, an *HIV prevention* project, a *teacher training* project, a *road construction* project, a *conditional cash transfer* project, a *privatization* project, a *community block grant* project, a *livelihoods* project. A class of projects designates a design space, which is the space of all of the possible instances of that class arrived at by specifying all of the choices necessary for a project to be implemented.

Design spaces of development projects are *high dimensional*.

[20] where Monitoring of outcomes becomes routine (i.e. for governments, monitoring can be outside of project areas (but inside its area of concern – i.e. the whole country) and sets up the opportunity for mini-research projects on what is really working. We don't necessarily have to wait 3 or 4 years to see how things are ultimately going to work.

Take the class of *Conditional Cash Transfer (CCT)* projects. Each *dimension* of the *design space* of a CCT project is one of the choices that have to be made to make a project implementable: who exactly does what, with what, for whom, and when. The operational manual of a "simple" project may run to hundreds of pages. Table 4 illustrates that even the simplest possible characterization of the design space of a CCT project has eleven dimensions. Even if there were only 3 discrete elements (which is a radical simplification as some dimensions have many more choices and some dimensions are continuous) in each of 11 dimensions there are 3^{11}=177,147 distinct CCT projects each of which is an *instance* of the *class* "CCT project."

The design space is also a complex space as the *elements* within each dimension—are often discrete and with no natural metric. For instance, in a CCT project the dimension of "magnitude of the transfer" has a natural metric in units of currency (or scaled as percent of household income in the project area) so that "more" and "less" have a natural and intuitive meaning. But what about the design space dimension of whether the transfer goes to the mother exclusively or to a legally designated head of household or to the father? How far apart are those in the design space dimension of "recipient"?

And CCTs are simple. Think of a "teacher training" project or a "micro-finance" project or a "road construction" project or a "livelihoods" project. Everyone who has ever had to design and implement a development project knows the fine *granularity* at which development happens on the ground.

Table 4: Design Space for CCT projects, illustrated with three specific CCT projects			
Dimension of design space of a CCT	**PROGRESA, Mexico (Oportunidades)**	**Red de Protección Social, Nicaragua**	**Malawi**
Who is eligible?	Poor households (census + socioeconomic data to compute an index)	Poor households (geographical targeting)	District with high poverty and HIV prevalence.
To whom in the household is the transfer paid?	Exclusively to mothers	Child's caregiver (primarily mother) + incentive to teacher	Household and girl
Any education component to the CCT?	Yes – attendance in school	Yes – attendance in school	Yes – attendance in school
What are the ages of children for school attendance?	Children in grades 3-9, ages 8-17	Children in grades 1–4, aged 7–13 enrolled in primary school	Unmarried girls and drop outs between ages of 13-22
What is the magnitude of the education transfer/grant?	90 – 335 Pesos. Depends on age and gender (.i.e. labor force income, likelihood of dropping out and other factors).	C$240 for school attendance. C$275 for school material support per child per year.	Tuition + $5-15 stipend. Share between parent ($4-10) and girl ($1-5) was randomly assigned.
How frequently is the transfer paid?	Every 2 months	Every 2 months	Every month
Any component of progress in school a condition?	No	Grade promotion at end of the year.	No
Any health component of the CCT?	Yes – health and nutrition	Yes - health	Yes – collect health information
Who is eligible for the health transfer?	Pregnant and lactating mothers of children (0-5)	Children aged 0–5	Same girls
What health activities are required?	Mandatory visits to public health clinics	Visit health clinics, weight gain, vaccinations	Report sexual history in household survey (self-report)
Who certifies compliance with health conditions?	Nurse or doctor verifies in the monitoring system. Data is sent to government every 2 months which triggers food support.	Forms sent to clinic and then fed into management information system.	

3.1.2 Rugged and contextual fitness functions

The impact of a development project (whether outputs or outcome or impacts) can be thought of as a *fitness function* over the design space. Conceptually a "fitness function" is a evaluative function over a design space (in evolution fitness this could be species survival over genetic designs, in software engineering fitness could be execution time over a design space in coding, in marketing fitness could be sales over a design space of alternative advertising, in cooking fitness could be meal tastiness over a design space of recipes, etc.). Learning about the efficacy of development projects is an attempt to empirically characterize fitness functions. There are two issues that will make learning from experimentation difficult.

The fitness function may be *rugged* in that seemingly "small" changes in project design can have big changes on outputs or outcomes or impacts.

Second, the fitness function may be *contextual* in that the mapping itself from design space to impact differs from context to context. Even doing *exactly the same project* as an instance in the design space can have different impacts depending on where and when it is done.

Rugged fitness functions: non-linear and interactive. Perhaps the most (if not only) thing that has been robustly learned from the "new experimentalism" in both behavioral economics and field experiments in development is that seemingly small changes in project design can have big impacts on outcomes. This is consistent with a fitness function that is rugged over a complex and hyper-dimensional design space. Here is an example.

Non-linear fitness functions. A number of experiments have found sharp non-linearity in impacts along a single dimension of the design space. For example, Cohen and Dupas (2010) that moving from 100 percent to 90 percent subsidy (from zero price to 60 cents) for insecticide treated bed nets reduced demand by sixty percentage points[21]. While cash transfers were shown by PROGRESA to impact school enrollment, an evaluation in Malawi (Baird, McIntosh, and Ozler 2009) found that the *size* of the cash transfer did not make a difference to the magnitude of the impact on enrollment, so that there is a non-linear impact where some cash has an impact but more cash (over the ranges tried) does not lead to more impact.

Interactive fitness functions. The second way in which the fitness function is "rugged" is that different design parameters are potentially *interactive* so that changes in some design parameters don't matter at all at some settings of design but do matter at others.

An experiment to look at cheating used students at Carnegie Mellon in a staged experiment. The subjects saw one person (a hired actor) clearly and publicly cheat with no consequences. When the cheating actor wore a plain white t-shirt then the public cheating led to 25 percent more students cheating. But when the actor wore a t-shirt that said "University of Pittsburg" (the cross-town rival of Carnegie Mellon) cheating only increased by 3 percent (Ariely et al. 2009).

A recent evaluation of providing extra teachers to reduce class size in Kenya found that providing an extra teacher did not improve child learning if the teacher was a regular civil

[21] In their review of findings from randomized experiments Holla and Kremer (2009) suggest that this unexpected and puzzling non-linearity around zero cash price has been found in a number of instances in health and education. This is particularly puzzling because in many instances the cash price is a small part of the opportunity cost (e.g. school fees as a fraction of total opportunity costs of schooling) so a sharp discontinuity around zero cash price is unexpected since there is no similar discontinuity in the total cost.

service hire but an extra teacher to reduce class size did improve student learning if the teacher was a contract hire (Duflo, Dupas, Kremer 2007).

The evaluation of cash transfers in Malawi discussed above (Baird, McIntosh, Ozler 2009) found that if the cash transfer went to the child and not the parent the impact on schooling was less when the transfer was unconditional but the impact was the same when the transfer was conditional, which shows the interactive of two design features (to whom the transfer is given with whether or not the cash transfer is conditional).

A very recent evaluation (Barrera-Osorio and Filmer 2012) examined the choice of recipients of scholarships in Cambodia between "poverty based" and "merit based" and found that while both raised enrollment only the "merit based" scholarships produced higher student learning.

An experiment in the impact of expansion of contraceptive access on contraception use and unwanted fertility in Zambia (Ashraf, Fields, Lee 2010) found that providing information and a voucher for contraceptives to *couples* led to no reduction in unwanted births compare to the control group. However, if the information and voucher was provided to a woman alone (without her husband present) there was a substantial increase in use of contraceptive methods that could be hidden from the spouse (e.g. injectables) and a decline in unwanted fertility.

A study of the uptake of consumer finance in response to mailed advertising found that including a picture of an attractive woman in the pamphlet increased demand by as much as a 25 percent reduction in the interest rate (Bertrand et al. 2010).

The ruggedness of the fitness function over a complex and high dimensional design space can account for the frequency of negative and seemingly contradictory findings. A review of the RCT evidence about HIV/AIDS prevention (Over 2011) found that "among the 37 distinct trials of 39 interventions to reduce HIV infection only five have found a benefit (Padian at al. 2010). Of these, three have produced strong evidence that adult male circumcision reduces a man's chance of infection by somewhere between 33 and 68 percent, one shows promise for a vaccine, and one, which finds HIV-prevention benefits to treating curable sexually transmitted infection (STI), is contradicted by other equally rigorous experiments."

Roberts (2004), writing about strategies of private firms, argues that we should routinely expect high degrees of interaction among various strategies of the firm as they have to cohere to be effective. Roberts uses the example of "performance pay" which is only one element of an organizations overall "human resources" strategy. Further, human resources strategies are themselves just one element of a private firms overall strategy, as they also have a marketing strategy, a production strategy, a financing strategy. One might call the collection of these strategies a corporate "culture." He points out that even if one experiments with randomized techniques to look at the impact of changes in "performance pay" one could consistently find no impact of performance pay—even experimenting with various performance pay designs—if performance pay was

inconsistent with other elements of the company's human resource strategy or corporate culture. But it is possible that *simultaneous* changes in linking pay to performance *and* changes in human resource and production process strategies could potentially have huge effects. Roberts argues that the practice of promoting "best practice" for firms element by element (e.g. 'best practice' performance pay, 'best practice' production process, 'best practice' marketing) makes no sense at all when there are, generically, interactions amongst these elements.

Similarly, Barder (2012) in his discussion of development and complexity illustrates the rich set of interactions between a large number of adaptive agents (people, firms, organizations, institutions) all of which are co-evolving. He argues that the "normal state of affairs is not linear systems, but complex non-linear systems."[22]

Just to visualize in an extremely simple case of a design space with two dimensions (over design parameter 1 and 2) and three design choices per dimension (A,B,C or I,II, III) for a total of nine possible designs (as opposed to the millions of design space elements in a real project). Figure 2a shows a "smooth" fitness function that is linear and non-interactive. The beauty of a known smooth fitness function is that an experiment comparing project A-I to A-II is also informative about A-II versus A-III (by linearity) and informative about B-I versus B-II (by non-interaction). Figure 2b illustrates a "rugged" fitness function (like the Swiss Alps). Clearly one experiment comparing project A-I versus A-II is completely uninformative about design space option A-II versus A-III and about B-I versus B-II.

[22] http://www.cgdev.org/doc/CGDPresentations/complexity/player.html

Figure 2: Comparing a "smooth" and "rugged" fitness function over a project design space

Figure 2a: Smooth—linear, non-interactive

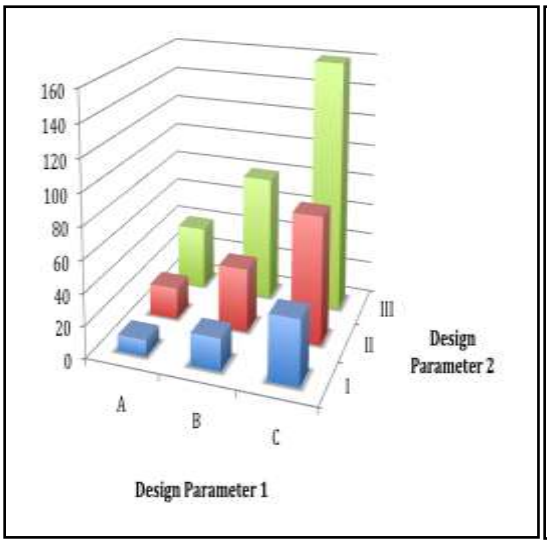
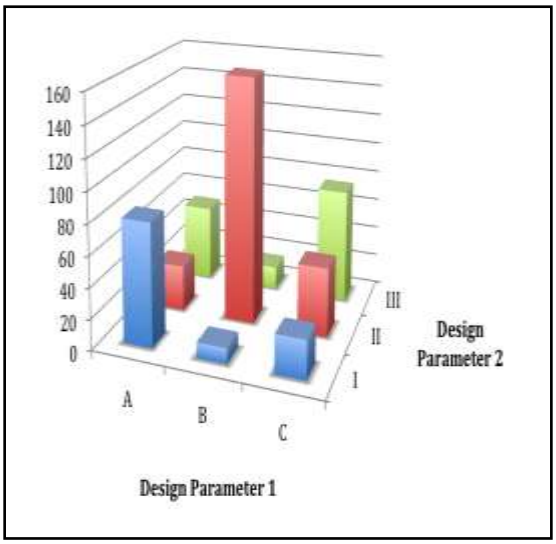


Figure 2b: Rugged—non-linear, interactive



Contextual (including dynamic) fitness functions. By "contextual" we mean that the shape of the fitness function over the design space may vary because of features of the *context* that are not under the control of project designers and hence not elements of the project design space. For instance, a project may require a mechanism for "enforcement"—like getting rid of staff who steal from the project. But while the project may produce a design to let such staff go employment law in the country might make such action theoretically possible but practically impossible. Even the exact same project from a *design* perspective (which, as seen above is itself difficult to reproduce given the complexity of the design space) may have very different outcomes depending on the context.

RIE/RCT evidence to date suggests fitness functions are contextual. Just as one example, there have now been a substantial number of rigorous estimates of the impact of "class size" on learning and they are completely different. Some find class size impacts large enough to suggest reducing class size is a cost-effective intervention (e.g STAR in Tennessee and the Maimondes rule in Israel). Others find class size impacts of exactly zero (e.g. Kenya and India).

3.1.3 What is learned from experiments in high dimensional design spaces and rugged and contextual fitness functions?

Every field experiment or impact evaluation of a development project must face all of these challenges, which we illustrate with just one example. Banerjee, Duflo, and Glennerster (2008) report on an experiment aimed at increasing the attendance of

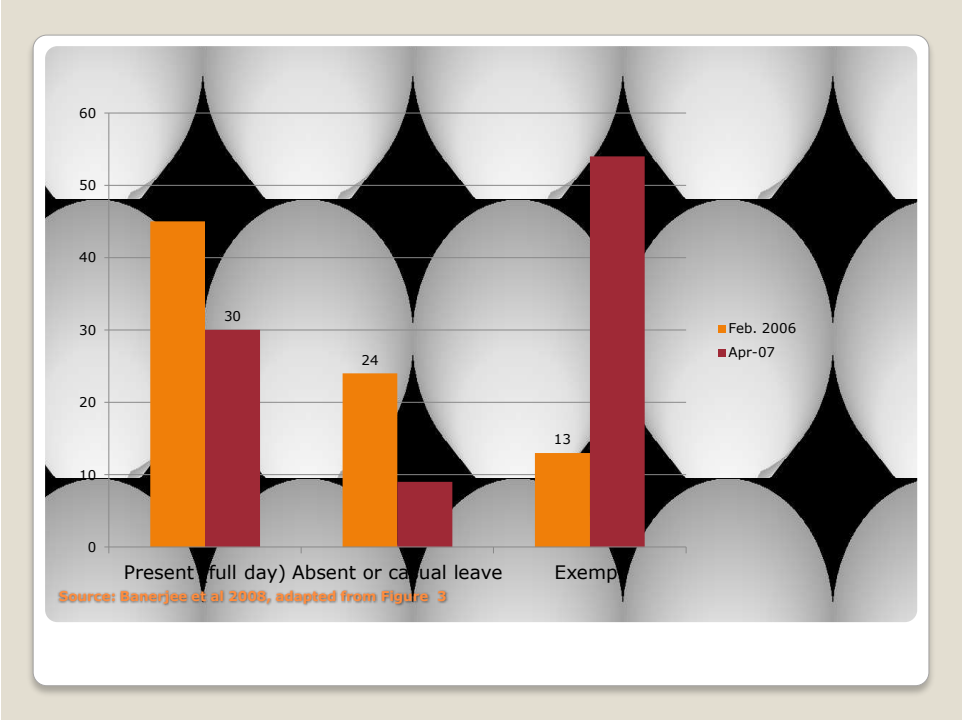
Auxiliary Nurse-Midwives (ANMs) at clinics (health sub-centers) in Rajasthan. This is an experiment about *implementation* as the mapping from *inputs* (funds of the Ministry of Health) to *activities* (e.g. introducing bonuses for attendance, placing time clocks, monitoring nurse attendance) to *outputs* (nurses present at clinics plus perhaps some health services). On a simplistic level this could be described as an experiment testing whether bonus pay would increase attendance of workers. But "bonus pay" is not a description of a *project*, it is a *label* for a *class* of projects. To specify the project as one instance of the class of "pay for performance" projects one has to fill in all of the elements of the design space, as in Table 5 below.

Table 5: (Simplified) design space of a "pay for performance" experiment

Elements of design space for a "pay for performance" policy	**Choice made in the BDG (2008)* experiment with ANMs in Rajasthan India**	**Other possible choices of design parameters**
Who will bonus apply to	Only additional (newly hired) nurses.	All nurses (including incumbents), nurses who "opt in", nurses in rural clinics, etc.
How much more (less) will be paid if attendance is adequate (inadequate)?	If absent more than 50 percent, pay reduced by number of absences recorded by NGO.	Continuum from small amounts (10 percent) to 100 percent of pay docked.
What is the threshold level of attendance needed to receive the bonus pay/not be docked in pay?	50 percent of the time on monitored days.	Continuum from small amounts (10 percent absence) to ever showing up.
How is attendance administratively recorded?	Introduction of time-date stamping machines.	Discrete alternatives: Status quo, biometrics, cameras, etc.
How are administrative attendance double checked for validity/ground-truthed?	Use of civil society volunteers to randomly show up at clinic and record physical presence of ANM.	Discrete alternatives: No double checking, community reports, peer monitoring, supervisors from Ministry, etc.
How are duties of ANMs defined with respect to physical presence at clinic?	Introduction of "clinic days" to reduce discretion of ANMs in attendance at clinic versus other duties.	Discrete alternatives: no change, specification of hours of the day, different frequency of "clinic days" (e.g. twice a week, once a month).
*Source: Banerjee, Duflo, and Glennerster. (2008)		

The results of their experiment are displayed in Figure 3 which shows that 16 months into implementation (August 2007) the attendance between the "treatment" and "control" additional ANMs on "monitored days" at the two ANM centers was indistinguishable—with less than a third physically present in either case. Attendance of the "treatment" additional ANMs had actually fallen steadily over the implementation period. The proximate explanation was the "exemptions" that allowed nurses to not be physically present while not affecting their absences for purposes of pay skyrocketed (to over half of all days).

Figure 3: Results of an experiment with bonus pay and increased monitoring of attendance for auxiliary nurse-midwives (ANMs) in Rajasthan India



What is learned from this experiment for pay for performance? That incentives don't work? No, there is too much evidence that incentives do work in many other contexts. That increased monitoring using new technology doesn't increase attendance? No, because another randomized experiment by one of the same authors does show that using time-stamped cameras as a new monitoring technology in classrooms in rural schools increased teacher attendance enormously—and this experiment involved the same NGO and some of the same researchers (Duflo and Hanna 2007). That bonus pay doesn't work for nurses? That the bonus pay was too small? That the time machines didn't work but biometrics would have? That civil society engagement was too weak? That enforcing attendance is impossible when it is possible through corruption to buy exemptions from attendance?

All that was learned was that this particular *instance* of the class of pay for performance schemes in this particular place at this particular time did not change attendance. It could have been that a minor change in design of the project would have led to massive impacts on attendance. It could be that *exactly* this design could work in another context (even another state of India). It could be that this project with many *fewer* features (e.g. without civil society engagement) could work in a different context. That I drilled to 50 feet down and did not hit water right here or that I gave this chemical compound to people and their health did not improve does not provide very much information.

3.1.4 The Limits of Experiments as a Learning Strategy

Think of learning about project effectiveness as an optimization algorithm, which is a plan for sequenced, iterative, outcome contingent, evaluation of the fitness function at various points of the design space. What do we know from optimization?

- If the design space is low-dimensional then a simple grid search is a feasible optimization procedure (especially if cost per evaluation of the fitness function is low in time and resources).

- If the fitness function is known from validated theory to be smooth (e.g. quadratic and non-interactive) then an optimization procedure can take advantage of that and a relatively small number of evaluations along a given dimension can pin down the fitness function's shape (e.g. marginal returns at various points) quite easily.

- If the fitness function is non-contextual (or its invariance laws are known) [23] then one can use evidence from one context to make predictions about fitness functions in another.

These three properties (low-dimensional, smooth, context free) are exactly what we know, always from theory and more recently amply buttressed by the new experimental (both impact evaluation, field experiments and laboratory experiments on behavior) evidence, are *not true* of development projects. We know from theory that development projects involve people, who are the ultimate complex phenomena, embedded in organizations, which are complex, and organizations are embedded in rules systems (e.g. institutions, cultures, norms) which are themselves complex. It would have been a staggering and wholly unexpected empirical discovery if, in spite of the known complexity of development projects, it had been shown there "the evidence" about "what works" made sense as a way of talking about development projects. It is not at all surprising that the existing experimental results so far mainly resist any simple summary—even in domains like micro-finance or education or 'incentive pay' where there have been many experiments.

In the literature on organizations there is a distinction between problems that are *simple, complicated,* and *complex.* Pritchett and Woolock (2004) and Pritchett (2012b) have extended this into development projects using analytic criteria to distinguish five different types of tasks, two of which (*policy making/elite services* and *logistics*) are implementation simple or complicated while the other three are complex. At this stage in the development process (at least 50 years into self-conscious promotion of development)

[23] Invariance laws describe how the measured quantity varies with respect to alterations in the conditions under which the experiment is carried out—e.g. is the experiment invariant with respect to non-accelerating reference frames--it is still possible only one experiment is needed even if actual observed quantities from experiments will vary but in entirely predictable ways (e.g. the boiling point of water is 100 C at only specified conditions but how that varies is predictable). Pritchett (2012a) explores the implications of the lack of invariance laws for experiments in development.

most development projects are addressing *complex* problems. This is good news, as, thankfully, in many country contexts many simple problems—those susceptible to logistical solutions (e.g. vaccinations, expanding enrollments)--have been solved (Kenny 2011).

Given the nature of the design space and fitness functions typical of development projects and the nature of algorithms facing complex problems, it is clear the standard impact evaluation approach is only one part of the learning strategy, for three reasons.

First, the use of RIE, and in particular RCTs, is *intrinsically* very expensive because the data required for impact evaluation on outcomes are *incremental* to the monitoring data as it has to collect data external to the implementing agency and (at least temporally) to the project itself (illustrated in Figure 4). This means the cost per evaluation in the search algorithm is very high, which is the exact opposite of what is needed. As we argue below learning that uses already available data that is part of routine data collection in "M" has much lower incremental cost.

Figure 4: Information requirements for various types of learning and their incremental costs and timing

Project: "M" Data
Internal to Implementation Agency

Time		Inputs	Activities	Outputs
Y 1	Q1	---	---	---
	Q2	---	---	---
	Q3	---	---	---
	Q4	---	---	---
Y 2	Q1	---	---	---
	Q2	---	---	---
	Q3	---	---	---
	Q4	---	---	---
Y 3	Q1	---	---	---
	Q2	---	---	---
	Q3	---	---	---
	Q4	---	---	---

Project: "E" Data
External

	Outcomes	
Time	Intended Beneficiary	Counter Factual
Before	---	---
Mid-term		
	---	---
After		
	---	---

External

Other Vital Statistics
At externally determined frequencies

Second, RIE (including RCT) on *outcomes* is slow—because usually the causal model of mapping from outputs to outcomes (e.g. micro-finance to sustained higher incomes, new innovations to adoption, education to wages) is slow, taking from years to decades. Figure 4 illustrates this as well, while in even moderately well implemented projects data on inputs, activities and outputs is available at very high frequency (at least quarterly) the data on counter-factual on outcomes is available once every few years (or at most once a year). This again implies very few evaluations of the fitness function at different points in the design space fitness are possible. Lag times between intervention and effect can be

long and variable, making the proper time for a follow-up survey hard to predict, at the cost of much wasted effort and money (Behrman and King, WBRO 2009).

Third, the estimate of the LATE produces an estimate of the *average* impact, which averages over all interactions between characteristics of potential users and the project itself. By expunging (through randomization for instance) the effects of all the known or unknown X's in order to better identify "β" (the LATE) it precludes learning about the characteristics of the fitness landscape other than those explicitly included as variants in the evaluated project[24]. In medicine for instance this sets up a direct conflict between researchers and clinicians (analogous to development practitioners) where the former will get some population average while the latter needs to know what will happen to their individual patient with specific characteristics.[25] Moreover, the *marginal* effect of expanding (or scaling up) a program, is the concept relevant to economic appraisal which may, or may not, be well approximated by the LATE.

3.2 The varieties of project failure and learning about project design

An evaluation of the *impact* of a development project on *outcomes* for intended beneficiaries embeds two completely distinct causal models—one of the mapping from inputs to outputs which is internal to the implementing agency and one of the mapping from outputs to outcomes. Many, many projects fail in the first stage—that is the project design fails to produce the intended outputs. In this case, having mounted the cost of baseline data on *outcomes* for the treatment and control areas is (more or less) completely wasted as there is no intervention to monitor.

For example, prospective impact evaluations of the Total Sanitation Campaign in India were hampered by lack of any reliable evidence that the activities of the program even took place or were done particularly well[26]. Subsequent data collected by routine monitoring of social indicators (not part of the project except for tallying some basic indicators like toilets built under the program) were helpful in assessing impact where a formal evaluation was not.[27]

Another example involved social mobilization to support local governments in Karnataka, India.[28] While protocols of the project were carefully specified, it was clear that implementation quality and timing varied relative to baseline and endline surveys. While this is certainly a fault of the evaluation itself, such delays, accelerations and variations in implementation are the actual world of development projects. What is more

[24] This is a methodological point about the trade-off between approaches that estimate impacts from non-experimental data by *conditioning out* other variables (which requires a specification of a more complete model of the underlying phenomena) versus the "rigorous evaluation" approaches that estimate impacts by *balancing* the other variables to avoid the need to correctly condition them out to achieve statistically asymptotically consistent estimates. This implications are explicated in Pritchett 2012a.
[25] Social Science and Medicine among many others.
[26] World Bank, Evaluation of the Total Sanitation Campaign in Maharashtra, 2005.
[27] Spears, 2012.
[28] World Bank, Project Completion Report, Karnataka Local Government Support Project, 2010.

relevant is that had indicators of the implementation schedule and inputs been more carefully kept as part of the monitoring of the project, a better assessment of both the goals of the project as well as its implementation could have been obtained. The missing variable of project implementation quality could be included in the evaluation, a variable that is not amenable to *ex ante* randomization but is crucial to project outcomes nonetheless.

A recent evaluation of approaches to improving the quality of policing in Rajasthan India suffered a similar fate (Banerjee, et al 2012). The "interventions" that could be implemented in a top-down fashion were actually carried out and some of them showed some impact on outcomes. However, several of the planned interventions just did not happen (in spite of the fact that administrative data often reported that they did happen) and hence the "with and without" project comparison of outcomes found no difference. This wasn't a test of the impact of the interventions on beneficiary outcomes were the intervention to be carried out, it was an unplanned test of whether or not the organization could—or would—carry out the intervention.

That said, rigorous evaluation can be done of how different designs on implementation feed into project success at least in producing outputs (if not outcomes). For instance, Olken (2007) evaluates different means of accountability with a community project that built roads (as one option) and found that formal audits did more to reduce corruption than did social accountability. This is not an "impact" evaluation because it simply assumes that the roads, if properly constructed, would produce desirable outcomes. But this type of learning is valuable in its own right.

3.3 How organizations and systems learn

There is an intellectual puzzle about why RIE in general and RCTs in particular have not been more widely used in development projects. After all, the latest wave of enthusiasm for RCTs is not the result of a methodological or technological advance as they follow roughly the techniques and practices used in policy experiments since the 1970s in the USA and elsewhere (e.g. the Matlab experiment in Bangladesh or radio based mathematics instruction in Nicargua). The reason why there have not been more RCTs before, and what continues to limit the expansion of RCTs now, seems to be that governments, development funders, and implementing agencies don't want them. (In fact, many of the new result are from "field experiments" not impact evaluations of on-going development projects.) Why aren't implementing agencies beating a path to the door of this technique?

One narrative (as a proto-positive theory) is that governments, development funders and implementing agencies are self-serving bureaucracies that avoid accountability and that only by increased pressure for accountability will these actors do more RCTs and become more effective. Of course an equally plausible narrative is that the advocates of RCTs are self-serving academics who want more money to do what they want to do, which is write and publish academic papers to promote their fame and glory.

All agree that RIE as a learning strategy is not embedded in a validated positive theory of policy formulation, program design, or project implementation. Ironically the "theory of change" of RIE as a development project falls prey to its own critique of other projects use of evaluation—that is the arguments for RIE focuses exclusively on "input to outputs"—RIE/RCT is a better way of using evaluation inputs to produce impact evaluations. But how the "knowledge" gained from the RIE/RCT will lead to changed behavior at any level—either by the "authorizing" principal in funding agencies or the implementing agency—has never been articulated, much less validated empirically.

Only now are researchers beginning to examine whether RCTs produce knowledge that is organizationally replicable—and the first findings are not optimistic. Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (2012) for instance attempt to replicate the findings of a positive impact of reducing class size with contract teachers in one region of Kenya (and one context of implementation) from Duflo, Dupas, and Kremer 2007 into a broader program across Kenya. They find that when the intervention was implemented by an NGO the positive findings on student learning were replicated. But when the Ministry of Education implemented *exactly* the same project design there was zero impact on student learning. The outcome was not a function of design alone but depended critically on the implementing agency. As of today, the only rigorous test of the theory that knowledge created by field experiments is useful as a guide to scale policy changes refutes that theory. Ironically, "the evidence" is against evidence based policy making as the estimates of impact do not have applicability to precisely the situation for which applicability is the most important—that a government (or larger organization) can implement the policy to scale and get the same results.

3.3.1 How Organizations Learn

Many implementing agencies (or at least significant proportions of the people in those agencies) want to do what they want to do, and do it well if possible. The difficulty is that the idea of "independent evaluation" often arises when a *principal* (e.g. funding agency) wants to select among alternatives and provide more support over time to "what works." When organizations (correctly) perceive that the role of evaluator is to be an instrument to cut their budget if they are "ineffective" rather than help them be effective, the enthusiasm for evaluation naturally wanes (Pritchett 2002). Therefore implementing agencies often are less than enthusiastic (even subversive) of rigorous impact evaluations of *outcomes.*

However, implementing agencies are often interested in evaluation of what works to produce outputs. The management of an implementing agency has some control over the *outputs* of a development project by managing inputs and activities. Hence for accountability purposes, both internal and external to organizations, there is a powerful logic for focusing on the evaluation of the accomplishment of "output" objectives. If a project intends to build roads, or train teachers, or produce research then tracking whether roads were built, teachers attended training, or papers written has a compelling logic. Perhaps the construction of the road will not have its intended *outcome* effect of reducing transport costs for goods, perhaps trucking is monopolized and reduced transport costs

translate entirely into higher profits for truckers and not lower costs for consumers. No one can (or should) hold the *manager* responsible for road construction accountable for that lack of the intended *outcome* due to the faulty model of how road outputs would affect individual outcomes.

A second reason development organizations would allow evaluations which focused on outputs is that *outcome* data is more costly than *output* data because it nearly always involves engagement with actors who are *external* to the development project. Take the example of a project that builds health clinics. The project can easily track whether clinics were constructed and even whether clinics were used, as tracking clinic usage is likely monitoring data internal to the organization. But to know whether *outcomes* (more overall usage of health clinics) improved one has to know whether the increased usage of the clinic was *incremental* or merely displaced the use of other (perhaps equally competent) providers. If the displacement effects are large then even if the project succeeded in *output* terms measured as clinic visits the *outcome* impact on health, or even health care utilization, could be small depending on how much these visits are merely displaced from another provider (Filmer et al 2002). But to know the answer to that question one needs to know about the behavior of the intended beneficiary of the project, who is *external* to the managerial structure of the development project. Indeed, it is not just the intended beneficiary that needs to be understood (and observed) but, for all non-traded goods (like most services), what the nature of the market the beneficiary and the project is part of, also needs to be known. For example the reaction of private suppliers of the same services and the elasticity with respect to either the location or the price of the new facility determines net increase in usage (Hammer 1997). Moreover, one has to collect information from that person that is *additional* to that that would be expected to be collected in the development projects interactions with the project. That is, collecting information is often a routine part of the service delivery process, such as schools keeping track of child attendance, and hence low incremental cost relative to information that is needed for monitoring and management purposes. But to assess *outcomes* one needs to know information like what school, if any, the child attended previous to attending the project school. This often requires tracking information over time that is both costly to collect and not a routine part of the job description of the organization's staff (see figure 4 above). In both the health and education cases, a population based survey is necessary. Information based on the project's own facilities is simply insufficient to determine the full effects of the project – again, the effects that were used to justify the project in the first place.

Moreover, a RIE/RCT generally does nothing to improve the quality and potential impact of "M" on projects or on learning. Since the focus of RIE is on the *counter-factual* of what happened to those who were *not* exposed to the project the data collection for RIE is often completely separate from that of the monitoring data within the project. The emphasis on RIE can even have the tendency to further undervalue "M" as the routine accountability data is seen as even less interesting and relevant.

The key question is: what will change the behavior of the agents in the implementing agency? Often the implicit model behind an RCT is that the management of the

implementing organization will change *design* on the basis of entirely technocratic "evidence" and that implementation will change by edict from above. An alternative is that implementing agents will change their behavior when they are convinced that the new behavior furthers their objectives, which include both self-interest but also some concern for the organization's outputs and outcomes. If this is the case then involvement of the implementing agency and agents in the learning process is essential to the impact of the learning.

3.3.2 How systems learn

A final concern with the RIE/RCT approach is that it doesn't make the distinction between organizational and ecological learning. For instance, evolution does not work because individuals learn, it works when those with superior fitness are more likely to reproduce successfully. Similarly, productivity doesn't just increase in markets because firms become more productive, the average productivity in a market can increase because more productive firms gain a larger market share. So *systems* can learn or improve even if no individual organism or organization improves if entry and exit (or market share) is a function of fitness.

The "top down" model of learning is that there is one, expensive but scientifically definitive, impact evaluation that provides the 'evidence" on which the top managers of an organization change design. This then leads to better results.

The "bottom up" model of learning is that lots of agents/organizations are authorized to conduct their own initiatives and crawl the design space subject to a fitness function that determines survival and expansion. This leads to ecological learning without any necessity for this learning to be codified in a "scientific" way (or published in academic journals). Indeed, valuable improvisations and modifications by the implementers disqualifies the exercise as a publishable paper.[29]

4. Structured Experiential Learning: Introducing little "e" into M&E

The first four sections have set up the need for adding structured experiential learning for implementing and funding agencies to add to the repertoire of tools that include monitoring and rigorous impact evaluation. Our proposal is to explicitly add a new "e" defined as *structured experiential learning*[30]: the process through which an organization learns during the period of project implementation. Development practitioners are well aware that a lot of learning from a project happens *after* the design, but well *before* any formal "evaluation" but as it is this learning is often haphazard and below the radar. The goal is to bring the currently informal processes of experiential learning, from project implementation, explicitly into the overall strategy of development organizations (both implementing and funding). This section now sketches out how MeE could work in practice. As we will see, MeE is the learning component of a larger shift in the way

[29] Unless such improvisation is allowed in evaluating the "intent to treat on treated" but rarely could the exact choices of implementers be modeled.

[30] We would like to thank Ruth Levine for coining the term *experiential learning* to describe "little e".

typical development projects (especially externally funded development projects) operate and is an integral component of the PDIA (Problem Driven Iterative Adaption) (Andrews, Pritchett, Woolcock 2012) approach to building organizational and state capability to implement effective policy.

Essentially, *structured experiential learning* is the process of disaggregating and analyzing data on inputs, activities and outputs chosen to be collected by the project to draw intermediate lessons that can then be fed back into project design during the course of the project cycle. The idea is to take the key insight about using randomization and other rigorous methods to identify impact and expand it dramatically—at lower cost—by using the development project itself as a learning device. Variations in alternatives within the design space *within the project* can be used to identify efficacy differentials in the efficacy of the project on the process of inputs to outputs, which can be measured at low incremental cost at high frequency intervals, for real-time feedback into implementation, at key decision junctures. Rather than thinking of projects as a single element of the design space, projects that are intended to be innovative are authorized strategic evidence-responsive crawls over (part of the) design space.

Let us say that you are Ms. Eager Beaver[31], manager of a development project in the fictitional country of Utopia and you were interested in learning from your project. What would you do? How would you develop a learning strategy that achieves *more than monitoring,* is *cheaper* than impact evaluations, has *timely dynamic feedback loops* built into the project and *extends* the insights gained into the design and management of projects?

In this section we propose a seven step dynamic approach of how "e" can be used to strategically crawl the design space of implementation and help Ms. Beaver learn from her development project.

4.1 The seven steps of MeE

The first two steps are Business As Usual (BAU) approaches used by many development organizations or private foundations and should look very familiar. While steps 1 and 2 are necessary for any development project, they are not sufficient. Therefore, our proposal to Eager Beaver is to augment steps 1 and 2 with 5 additional steps to achieve the desired result of effective development projects.

Step 1: Reverse engineer from goals—framed as solving specific problems--back to project instruments

Eager Beaver is convinced that development projects should be problem driven and not solutions driven. She firmly believes that you cannot *solve* a problem if you cannot *define the problem*. We agree with Eager Beaver as many development projects are designed around solutions looking for problems rather than vice versa.

[31] Ms. Eager Beaver is the companion of Ms. Speedy Analyst (Ravallion 1999).

1(a). Begin with a clear definition of the specific problem you are trying to solve. Then state a *goal* as the *magnitude* of the desired impact. Often the *magnitude* of the goal is left unstated (e.g. "improve quality of education" versus "X percent of children in grade 3 will fluently read grade appropriate material").

When setting the magnitude, it is important to also set an achievable threshold level for your desired impact. Basically, you want to be sure that the magnitude is above the threshold you have set and the impact is achievable with the resources you have available. Setting a reasonable magnitude of impact is important because some development projects have desired impacts that are simply not achievable (e.g. "eliminate corruption"). So even if the project were to be implemented perfectly, the desired impact would not be achieved. This then leads to projects being deemed as failures when the real failure is that of an overly ambitious and unattainable magnitude of the impact (Pritchett 2011)[32]. Therefore setting an achievable goal and magnitude is crucial.

1(b). Reverse engineer from your goal to project instruments. In this step you define the links in your causal chain. Ideally a project should have a clear objective (what problem you are addressing), a clear idea of how these objectives will be achieved (what is your story line/hypothesis/causal chain/theory of change) and clear outcomes (what visible changes in behavior can be expected among end users thus validating the causal chain/theory of change). It is important to emphasize that there are two causal models that neeed to be clearly articultated. The first is the causal model of implementation or the positive behavioral model of implementors that will turn inputs into outputs. The second is positive behavioral model of intended beneficiaries that will turn outputs into outcomes and impacts (see figure 1).

Eager Beaver has identified two key problems that she would like to address in Utopia. The first is related to education. A Utopian NGO recently found that a significant fraction of children in 8th grade could not read or write at acceptable levels. She begins by setting the goal for her project to be – all children can read by Grade 3. She then reverse engineers her goal, using her theory of change and working hypotheses, at each link in the causal chain. Figure 5 illustrates her work.

[32] Clemens and Moss (2005), Easterly (2009) show that the MDGs were *ex ante* designed as too ambitious for Africa and hence it is not at all surprising that *ex post* Africa is "failing" at the MDGs which creates a negative reaction and message about Africa whereas the reality might be that the MDGs failed Africa.

Figure 5: Reverse Engineer Goals to Instruments based on working hypotheses about the causal links of inputs, outputs and outcomes

Figure 5a: Education example of working from goal to instruments

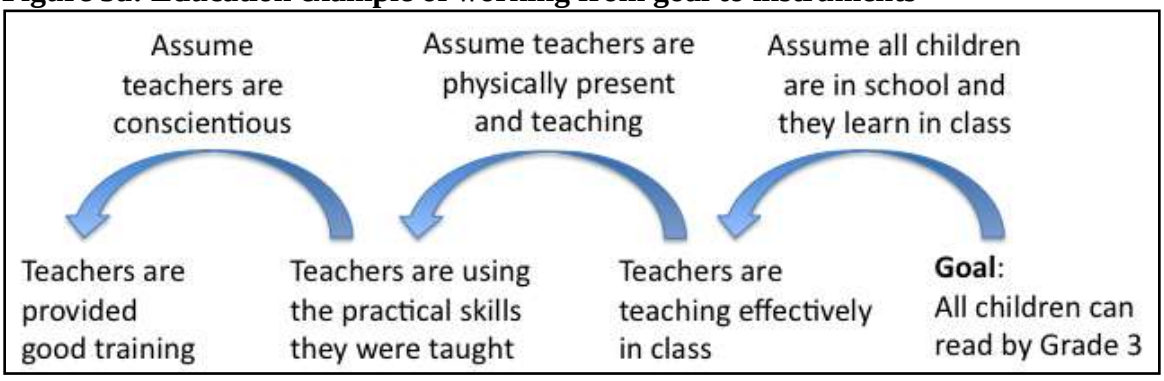
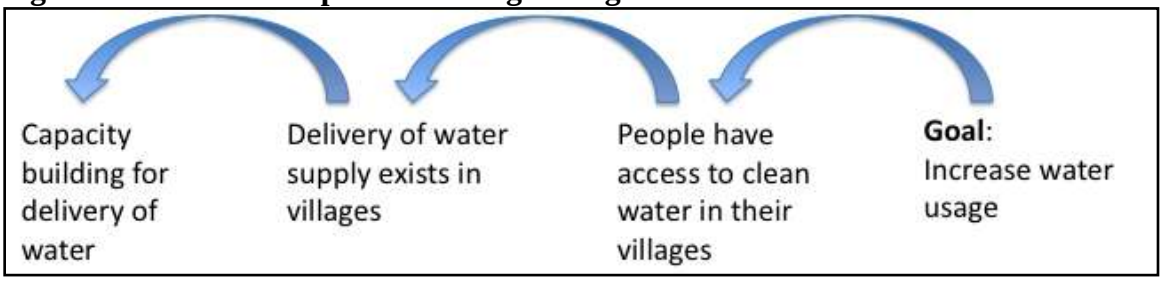


Figure 5b: Water example or working from goal to instruments



The second problem is that much of rural Utopia lacks access to clean water which impacts health, productivity and several other areas. Eager Beaver has worked hard on this problem and decided that the goal of her development project will be to increase clean water usage. Figure 5b shows a simple causal chain.

Step 2: Design a Project

Based on the analytics of Step 1, Step 2 is to design a project that will help achieve the goals. A concrete project design is creating an *instance* of the class of possible projects to achieve your goals—a teacher training project, a community empowerment project, a conditional cash transfer project, a curriculum reform project—are all possible outcomes of a project design with a goal of increasing student learning. Of course a large and comprehensive project will have a number of sub-projects and our use of the word "project" often fits "project component" in large projects.

As part of the project design, specify the timing, magnitude and gain from the project for each link in the causal chain. A development project is a set of decisions about inputs, activities, outputs and a specification of why those will lead to the desired outcomes and impacts. This is often referred to as a logical framework, results based framework, a "complete, coherent, causal chain" or a theory of change, and is often required from implementing agencies for project approval either within an organization, or by an

outside funder[33]. This is true for governments, large multilateral organizations like the UNDP, World Bank, DFID[34], private foundations, as well as internally within NGOs.

One of the biggest, but underemphasized, gains from the increase in RCTs has been at this *ex ante* project design stage. In order to design an experiment to test a project, one has to articulate the project's outcomes and how they would be measured in a more precise way than was often required in traditional logframes. Moreover, to determine the sample size for the statistical power calculations for a prospective RCT design, the *magnitudes* of the expected gains have to be specified. So "improved learning" is not a goal that can be subjected to an RCT evaluated, but "we expect the project to raise the score on this particular instrument that assesses competence in reading (or mathematics or science) by 20 percent" is. Moreover, to specify the outcome gains one has to specify the magnitude of the expected output gains and the link—if a teacher training project is to increase learning then how much the teacher training will augment teacher performance has to be specified to get to the learning gain.

Ms. Eager Beaver designs her two development projects and creates the following diagrams with indicators. She has now completed steps 1 and 2 and is ready to submit her organization's project for funding approval.

[33] http://www.theoryofchange.org/background/basics.html

[34] DFID's Logical Framework from *DFID's* "*Guidelines on Humanitarian Assistance*", *May 1997.* http://webarchive.nationalarchives.gov.uk/+/http://www.dfid.gov.uk/FAQS/files/faq11.htm

Figure 6: Framework and Measurable Indicators of inputs, activities, outputs, outcomes, and impacts

Figure 6a: Education Example

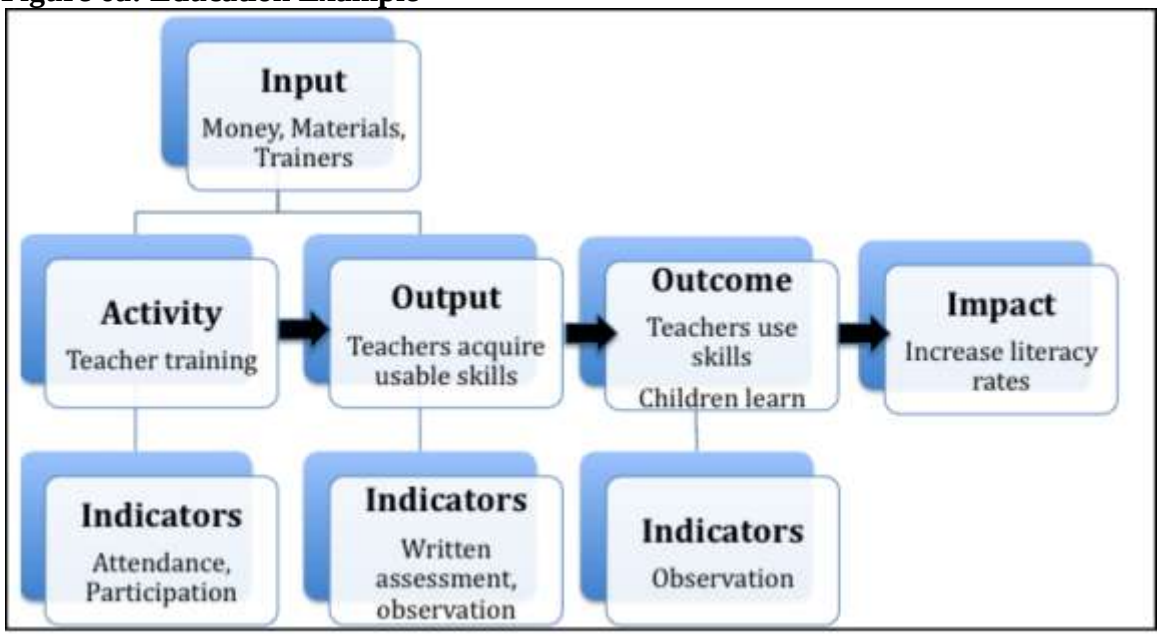
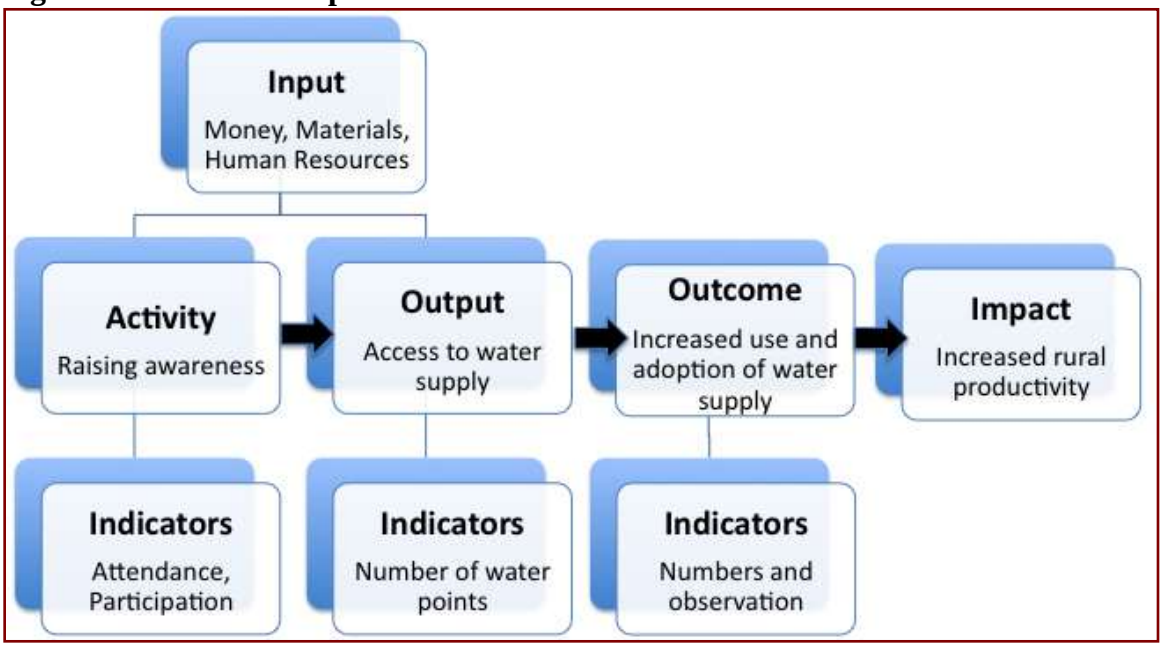


Figure 6b: Water Example



Step 3: Admit we do not know exactly which project design will work and design a crawl of the design space to be authorized as a project

This is the hardest step. The reality is that with complex endeavors—projects in high dimensional design spaces over rugged and contextual fitness functions—no one can know what will work in advance. Development project managers do not know if the inputs will lead to useful outputs (internal area within their control) or if the outputs created will in turn lead to outcomes and impacts (not within their control). As we have argued above, given the level of granularity at which projects have to be designed one cannot be "evidence based"—even if one draws on all of the available information (both from RCTs, RIEs, and otherwise). Development projects are not like chemistry—which is complicated but not complex--where we can predict exactly how interactions will work under specified conditions because we have empirically validated invariance laws that cover all the relevant contingencies.

However there are biases at the individual, organizational, and systemic levels that lead to claims of excessive certainty about what will work. Even though development project managers recognize and accept that they cannot know ex ante about exact project design, the organizational and systematic context in which they work does not allow them to admit that. Often the approval process demands specificity about project design and expected outputs and outcomes from that design at a level of granularity that far exceeds the available, context relevant, evidence.

Our proposed solution to the contradiction between the funding organizations need for specificity of project design as inputs, activities, outputs and outcomes and the intrinsic uncertainty about efficacy facing implementing organizations is to create classes of projects.

Some projects really are just logistics, the solutions have been tried out and proven in context (both overall and organizational), and hence the purpose of the project is just scaling. This is wonderful and this will be part of every funding organizations portfolio. However, not all projects are just the logistics of implementing known solutions and hence processes that insist that all projects present themselves either as logistics or as small scale pilots or field experiments create unnecessary fictions and confusions.

Some projects have to be authorized as a structured crawl over the promising parts of the design space (see step 5 below). This approach to projects balances giving project implementers the flexibility they need to find out what works in organizational and country/regional context with the accountability over use of resources that "authorizers" need to justify decisions.

What does optimization theory say about addressing complex problems? In computational theory there are NP-complete problems which are NP problems (problems solvable in finite (polynomial) time) but also NP-hard (so there are no known, general, quick algorithms). In addressing these hard problems programmers have adapted the principles of evolution—random mutation plus differential replication based on success

in a fitness function to a range of problems. Steve Jones, an evolutionary biologist, helped Unilever create a better nozzle for soap production. He basically made 10 copies of the nozzle with slight distortions at random and tested them all. He then took the most improved nozzle and made another 10 slightly different copies and repeated the process. After 45 such iterations, they had a nozzle with a complex and unexpected shape that worked significantly better than the original.[35]

Drawing on the intuitions from this enormous literature on computational complexity and algorithms (e.g. Cormen, Leiserson, Riveset, Stein 2009)—and on the analogous work on how organizations cope with NK[36] or NP-complete problems in practice—where it is known there are no general solutions available there are several principles to the type of heuristic algorithms used for these problems.

One: try and get into a desirable part of the fitness landscape with good guesses.

Two: rapid iterations are essential to crawl the space (and hence low cost, rapid evaluations of the fitness or objective function are preferred to high cost and slow evaluations).

Three: avoid too early lock-in to a single region of the design space. Standard algorithms that produce local optimum often cannot crawl sequentially to other regions of the fitness landscape.

We propose *structured experiential learning* as the development project counter-part of this. First, the *ex ante* design process is to try and get into a favorable part of the design space. Second, using variations within a project to identify differentials in the efficacy of the project on inputs and outputs for real time feedback into project implementation lowers evaluation cost and feedback loop time. Third, locking yourself into one project limits the potential learning from both the upside and the downside.

Let us give one of potentially thousand examples of the need for mid-course corrections. In the Poverty Alleviation Fund (PAF), a project giving untied funds to groups in Nepal, many villagers chose to purchase and raise goats. As time went on they realized that the goats were getting sick or dying thus defeating the objective of the project. First, this led to a variation on the data that was systematically collected – something that was not initially part of the data, no one being clairvoyant. Armed with data, administrators of the program (which was being evaluated) could successfully put pressure on the Agriculture ministry to increase veterinary services to help address this. Without the feedback of data and the flexibility of reallocating project resources into another dimension of the design space (providing complementary services) the whole project would have failed.

Step 4: Identify the key dimensions of the design space

[35] Jones (1999)

[36] A particular simple type of NP-complete is the NK problem class as introduced by Stuart Kaufmann (1989) that has a fitness landscape with "tunable ruggedness" which has been influential on our thinking (and language).

After admitting you don't know exactly which project will work, it is time to articulate the key dimensions/elements of your design space with multiple alternative options for each.

Let us say that Eager Beaver found out that the real constraint for children's learning in Utopia was teacher training. She begins to think about what the design space would look like: (i) Where should the teacher training take place? (ii) What content should you use? (iii) What will the duration of the training be? (iv) What follow-up activity will you have? It is important to note that with each design parameter you add, you complicate the dimensionality of the design space.

Eager Beaver then narrows it down to three key design parameters for teacher training, with two options each (there can be multiple options):

1. Location: Centrally (A) or in School (B),
2. Content of teacher training: Subject matter (α) or Pedagogy (β), and
3. Follow-up: Semi-annually (I) or Annually (II).

Her design space would then be the total of all possible combinations of her design parameters and would look like table 6. Let the project P1, selected in step 2 be D1(A, α, I).

Table 6: A simplified and illustrative design space for a teacher training (sub)project

Design Parameters	***Design Space***							
	D1=P1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**
Location (A,B)	A Central	A Central	A Central	A Central	B School	B School	B School	B School
Content (α,β)	α Subject	α Subject	β Ped.	β Ped.	α Subject	α Subject	β Ped.	β Ped.
Follow-up (I, II)	I Semi	II Annual	I Semi	II Annual	I Semi	II Annual	I Semi	II Annual

The specification of the design space can be one of the most valuable parts of the project design exercise. It can provide a way in which the various stakeholders in the project are able to articulate their views in ways that at least potentially can be heard. Again, an unacknowledged but important gain in the expansion of the RCTs is that researchers and academics (and others) can be brought into the project design process so that new ideas can be floated and discussed before they are locked in.

Step 5: Select alternate project designs

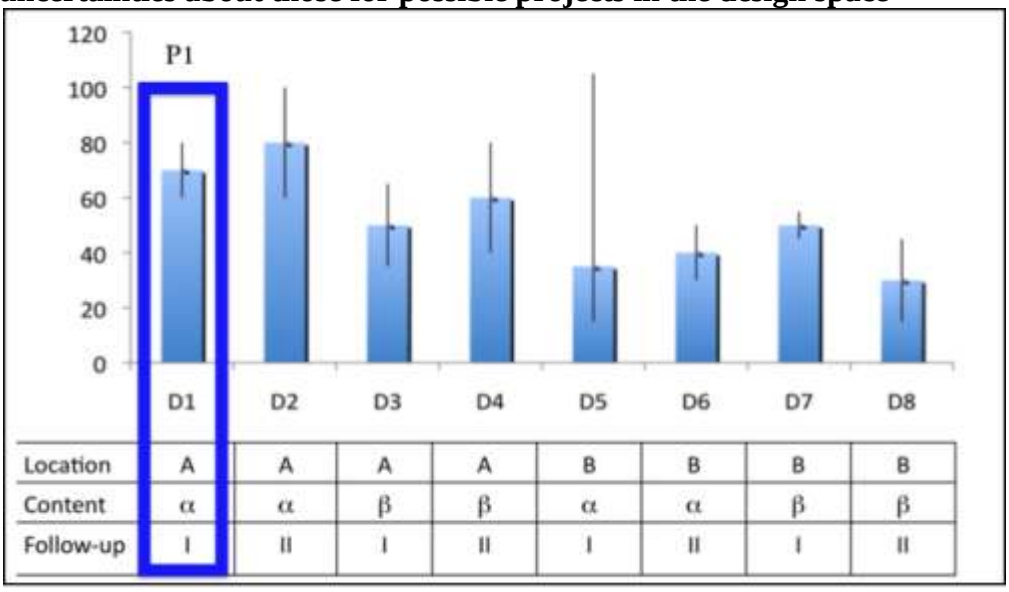
Specify the timing, magnitude and potential gain for each of these possible project variants D2 through D8. Create new project designs by varying design parameters with a clear view to the different hypotheses and theories of change for each project design. For instance, choosing between subject matter or pedagogical content of the teacher training

is based on a view of which constitutes a bigger constraint to effective teaching combined with a view about which would be more responsive to training.

Identify the biggest uncertainty within each link in the chain. Where in the causal chain are you more uncertain that you will get the desired outputs? Where is the highest variance?

Eager Beaver does all the calculations and makes informed guesses about the magnitudes of the project variants, where P1 = D1 (see figure 7).

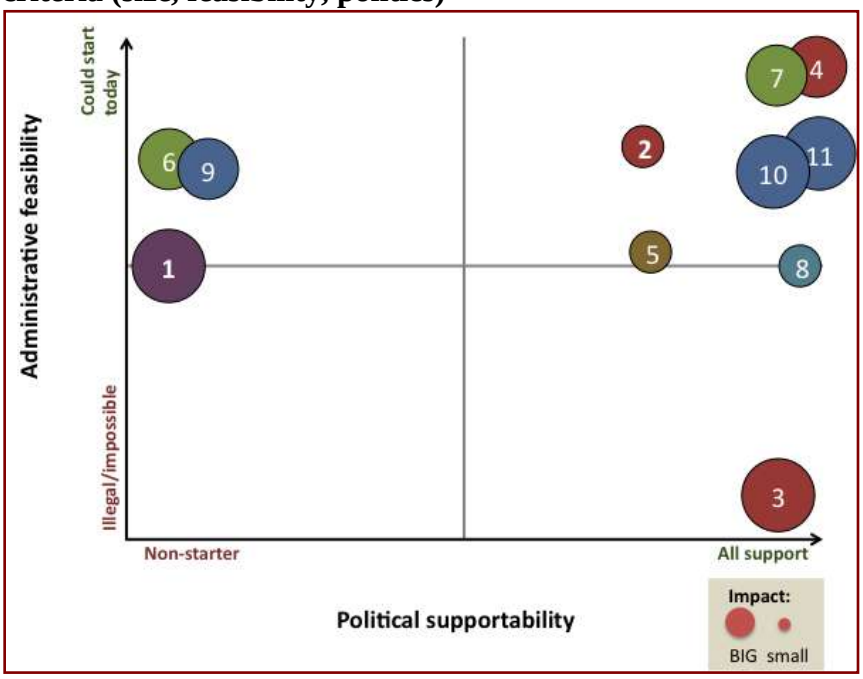
Figure 7: Guess as the magnitudes of gain in performance metric—and the uncertainties about those for possible projects in the design space



Deciding which are the most attractive project design variants is going to be complex and, not surprisingly, there are no simple rules for this.

One approach is to evaluate possible project designs by their likely size of impact, feasibility, and political support so that easier and more attractive alternatives are examined for efficacy first while saving harder and more intractable problems for later when more confidence has been built. Fraker and Shah (2011) use this approach to search the design space and filter out the best project design options to test. Figure 8 highlights that some projects are attractive (can be done, have support, conjectured impact is large) while other projects with equal potential magnitude of impact are "non-starters."

Figure 8: Evaluating potential project design alternatives for sequencing on multiple criteria (size, feasibility, politics)



Source: Fraker and Shah (2011)

Another approach is to consider uncertainty and particularly upside potential. In figure 7 for instance, while D5 has a lower expected return there is massive uncertainty and hence if it works it could be very high performing. Therefore one might want to include that as a project to be tried before say, D3, which has higher expected gain but lower upside.

The goal is to select the project designs that are worth exploring based on some criteria of the attractiveness of testing the design out—which could be low political cost, could be administrative ease, could be upside potential.

Step 6: Strategically crawl your design space: Pre-specify how implementation and learning will be synchronized

As discussed earlier, all development projects collect monitoring "M" data for fiduciary responsibility and for organizational accountability. This data is often stored in text documents or in report formats required by the donor and hardly ever analyzed, often because those engaged in project implementation do not see the value of this data. The process of determining what data should be collected and why, often remains donor centered with very little participation (if any) from the implementers despite their deep understanding of the reality on the ground. So for many implementers, monitoring is just another item to check off their long list of activities rather than being seen as value added to them. In addition, to the implementers, the findings of impact evaluations often come too late – after the project has closed or ended. So neither "M" nor "E" are necessarily

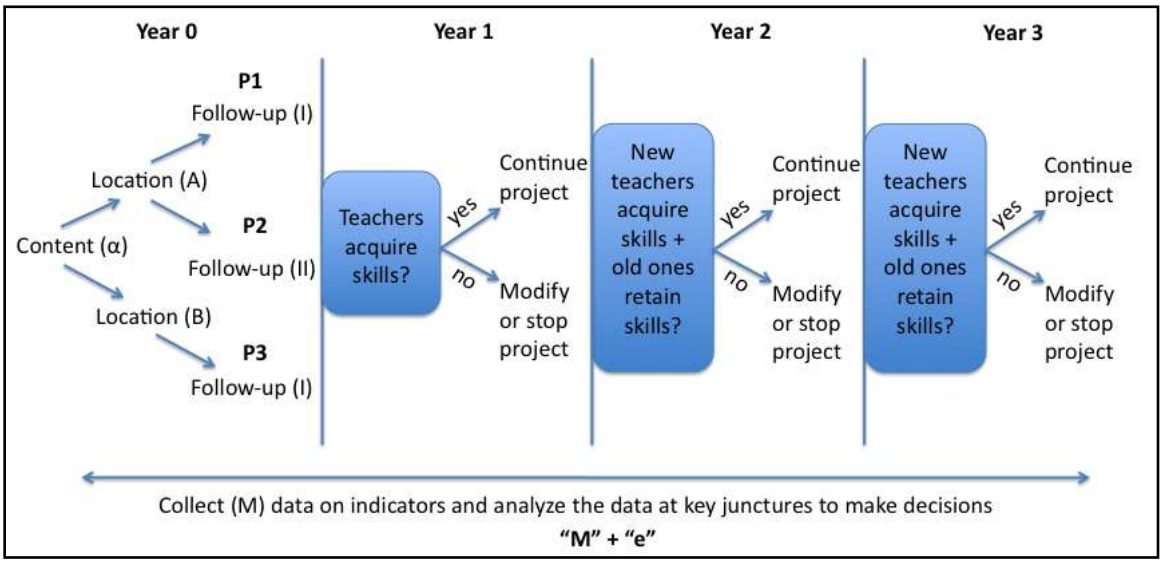
perceived as a useful exercises or uses of resources for project managers and implementers[37].

What development project managers, like Ms. Eager Beaver, need is a management tool to help them make decisions on what resources or inputs to shift; which interventions to implement; and ultimately, identify the priority questions for evaluation. They need a mechanism that helps with planning, provides an opportunity to institutionalize learning and creates a legitimate space for failure.

Introducing *structured xperiential learning* in our approach, builds a dynamic feedback loop into the project where decisions can be made at each step. This could be multi year and/or multi phase. We refer to this as a sequential crawl over the design space.

Ms. Eager Beaver tries to draw the sequential crawl for her teacher training project.

Figure 9: Sequential crawl for teacher training project in which variants are tried out and at project decision points alternatives are scaled, dropped, or added based on results



[37] There is in all policy experiments a trade-off between the "integrity of the experiment" and flexibility. If project implementers change any design element during the course of implementation then it is hard to specify what exactly was evaluated, the project design or some combination of project design and responses to ongoing issues that surface in implementation. While one of us was visiting an NGO project doing an impact evaluation the head of the NGO introduced me to the on-site representative of the evaluation as "This is the guy that makes sure we don't help any children with what we know" as six months into the experiment it was obvious that certain components were not working but the research organization did not want to respond to that information to protect the experimental results.

Step 7: Implement the approved sequential crawl and learn

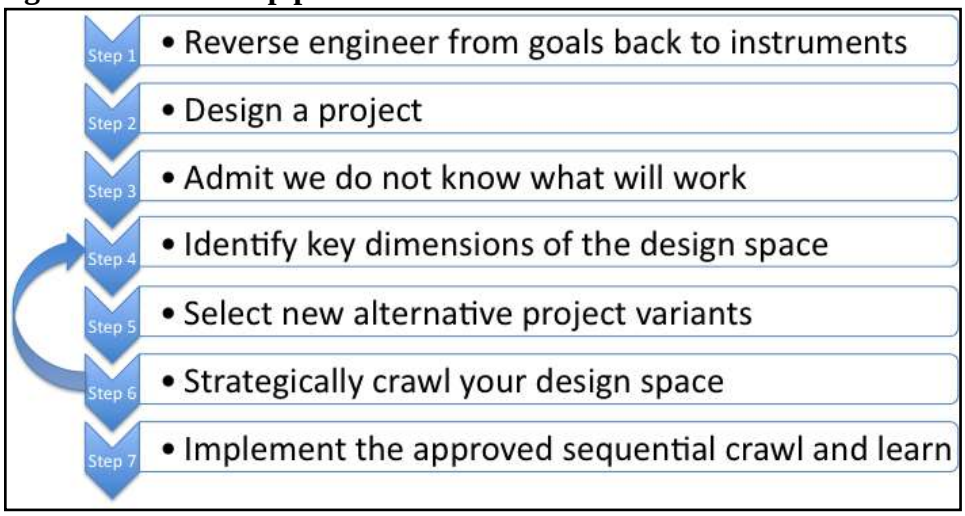
The final step is to implement the project, not as a static design but as a sequential process. For the duration of time specified the initial set of design space variants (D1, D2, D5) are implemented.

The monitoring data is tailored to collect all input and output indicators for all three projects. During implementation the monitoring data (including additional data introduced on the basis of feedback from ground personnel) is analyzed to feed into decision points in the pre-specified sequential crawl. The point is that you keep watching and adjusting the design parameters and inputs as you discover what their impact is on outputs through time. Indeed, the data you collect on outputs can adjust to the realities of the project as well. The advantage of using experiential learning to inform design rather than evaluations is that you do not have to worry about contaminating your sample and you can easily make mid course corrections during project implementation.

After project implementation, an integral part of the project outcomes will not only be the achievements on goals but also the information from the evaluation of various alternatives. At that stage there will be another set of options:

- If there is a variant that appears to be successful in expanding outputs one may want to move to an impact evaluation of a particular project design, or
- If no variant has been successful it will be necessary to either
    - crawl the design space along different dimensions, or
    - shut down the project altogether.

Figure 10: The 7 step process of MeE



5. MeE as an additional tool for implementers and funders

Our point is not that MeE is for every organization in every project. Rather, Ms. Eager Beaver can now decide what makes the most sense for her organization and her development project. With perfectly routine projects "M" alone can be enough, for others, M+e (innovative projects, pilots), or M+E (mature project designs looking to scale) or M+e+E (complex projects with many components). But the view that every development project needs a full scale independent impact evaluation is, at this stage, pure ideology and is not based on evidence of any type.

5.1 Advantages of MeE approaches

There are several advantages to using structured experiential learning ("e") as a complement (not substitute) to "M" and "E".

First, an "e" approach acknowledges and strengthens what already happens informally. Everyone with development experience knows that, just like detailed plans for a battle, the plan evaporates when the first shot is fired. Empirical evidence from over 6,000 World Bank projects shows that *the quality of the task manager*, the extent of project supervision, and early-warning indicators that flag problematic projects, are as important as nearly any other factor in determining project success (Denizer et al., 2011). Implementing a development project, whether in government, in an NGO, or as a funder, requires a great deal of creativity to deal with obstacles and issues that often arise during implementation. Unfortunately processes of project "authorization" explicitly limit flexibility. A MeE approach could potentially balance the needs for both accountability and project flexibility.

Moreover, the acknowledgement of the importance of real time learning from project implementation as part of the organizational strategy, and a legitimization of this as "learning" as opposed to just *ad hoc* temporizing to make a badly designed project work, might help reverse priorities in organizations from *ex ante* to *real time.* Organizations like the World Bank perpetually over-emphasize, over-reward, and over-fund *ex ante* project design over implementation. This is because in the standard model, implementation is just faithful execution of what has already been designed, whereby the thinking is done up front and the implementation is just legwork. However, de facto many successful project designs are discovered when project implementers are given the flexibility to learn, explore and experiment.

Second, the process of articulating the design space and proposing project alternatives with concrete performance objectives makes the *ex ante* project design process more useful. The reality of the project selection process, inside government organizations and between government organizations, tends to be an adversarial process of choosing among projects, which puts project advocates in the position of making much stronger claims for project benefits than can be supported, and being more specific than they would like to be. This is also true of multi-sector funding organizations like the World Bank, in which different types of projects "compete" for their place within the portfolio. In fact, the

section in the World Bank project documents called "alternatives considered and rejected" is often a complete afterthought since the project being proposed is sure to work, so why would any alternative ever have to have been considered?

Third, using *internal* variation in project design (i.e. P1, P2 and P3) to measure effectiveness is enormously more cost effective than impact evaluation if the questions are about mappings of inputs into activities into outputs—which often are the key questions. As illustrated in Figure 4 above, "E" is costly because of the need to create a "non-project" counter-factual, which means collecting *outcome* data on individuals/regions that have *no connection* to the project. Therefore, even if a project has thousands of beneficiaries and keeps track of those individuals on many dimensions as a routine part of project implementation, the statistical power of project effectiveness is determined in part by the size of the counter-factual sample. In addition, if the design space is "rugged" where different designs work better (see figure 2 for an illustration), they can easily be discovered by using "within project" variation at an *incremental* cost over and above the actual cost of routine "M". You would still need to think about statistical power, however, the power per incremental dollar for "e" is much lower than for "E" because if you do "M" properly you should be tracking your inputs and outputs regularly.

Fourth, experiential learning is in the interests of both the implementing organization and the external funders. *Experiential learning* is about doing what the implementing organization *wants to do*, better, while independent impact evaluations are double edged swords. One of the most common issues experienced by those responsible for implementing impact evaluations is the disinterest, if not outright hostility, of the project implementation management to the evaluation team. It is worth noting that much of the impetus for RCTs has been shunted into "field experiments" not "impact evaluations" as there is more enthusiasm for RCTs among academics and their funders than among people who make and implement policy who, for the most part, have yet to be convinced impact evaluations are worth the time, effort, expense, and risk.

Fifth, more emphasis on experiential learning can improve and strengthen *monitoring.* One reason why "M" data is often ignored is that it doesn't provide timely answers to management decisions that project implementers need to make. In fact, a vicious cycle could be induced whereby project implementers find the "M" data, less and less useful or relevant. Unfortunately "E" can also undermine, not strengthen, "M". Again, since the value added of "E" is the *counter*-factual which they need to collect from non-project places, the instruments used are not the same as those used to collect the "M" data. This means that there is complete separation of the "M" data and the "E" data, which means that "M" is even less relevant than it was before.

5.2 Organizational mixes of M, e, and E

An organizational *learning strategy* consists of a project specific mix of MeE:

- *Monitoring* provides needed fiduciary and organizational accountability as well as real time information for active management.
- *experiential learning* creates dynamic feedback loops at key decision junctures, that allow adjustments of development projects to be made to the original program plan, in order to find the one with the highest impact. This middle path is a way to bring the informal process of experiential learning, from project implementation, explicitly into the overall strategy of development organizations.
- *Rigorous Impact Evaluation* provides the most rigorous estimates, of the causal impact of projects on outcomes possible, given the nature of the project.

The optimal MeE strategy will depend on the type of organization and what your objectives are. What do you need to learn? What is your fiduciary reporting?
The problem is that organizations lack a differentiated MeE strategy. Furthermore, as stated earlier, fiduciary reporting is in direct conflict with the idea of learning as an organization and often there is no tolerance for failure.

The learning and evaluation problem is most difficult where funds are allocated across various sectors—which is true of every government—but also true of large development organizations and of large foundations. In this case organizations are often coalitions of advocates for various sectors and/or specific approaches. The single sector implementers/advocates want to discover the most effective projects at accomplishing their desired objectives, but will resist "external" evaluations designed to threaten funding support. The executive components (e.g. Planning Ministries) want a basis to compare effectiveness across sectors, but also want to create a space in which the sectors can search for the most effective projects within their sectors. This creates conflicting objectives within the organization and can often stymie evaluation, which requires the cooperation of both the expertise of evaluation but also the interest and cooperation of sector experts and project implementers who often feel that impact evaluation is a hostile endeavor. MeE is an attempt to reconcile these so that there is an organizationally realistic approach to learning that has the enthusiastic cooperation of sectors and implementers in an "evidence based" approach for searching for what is most effective.

We suggest a portfolio approach where you match resources to what you need to learn. So when planning your portfolio of projects, you want to use a lot of "M" (and analyzing the data you collect) on routine type projects; "e" to help crawl the design space for innovation projects, or ones that have large uncertainties; and "E" for flagship projects which are large, scalable with the potential to affect the system, and novel.

We explore four different types of development organizations to illustrate how their learning portfolio will vary in their optimal MeE strategy by project type. The numbers used in the tables are best guesses and are mainly for illustrative purposes.

Country Governments. Governments will have a large share of their learning strategy in routine type projects. The value of routine collection of high quality data on outcomes at a level of disaggregation that would allow for sensible comparisons in real time cannot be overestimated. Since people associated with very specific projects will have an incentive to economize on data collected in "non-program" areas and on data concerning variables they are not immediately interested in examining, such as the level of education, income, road density, etc. In a province, district or village (or whatever is the sensible unit of analysis), either within or especially outside the project area, it falls to government to collect such data.

In this sense, the government (or their statistical agencies) builds a rolling baseline over time that can be used to compare jurisdictions on their progress toward ultimate goals. It also generates data that can be used to tell where an intervention is likely to succeed and, ultimately, contribute to a model of why it is likely to succeed. This will be useful for the huge number of policies, more so than for discrete projects, that are simply impossible to evaluate with RIE's/RCT's.[38]

Large aid organizations. The World Bank operates in 6 regions and in 2006 had 1,282 new projects under preparation; 2,372 projects under supervision; and $20-25 billion in new lending as well as $9 billion in trust funds. All World Bank projects have funds for "M&E" and all collect "M" data. It is simply not feasible, desirable or cost effective to conduct an impact evaluation of each and every one of these projects. Since the World Bank has decades of experience with projects, there are several categories of projects that could be combined in the category of "routine" projects. These could include infrastructure projects like building roads, schools etc. On routine projects, "M" could be sufficient, provided you collect the relevant data and you analyze it and use it to make decisions. Then, there are "flagship" learning which were mature project designs going to scale. On these impact evaluation ("E") is the key to determine a rigorous estimates of the causal impact on outcomes and hence test the causal models there.

Finally, within assistance organizations there are a much smaller number of projects that are new and innovative. For these projects "e" would be a very helpful tool to help crawl the design space to find the project with the highest impact. Innovation funds are on the rise. In fact, USAID recently launched a promising program called Development Innovation Ventures (DIV) to invest in projects that will produce development outcomes more effectively and cost-efficiently while managing risk and obtaining leverage[39].

[38] For example: in the U.S. the continuing debate over whether gun control or sentencing laws increase or reduce murder rates is of critical importance. However, it will never be resolved by subjecting it to analysis by experimental methods. It can only be analyzed with observational methods even though it may never be finally and conclusively determined by them, but must be discussed in political debate and absolutely requires regular data on murders, incarceration rates, income, unemployment, etc. at state or smaller jurisdictions. This data generates if not disposes of hypotheses of much wider import than discrete projects.

[39] See www.usaid.gov/DIV

The optimal MeE strategy for a large aid organization could therefore have a majority proportion on doing "M" but at the same time designated projects could be "e" or "E" focused.

Large private foundations. Large private foundations like the Bill and Melinda Gates foundation, or the William and Flora Hewlett foundation, operate in several countries and give over $1.5 billion, and over $400 million annually. Unlike organizations like the World Bank or USAID, these private foundations are not accountable to country governments and are therefore able to take more risks and be more innovative. Bill Gates recently announced that he would be investing $41.5 million to reinvent the toilet—clearly an endeavor that requires out of the box thinking. Foundations can plausibly have a higher tolerance for risk and may want to focus its learning strategy on innovation type of development projects where "e" is then the largest share of their portfolio.

Single sector implementation organizations. There are a variety of "single sector" or even "single project type" organizations. These are typically worried about both attracting more resources into their sector/activity (e.g. girl's education, micro-credit, family planning, maternal mortality) and about effectiveness of the use of those resources (e.g. what works best to keep girls in school). Therefore these organizations are typically more interested in "e"—the experiential learning from "crawling the design space" to find the most effective project type in the given context. Only as they feel they have moved towards an effective intervention will they need to move towards RIE as a means of creating evidence to scale their projects.

Table 7 summarizes the allocations of the various types of organizations. As stated earlier we are not proposing that each individual project activity adopt an MeE approach but rather the organization as a whole have an explicit learning strategy that involves a mix, depending on their risk tolerance, the extent to which learning is an objective, and their capacity to authorize and support innovation. These same type of portfolio allocations can happen within sector or sub-sectors or ministries. We are not recommending a Ministry of Education or Ministry of Health or infrastructure agency abandon the core functions and become innovators. But, at the same time, some part of the organization should be devoted to innovation that is rigorous and evidence based to address the new and recurring challenges they are facing.

Table 7: Choices amongst M, e, and E across types of organizations

Type of Projects and Learning Strategy Category	**Governments** *(Risk averse, hard accountability constraints)*	**Large external development agency** *(Knowledge generation an explicit organizational objective)*	**Large foundation** *(Can promote risky endeavors, knowledge more than scaling an objective)*	**Individual Implementing organization** *(Risk averse about funding support, limited breadth of mission)*
	Percent of portfolio by numbers of projects (not grant/lending volume)			
Routine: "M" *(Projects based on firm evidence at the logistical stage of implementation)*	80%	70%	50%	40%
Innovations: M + e *(Projects (or sub-projects) where learning by exploring new approaches is itself a primary outcome of the project)*	10%	20%	30%	50%
Flagship Learner: M + E *(Projects testing mature project designs ready to go to scale and replicate)*	10%	10%	20%	10%

Conclusion

We feel development desperately needs a "science of implementation" (Kim 2012). But everyone engaged in development needs to acknowledge that the practice of development will be a "science" in that way that medicine is "science"—a set of accepted practices in a community of doers that are based as best as possible on a evidentiary foundation that draws on a range of scientific disciplines—not in the way that academic chemistry is a science. "M&E" as currently practiced is insufficient as a learning tool about complex development projects.

Our approach of MeE is just one way of describing the ideas similar to many other proposals and we are not claiming exclusivity but rather are emphasizing the commonality. Blattman (2008), for instance, makes the case for "Evaluation 2.0" which takes into account context specificity and the need for evaluation to focus on "performance management and process learning." Pawson et al. (1997, 2004) argue for

"Realist Evaluation that asks not, 'What works?' but instead asks, 'What works, for whom, in what circumstance, in what respects, and how?' Szekely (2011) argues that development is a moving target and therefore more integrated approaches are needed to institutionalize learning[40]. He suggests Results-Based Social Policy Design and Implementation (RSPDI) systems, which could look like the *diagnosis-design-implementation-evaluation-analysis-finetuning-implementation* described by Greenberg (1968). Khagram et al (2009) suggest diagnostic, contextual approaches to experimentation and innovation for development in the twenty-first century – Impact Planning Assessment Reporting and Learning Systems (IPARLS)[41].

MeE is not a panacea or a development strategy but the pragmatic project level learning tactical counter-part of an emerging strategic approach to development that might be called: "guided incremental experimentation" that emphasizes that the development process is a highly complex and contingent process that can be guided by principles, but is not reducible to simple rules or programs[42]. We feel that MeE (or a variant of it)—encouraging the extension of the principles of RCTs inside the project implementation process—can be a valuable component of making development more pro-actively evidence based, especially if embedded in a generally more organic, open, and performance based approach to the hard slog of development.

[40] Szekely (2011) states "evaluators may prioritize academic purity, professional prestige, recognition, knowledge generation, academic success (publications), etc., that may be incompatible with evaluations that are timely, credible, relevant, pertinent, and communicable from the point of the users."

[41] The Impact Evaluation For Development (IE4D) Group's 2011 "Principles for Action" states: "Evaluation, like development, needs to be an open and dynamic enterprise. Some of the current trends in evaluation limit unnecessarily the range of approaches to assessing the impact of development initiatives. We believe that impact evaluation needs to draw from a diverse range of approaches if it is to be useful in a wide range of development contexts, rigorous, feasible, credible, and ethical."

[42] This approach has emerged from a number of sources in different domains of development: "second-best" approaches to reform and institutions or "one economics, many recipes" (Rodrik, 2007), the search for "high-bandwidth" economic policy making (Hausmann, 2008), the "good enough governance" approach in the political and social policy sphere (Grindle, 2005, 2011), the shift from "best practice" to "best fit" in project design (Booth, 2011), the dangers that *the solution* becomes the problem and leads to "capability traps" in administrative capability (Pritchett and Woolcock, 2004; Andrews, Pritchett and Woolcock, 2011; Andrews, 2011; Filmer et al 2000, 2002).

References

Andrews, M., Pritchett, L., and Woolcock, M. (2012), "Escaping Capability Traps through Problem-Driven Iterative Adaptation (PDIA)", UNU-WIDER working paper 2012/64.

Andrews, M., Pritchett, L., and Woolcock, M. (2012b), "Looking Like a State: Techniques of Persistent Failure in State Capability for Implementation", UNU-WIDER working paper 2012/63.

Andrews, M., Pritchett, L., and Woolcock, M. (2010), "Capability Traps? The Mechanisms of Persistent Implementation Failure", Center for Global Development, Working Paper 234.

Angrist, J., Bettinger, E., Bloom, E., King, E. and Kremer, M. (2002), "Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment", NBER Working Paper 8343.

Angrist, J. and I. Fernandez-Val (2010), "Extrapo-LATEing: External Validity and Overidentification in the LATE Framework," NBER Working Paper No. 16566.

Ariely, D., Ayal, S., and Gino, F., (2009), "Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel." Psychological Science.

Ashraf, N., Field, E., and Lee, J., (2010), "Household Bargaining and Excess Fertility: An Experimental Study in Zambia."

Baird, S. M., Craig; Ozler, Berk (2009). "Designing Cost-Effective Cash Transfer Programs to Boost Schooling among Young Women in Sub-Saharan Africa." World Bank Policy Research Working Paper Series(WPS5090).

Barder, O. (2012), "Development and complexity," Presentation and podcast made at CGD. http://www.cgdev.org/doc/CGDPresentations/complexity/player.html

Barrera-Osorio, F., and Filmer, D., (2012), "Incentivizing schooling for learning: Evidence on the impact of alternative targeting approaches".

Banerjee, A., Chattopadhyay, R., Duflo, E., Keniston, D., and Singh, N., (2012), "Can Institutions be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police", NBER Working Paper No. 17912.

Banerjee, A., Duflo, E., and Glennerster, R., (2008), "Putting a Band-Aid on a corpse: incentives for nurses in the Indian public health care system", Journal of the European Economic Association, Volume 6, Issue 2-3, pages 487–500.

Bertrand, M., Mullainathan, S., Karlan, D., Shafir, E., and Zinman, J., (2010). What's Advertising Content Worth?Evidence from a Consumer Credit Marketing Field Experiment The Quarterly Journal of Economics. 125(1):263–305.

Behrman, J., and King, E., (2009), "Timing and Duration of Exposure in Evaluations of Social Programs," World Bank Research Observer, World Bank Group, vol. 24(1), pages 55-82, February.

Bjorkman, M. and Jakob, S. (2007), "Power to the People: Evidence From a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda." *CEPR Discussion Paper No. DP6344*.

Blattman, C. (2008), *Impact Evaluation 2.0*. http://www.chrisblattman.com/documents/policy/2008.ImpactEvaluation2.DFID_talk.pdf

Briscoe John et al. (1993), "The demand for water in rural areas: determinants and policy implications", World Bank Research Observer, Vol. 8, No. 1, pp 47-70.

Cohen, J. D., Pascaline (2010). "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." Quarterly Journal of Econonics **125**(1): 1-45.

Cormen, T., Leiserson, C., Rivers, R., and Stein, C., (2009), *Introduction to Algorithms*.

Das, J., and Hammer, J. (2007), "Money for nothing: The dire straits of medical practice in Delhi, India," Journal of Development Economics, vol. 83(1), pages 1-36.

Devarajan, S., Squire, L., and Suthiwart-Narueput S. (1997), "Beyond Rate of Return: Reorienting Project Appraisal", World Bank Research Observer, 12 (1).

Dasgupta, P., Marglin, S.A., and Sen A.K. (1972), "Guidelines for Project Evaluation", United Nations, New York.

Denizer, C., Kaufmann, D., Kraay, A. (2011), "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance", Background paper for IDA at 50 Report, World Bank.

Duflo, E. D., Pascaline; Kremer, Michael (2007). Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya. http://isites.harvard.edu/fs/docs/icb.topic436657.files/ETP_Kenya_09.14.07.pdf.

Duflo, E., Hanna, R., (2007). "Monitoring Works: Getting Teachers to Come to School," Natural Field Experiments 38, The Field Experiments Website.

Filmer, D., Hammer, J., and Pritchett, L., (2000), "Weak Links in the Chain: A diagnosis of health policy in poor countries", World Bank Research Observer, vol. 15, no. 2, pp.

199–224.

Fraker, A., and Shah, B. (2011), “Learning to Learn: A New Approach to Policymaking in Hyderabad, India”, Harvard University, SYPA, March 2011.

Greenberg, B. G. (1968), “Evaluation of Social Programs”, Rev Int Stat Inst 36: 261.

Grindle, M. (2002) “Good Enough Governance: Poverty Reduction and Reform in Developing Countries”, http://www.gsdrc.org/docs/open/HD32.pdf

Hammer, J. (1997), “Economic Analysis for Health Projects”, World Bank Research Observer, 12 (1).

Hausmann, R.,(2008), “*High Bandwidth* Development Policy”, RWP08-060.

Holla, A. K., Michael (2009). "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." CGD Working Paper Series(158).

Jensen, R. (2010), “The (Perceived) Returns to Education and the Demand for Schooling”, Quarterly Journal of Economics 125 (2): 515-548

Jones, S. (1999), Almost like a whale: the origin of species updated. London: Random House.

Kim, J.Y., (2012), “Delivering on Development: Harnessing Knowledge to Build Prosperity and End Poverty”, World Bank Group President Keynote Speech to World Knowledge Forum, October 9, 2012 - Seoul, Korea.

Khagram, S., Thomas, C., Lucerno, C., Mathes, S. (2009), “Evidence for development effectiveness”, Journal of Development Effectiveness, 1:3, 247-270.

Khagram, S., Rogers, P., Bonbright, D., Earl, S., Carden, F., Ofir, Z., Macpherson, N. (2011), “Principles for Action”, The Impact Evaluation for Development (IE4D) Group.

Little, I. M. D., and Mirrlees, J. A. (1969), “Social Cost Benefit Analysis: Manual of Industrial Project Analysis in Developing Countries”, Vol.II, OECD.

Moore, M. (1995), *Creating Public Value Strategic Management in Government*, Harvard University Press.

Moore, M., and Khagram, S. (2004), “On Creating Public Value: What Business might Learn from Government about Strategic Management”, Corporate Social Responsibility Initiative Working Paper 3, Harvard Kennedy School.

Munnell, A. H. (1987), "Lessons from the income maintenance experiments: an overview," New England Economic Review, Federal Reserve Bank of Boston, issue May, pages 32-45.

Newhouse, J. P. (1996), *Free for All? Lessons from the RAND Health Insurance Experiment*, RAND.

Ostrom, E. (1996), "Crossing the Great Divide: Coproduction, Synergy, and Development." *World Development* 24(6): 1073-87.

Padian, N., McCoy, S., Balkus, J., Wasserheit, J., (2010), "Weighing the gold in the gold standard: challenges in HIV prevention research", AIDS 2010, 24**:**621–635

Pawson, R. and Tilley, N. (1997), Realistic Evaluation. London: Sage Publications.

Pawson, R. and Tilley, N. (2004), "Realist Evaluation", funded by the British Cabinet Office.

Pritchett, L. (2002). "It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation", *Policy Reform*, Vol. 5(4), pp. 251–269

Pritchett, L., and Woolcock, M., (2004), "Solutions when the Solution is the Problem: Arraying the disarray in development", The University of Manchester, e-journal, 32/2: 191-212.

Pritchett, L., with Samji, S. (2009), "Political economy of aid evaluation: How to build a sustainable *and* effective movement", Presentation. Harvard Kennedy School.

Pritchett, L., (2011), "The Financial Crisis and Organizational Capability for Policy Implementation." Chapter 9 in *New Ideas on Development after the Financial Crisis*, edited by Birdsall N., and Fukuyama F.

Ravallion, M. (1999), "The mystery of vanishing benefits: An introduction to impact evaluation", World Bank Economic Review, Vol. 15, No. 1, pp: 155-140.

Ravallion, M. (2011), "On the implications of essential heterogeneity for estimating causal impacts using social experiments", World Bank.

Roberts, J., (2004), "The Modern Firm Organizational Design for Performance and Growth", Oxford University Press.

Rodrik, Dani (2007), *One Economics, Many Recipes: Globalization, Institutions and Economic Growth*, Princeton, NJ: Princeton University Press.
Baird, S. M., Craig; Ozler, Berk (2009). "Designing Cost-Effective Cash Transfer Programs to Boost Schooling among Young Women in Sub-Saharan Africa." World Bank Policy Research Working Paper Series(WPS5090).

Bhide, A. (2000). The Origin and Evolution of New Businesses. NY NY, Oxford University Press.

Cohen, J. D., Pascaline (2010). "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." Quarterly Journal of Econonics **125**(1): 1-45.

Duflo, E. D., Pascaline; Kremer, Michael (2007). Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya. http://isites.harvard.edu/fs/docs/icb.topic436657.files/ETP_Kenya_09.14.07.pdf.

Easterly, W. (2006). The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good. NY NY, Oxford University Press.

Holla, A. K., Michael (2009). "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." CGD Working Paper Series(158).

Kenny, C. (2011). Getting Better: Why Global Development Is Succeeding--And How We Can Improve the World Even More. NY NY, Basic Books.

Mintzberg, H. W., James A.; "Of Strategies, Deliberate and Emergent." Strategic Management Journal **6**(3): 257-272.

Rodrik, D. (2008). "The New Development Economics: We Shall Experiment, but How Shall We Learn?" HKS Faculty Research Working Paper Series(RWP08-055).

Samji, S., and Sur, M. (2006), "Developing a high quality baseline", World Bank.

Savedoff, W., Levine, R. and Birdsall, N. (2006), "When will we ever learn: Improving lives through impact evaluation", Report of the Evaluation Gap Working Group, Center for Global Development, Washington, D.C.

Squire, Lyn, and Herman G. van der Tak (1975), *Economic Analysis of Projects*, World Bank Publications.

Szekely, M. (2011), "Toward Results-Based Social Policy Design and Implementation", Center for Global Development, Working paper 249.

Victora, C. G. (1995), "A Systematic Review of UNICEF-Supported Evaluations and Studies, 1992–1993", Evaluation and Research Working Paper Series 3. United Nations Children's Fund, New York.

White, H. (2006), "Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank", World Bank, Washington, D.C.

Baird, S. M., Craig; Ozler, Berk (2009). "Designing Cost-Effective Cash Transfer Programs to Boost Schooling among Young Women in Sub-Saharan Africa." World Bank Policy Research Working Paper Series(WPS5090).

Bhide, A. (2000). The Origin and Evolution of New Businesses. NY NY, Oxford University Press.

Cohen, J. D., Pascaline (2010). "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." Quarterly Journal of Econonics **125**(1): 1-45.

Duflo, E. D., Pascaline; Kremer, Michael (2007). Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya. http://isites.harvard.edu/fs/docs/icb.topic436657.files/ETP_Kenya_09.14.07.pdf.

Easterly, W. (2006). The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good. NY NY, Oxford University Press.

Holla, A. K., Michael (2009). "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." CGD Working Paper Series(158).

Kenny, C. (2011). Getting Better: Why Global Development Is Succeeding--And How We Can Improve the World Even More. NY NY, Basic Books.

Mintzberg, H. W., James A.; "Of Strategies, Deliberate and Emergent." Strategic Management Journal **6**(3): 257-272.

Rodrik, D. (2008). "The New Development Economics: We Shall Experiment, but How Shall We Learn?" HKS Faculty Research Working Paper Series(RWP08-055).