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

Lant Pritchett
Salimah Samji

and

Jeffrey Hammer

June 22, 2012

Abstract: Organizations that fund development projects—whether they be governments, multi-laterals, bilateral agencies, or NGOs—have to make hard decisions about what to fund. In this decision there is an inherent tension between funding activities that have solid evidence about effectiveness and funding innovative activities that promise even greater effectiveness but are untested. “Evidence based” approaches that promote greater use of Rigorous Impact Evaluations (including randomized control trials) and evidence from those evaluations in policy and programming have added more rigor to the E (evaluation) in traditional M&E. Here we extend the basic idea of rigorous impact evaluation—the use of a valid counter-factual to make judgments about causality—to evaluate project design and implementation. This adds a new learning component of experiential learning or a “little e” to the M&RIE so that instead of just M&E development projects are all about MeE. Structured experiential learning 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 project design and implementation. The key insight is that within-project variations can serve as their own counter-factual which dramatically reduces the incremental cost of evaluation and increases the usefulness of evaluation to implementing agencies. The right combination of MeE provides for rigorous learning while the providing needed space for innovation.
1. Introduction

An effective development project must have an adequate “theory of change”-- a complete, coherent, and correct causal model from funding to inputs and activities to outputs to outcomes and impacts. The design of a development project needs to answer two key “why” questions. First, why will the agents of the implementing organization translate funding into inputs into activities that will create useful outputs? Second, why will the outputs produced by the project increase the well-being of the intended beneficiaries? These “why” questions require a positive behavioral models of how people (implementers and intended beneficiaries) respond to the new structure of incentives and 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 better water supply for the farmers or because water wasn’t a key constraint to farmer output; education projects 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, etc.

The traditional approach to monitoring and evaluation (M&E) in development projects has been under withering attack, on two fronts. First, that traditional M has been too focused on input utilization and process compliance and does not actually contribute useful information to project management. Hence the collection of monitoring data is not a priority for project management and therefore monitoring data is not reliable or timely. Second, that evaluation practice was based on crude “before and after” estimates and lacked a 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)\(^1\), and increased pressure on funding organizations that their activities be based on “rigorous” evidence about “what works.”

Rigorous impact evaluation is a tactic not a strategy. RIE can be a tactic in traditional top-down strategies and their approach to project design, implementation, learning, and scaling up. However, many argue that the traditional top-down approach is itself a fundamentally unsound general strategy to development (whether supplemented by RIE or not)\(^2\) and is not a generally viable strategy for creating capable\(^3\) and accountable\(^4\)

---

1. 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.
2. Pritchett and Woolcock (2004) catalogues the many critiques and proposed alternatives to “the solution” (top-down implementation by a Weberian bureaucracy of a technocratically determined project design) as the solution.
3. Pritchett, Andrews and Woolcock (2012) argue that traditional approaches to project design, implementation, learning and scaling actually facilitate continued failure and perpetuate “capability traps”

Preliminary and Incomplete:  Page 2
For Comments Only June 22, 2012
public sector organizations, particularly for the provision of implementation intensive services or imposition of obligations, and needs to be replaced by an alternative strategy that entirely rethinks the role of development projects in a development theory of change. The criticism of the lack of rigor of impact evaluation was a very conservative critique as, perhaps inadvertently, it reinforced mainstream development approaches by raising the possibility at all that was wrong with traditional strategy of development projects was the way they were evaluated.

Andrews, Pritchett, and Woolcock (2012) for instance outline a “searcher” 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 (the ideal of “planners”) - but rather as instruments for learning 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.

In this paper, we extend the fundamentally sound ideas behind rigorous impact evaluation and RCTs into a “searcher” or PDIA strategy for development with structured experiential learning. We introduce experiential learning or “little e,” which is an addition to M&RIE, that merges rigorous evaluation with practice and the ways implementing organizations and communities of practice actually learn. Our structured experiential learning is a seven step dynamic approach that explicitly builds learning objectives into project design and allows project implementers the flexibility to “crawl the design space” in search of the most effective project designs. This integrates monitoring data into decision loops of project implementation by using within project design variations for a sequential and contingent plan for implementation. Increased relevance of monitoring to the implementers is likely to improve its quality. The use of MeE by implementing agencies allows funders of development to balance the search for innovation with the need for rigorous evidence.

---

4 The World Development Report 2004, Making Services Work for the Poor provided an analytical structure for cataloging the ways in which the provision of services by public sector organizations (the “long route”) failed to produce the needed accountability relationships (see especially chapter 3).

5 For instance, Easterly (2006) raises a fundamental distinction between “planners” and “searchers” in development. In the debates over the use of RIE it is clear that some see the “planners” approach as fundamentally sound if only supplemented by more RIE whereas others argue that “planners with more RIE” might be a modest improvement on “planners without more RIE” but that the adoption of RIE does not change planners into searchers. This paper is about “searchers with RIE.”
2. First Generation M&E\(^6\)

We 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\(^7\). These development projects have the goal of improving the well-being of some target population, the *intended beneficiaries*\(^8\).

Our definition of a development project intends to include everything people consider a development project--and more. Building physical infrastructure or facilities (e.g. roads, schools, ports, health clinics, nuclear power plants) are development projects. Social programs (e.g. conditional cash transfers, micro-lending) are development projects. Policy advocacy is a development project. Empowerment is a development project. It is important to stress that rigorous evaluation of a development project is itself a development project. Evaluation uses funds to finance inputs and *activities* (collection of data, analysis of data) that produce *outputs* (reports, research papers, policy advocacy) by an *implementing agency* 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 an array of examples.

---

\(^6\) 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.
Table 1: Examples of the wide range of development projects

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

\[\text{Figure 1: M&E for a development project}\]

\[\text{Diagram showing the flow of data from funders to implementing agency to intended beneficiaries.}\]

---

\(^9\) 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? 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.

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

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

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

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

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

**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 budget available (Little and Mirrlees 1969, Dreze and Stern, 1969). The main intellectual problem was the valuation of the outcomes of the project. 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, particularly when 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 when market prices for inputs and outputs were distorted and hence “shadow prices” were needed for valuation, when markets were

---

\(^{10}\) 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.
non-existent, when valuing non-marketed outcomes and addressing distributional concerns (Squire and van der Tak 1976).

**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. bilateral donor agencies their own budgeting and accountability mechanisms and ultimately taxpayers). Nearly all funding organizations have implementation evaluation as a required part of the project cycle.

**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 counterfactual—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**

<table>
<thead>
<tr>
<th></th>
<th>Valuation Evaluation：Do the benefits (properly weighted and discounted) exceed the costs?</th>
<th>Implementation Evaluation：Did the implementing agency succeed in doing what they said they would do in processes, activities and outputs?</th>
<th>Impact Evaluation：Did the project lead to the desired outcomes and impacts on the intended beneficiaries?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construction of a road</strong></td>
<td>Does the predicted road volume justify the costs of reducing travel time by constructing the road?</td>
<td>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</td>
<td>Road was constructed but projected traffic volume did not materialize – why? - Economy wide recession - Monopoly of truckers</td>
</tr>
<tr>
<td><strong>Promotion of better practices to improve nutrition</strong></td>
<td>Are the costs of personalized promotion too high versus other ways of producing same nutrition gains (cost effectiveness)?</td>
<td>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</td>
<td>Nutritional outcomes did not improve – why? - Beneficiaries, having received messages do not change practices - Messages were wrong</td>
</tr>
<tr>
<td><strong>Micro-credit</strong></td>
<td>Do the costs of providing credit at micro level have higher net returns than other uses of capital?</td>
<td>Why weren’t loans made? - Loan officers do not generate lending activity - Low repayment rates decreases the total possible lending</td>
<td>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</td>
</tr>
<tr>
<td><strong>Impact Evaluation</strong></td>
<td>Is the scope of the findings sufficient to justify time and cost of evaluation?</td>
<td>Evaluation not completed even after baseline is done - Project is not carried out - Contaminated experimental design - Poor quality data collection</td>
<td>Evaluation results have no impact on beliefs or behaviors of key actors.</td>
</tr>
</tbody>
</table>
3. Second Generation: The rise of the randomistas

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. The criticism is of M&E practice, a critique that has two key elements:

- evaluation was too ex ante and need 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.

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-
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 Kenya13. 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 schooling promoted by the project) went down in absolute numbers and more than 100 percent of the rise in schooling was in the private sector.

The second problem with using either “before and after” or “project and non-project” areas is the 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 benefits14. 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 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?

---


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

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
<th>Difference over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>$\bar{Y}^{Project, T}$</td>
<td>$\bar{Y}^{Project, T+N}$</td>
<td>$\bar{Y}^{Project, T+N}_t - \bar{Y}^{Project, T}_t$</td>
</tr>
<tr>
<td>Non-Project</td>
<td>$\bar{Y}^{No, T}$</td>
<td>$\bar{Y}^{No, T+N}$</td>
<td>$\bar{Y}^{No, T+N}_t - \bar{Y}^{No, T}_t$</td>
</tr>
<tr>
<td>Difference non-project and project exposure</td>
<td>$\bar{Y}^{Project, T}_t - \bar{Y}^{No, T}_t$</td>
<td>$\bar{Y}^{Project, T+N}_t - \bar{Y}^{No, T+N}_t$</td>
<td>$(\bar{Y}^{Project, T+N}_t - \bar{Y}^{Project, T}_t)_t - (\bar{Y}^{No, T+N}_t - \bar{Y}^{No, T}_t)_t$ (Differences in differences)</td>
</tr>
<tr>
<td>Difference in outcome with and without the project</td>
<td>$(\bar{Y}^{Project, T+N}_t - \bar{Y}^{Counter-factual without the Project, T+N}_t)$ (the latter is unobservable)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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” or “impact” of a project—propensity matching, regression discontinuity, instrumental variables—even without an ex ante experimental design (Angrist 2011). 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).

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, and especially RCTs, in development projects\textsuperscript{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\textsuperscript{16}. J-PAL (as of June 2011) has 116 studies completed or underway, IPA has over 500 staff working around the world. In 2004 the Center for Global Development (CGD), with support from the Gates and Hewlett foundations, launched the “Evaluation Gap Working Group” headed by Ruth Levine at CGD to examine what could be done to improve evaluation in development projects. This group

\textsuperscript{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)).

\textsuperscript{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 Oportunidades), in Mexico. This was influential as it was a rigorous evaluation of an ongoing government program carried out at scale.
produced a report in 2006—“When Will We Ever Learn?” that made recommendations for improving the support for impact evaluations. This in turn has led to the creation of 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.”

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” of specific techniques within such projects and the assessment of the causal impact are carried out by researchers rather than the people who are usually expected to implement a project.

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.¹⁷ That being said, many agencies do not publish or document failures. This also includes not publishing projects that had planned to be evaluated using an RCT but which failed in the implementation phase.

4. Third Generation: Learning from Experimentation

The use of more RCTs and more RIE in development funding organizations is an important advance. 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 tactic¹⁸ which is still wedded to outmoded top-down strategies for innovation and learning in development projects.

¹⁷ For instance, the World Bank’s research group has been promoting building RCTs into Bank operations since at least the mid-1990s.

¹⁸ The notion that RIE is, in and of itself, the basis of an alternative strategy for poverty reduction or improvement in human well-being is, at best, surreal. The evidence on all levels, macro and macro, is compelling that far and away the major determinant of reduction in absolute poverty is the growth rate of the economy a household lives in (Kraay 2002, Narayan, Pritchett and Kapoor 2008). The contribution of “programmatic” interventions to poverty reduction is tiny by comparison and hence the contribution of “better programs/projects due to RIE” must be only some fraction of that. While “we will rigorously evaluate everything we do” sounds reasonable it implies the converse: “we will only do what we can rigorously evaluate” which is just plain wrong as a development strategy. The rapid and sustained economic growth in China and India has been the major driver behind the reduction in global poverty. The annual social welfare gain from the rapid growth in India since 1991 versus a counter-factual of average growth is trillions of dollars, possibly as big as US GDP. Any development activity that contributed to that growth paid for itself, hundreds of times over. But no one will ever be able to rigorously establish what contributed to India’s economic growth, even retrospectively, much less prospectively. Imagine the claim “there is no rigorous evidence that Martin Luther King contributed to the improvement of civil rights for African-Americans in the USA.” If one construes “rigorous” to mean “supported by an RCT” this statement is true. But it’s doubtful that anyone who supported Martin Luther King, financially or otherwise, loses any sleep over the thought they wasted their efforts because there is no “rigorous” evidence of impact. To only focus on what can be “scientifically” investigated via RIE/RCT might be an option for academics whose output is measured in published papers, but cannot be a strategy for any organization whose output is measured in impact on human well-being.
There are three fundamental reasons why M&RIE needs to be supplemented by MeE:

- A rugged\textsuperscript{19} and contextual fitness function over a high dimensional and complex design space implies that learning “what works” has to be flexible and dynamic.
- 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 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, what we propose is an extension of the use of alternatives with a counter-factual and rigorous measurement to learn about causal models inside the implementing agencies and organizations\textsuperscript{20}.

### 4.1 Crawling a rugged and contextual design space

Imagine you run the experiment of drilling for water at spot X on the surface of the earth on September 1\textsuperscript{st} 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 patterns rainfall and run-off and knowledge of the surface and underground topology your experiment taught you nothing useful. Every useful statement is about the future and experiments can only make statements about the past.

Banerjee, Glennerster, and Duflo (2008) report on an experiment aimed at increasing the attendance of auxiliary nurse-midwives 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). 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 4 below.

\textsuperscript{19} The function relating an intervention may be deemed “rugged” if it is very sensitive to details of context or of implementation. This is in contrast to “smooth” functions that do not vary much across circumstances.

\textsuperscript{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.
Table 4: Design space for “pay for performance” projects

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

Source: Banerjee, Glennerster, and Duflo (2008)

The results of their experiment are displayed in Figure 2 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 fallen steadily over the implementation period.
Figure 2: 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? That incentives don’t work? No, there is too much evidence that incentives work. That increased monitoring 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 run by the same NGO as involved in the nurse project increased teacher attendance enormously (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 with any “rigor” at all 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.

This example illustrates two generic features of development projects and policy: (a) high dimensional and complex design space and (b) contextual and rugged fitness spaces (mappings from elements of design space to results).

4.1.1 High dimensional and complex design spaces

Names of development projects are labels for classes and a specific project is an instance of a class, that is, a project as implemented is just one point or element in the design

Note: “Run by an NGO” not through the usual hierarchy that a large system has to rely on.
space that defines the class of projects. 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 *conditional cash transfer* project, a *community block grant* project, a *livelihoods* project. None of these names of classes of projects conveys enough information that anything could be learned about the class from an evaluation of an instance. Design spaces of development projects are *high dimensional*.

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, exactly with what, for whom exactly, and when exactly. The operational manual of a “simple” project may run to hundreds of pages. As illustrated in Table 5 even the simplest possible design of a CCT has eleven dimensions. Suppose that for each of these eleven dimensions of the design space there are only 3 possible elements (which is a radical simplification as some dimensions have many more choices and some dimensions are continuous). Even with this radical simplification of reality there are $3^{11}=177,147$ distinct CCT projects as *instances* of the *class* of a “CCT project.” With so many fewer CCT projects in existence, it is logically impossible to infer causality even to whole subsets of conditions. For that we need a theory (a theory of change) to fill in those things that cannot be discovered in such a piecemeal fashion.

And CCTs are simple as what is “delivered”—cash—is simple. Think of a “teacher training” project or a “micro-finance” project or a “road construction” project or a “livelihoods” project. Each of these has at least as many dimensions of the design space and at least as many possibilities per dimension hence there are hundreds of thousands of possible project variants for each type of project. Everyone who has ever had to design and implement a development project realizes that the granularity at which development happens is incredibly fine.

The design space of development projects is not only high dimensional but also a conceptually 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”? 
### Table 5: Design Space for CCT projects

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

### 4.1.2 Rugged and contextual fitness functions

The impact of a development project (whether outputs or outcome or impacts) is a fitness function over the design space of the project. Learning from development projects has two elements. First, 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 is *contextual* in that the mapping itself from design space to impact differs from context to context. Doing *exactly the same project* can have different impacts depending on where and when it is done.
Rugged fitness functions: non-linear and interactive. Perhaps the most striking thing that is being learned from the “new experimentalism” of behavioral economics is that seemingly small changes in design can have big impacts on outcomes. An experiment to look at cheating at Carnegie Mellon staged an experiment in which the subjects saw one person (a hired actor) clearly and publicly cheat with no consequences to examine whether this affected whether this affected student cheating. When the cheating actor wore a plain white t-shirt then the public cheating led to 25 percent more students cheating than the experiment without the actor. 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 2009). In picturing the design space and fitness function what might seem like a slight variant in the design space (all other features constant except the t-shirt the actor doing the cheating is wearing) causes enormous differences in the outcome.

Non-linear fitness functions. A number of experiments have found sharp non-linearity in impacts. 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. 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. 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 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 2007) 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 (two whom the transfer is given with whether or not the cash transfer is conditional).

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

---

22 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.
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 uses this argument to point out that the practice of promoting “best practice” for firms element by element (e.g. “this is ‘best practice’ performance pay, this is ‘best practice’ production process, this is ‘best practice’ marketing) makes no sense at all when there are, generically, interactions amongst these elements.

Discovering the best project design requires a crawl over the design space to evaluate the fitness function at various points. Think of learning about project effectiveness as an optimization algorithm, which is a plan for sequenced and iterative movement over the design space. In a discrete space with few possibilities and a low cost (in time and resources and capability) per evaluation a simple grid search is a feasible optimization procedure. The special case of that is if it is known a priori (from validated theory) that the fitness function is flat over the relevant dimensions then only one good experiment is needed to find “the evidence” about “what works.”

Alternatively, if one is maximizing in dimensions in the design space with continuous variation and the fitness function can be safely (based on theory) to be “regular” (perhaps even linear as in Figure 3 or a simple quadratic) one can pursue a process of iterative search where a point on the fitness function is evaluated and then each new move attempts to move to a higher place on the fitness function.

---

23 A conceptually easy extension is if theory provides “invariance laws” the describe how the measured quantity varies with respect to alterations in the conditions under which the experiment is carried out it is still possible only one experiment is needed even if actual observed quantities from experiments will vary but in entirely predictable ways.
Figure 3: A “regular” fitness function (linear in each dimension, no interactions) over a simple (two dimensions, three possibilities in each dimension) design space

If the fitness function is rugged over a high dimensional and complex design space then learning will be difficult. Figure 4 illustrates a “rugged” (like the Swiss Alps) fitness function over a two dimensional design space with three elements per dimension (A,B,C in dimension 1 and I,II,III in dimension II).

Figure 4: A “rugged” results surface over a design space

Contextual fitness functions. What is learned from the rigorous evaluation of a project? Rigorously we learn nothing at all useful because all we learn is what happened in a particular instance in a particular place in the past. All extrapolation from one context to another context (including from the context of the past to the context of “now”) is based on theory and evidence is only as good as the theory used to interpret the evidence from one context to another context.
Technically experiments estimate a Local Average Treatment Effect (LATE). What is “local” is not itself a matter of evidence but of theory. Conceptually a “fitness function” is a evaluative function over a design space (in evolution fitness is species survival over genetic designs, in computer engineering the fitness could be time to complete a specified task over a design space of machine design, in marketing fitness is sales over alternative media buys, in cooking fitness could be meal tastiness over recipes, etc.).

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 designers and hence not elements of the 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 design to let such staff go, employment law in the country might forbid such action or such action is subject to legal challenge. In this case the exact same project from a design perspective is implemented in a different context and since the fitness function depends on the behavior of the implementing agents and this behavior depends on both design and contextual features the impact of a project as evaluated from a fitness function over the design space might not just be rugged but also contextual.

Evidence emerging from the RCTs and RIEs done so far is suggesting that results are contextual. In the education domain for instance there have been a number of rigorous estimates of the impact of class size—and they are completely different, ranging from impacts that make class size seem plausibly a cost-effective intervention (e.g STAR in Tennessee and the Maimonides rule in Israel) to finding exactly zero impact from smaller class sizes with regular civil service teachers in places in Kenya and India. Similarly, “pay for performance” schemes have found wildly different impacts on student learning in Israel, Kenya, and Andhra Pradesh India.

4.1.3 Crawling a high dimensional and complex design space with rugged and contextual fitness function

Of course the readers acquainted with computational complexity theory will recognize the characterization of the problem of finding the development project with the optimal impact as in the class of NP-complete (which are both in the class of NP (solvable in finite (polynomial) time) but also NP-hard (so there are no known, general, quick algorithms). A particular simple type of NP-complete is the NK problem class as introduced by Stuart Kauffman (references) that has a fitness landscape with “tunable ruggedness.” Both of these are related to the notion of evolution and complex systems.

Drawing on the intuitions from this enormous literature on computational complexity and algorithms (e.g. Cormen, Leiserson, Riveset, Stein 2009)—and the analogous work on how organizations cope with NK or NP-complete problems in practice—where it is known there are no general solutions available there are several principles to the type of approximation and heuristic algorithms that are used. One, try and get into a desirable part of the fitness landscape with good guesses. Two, rapid iterations are essential to crawl the space. Three, too early lock-in to a single region of the design space is to be
avoided as standard algorithms that produce local optimum often cannot crawl sequentially to other regions of the fitness landscape.

The usual approach to project design with impact evaluation (whether RIE or even RCT) is of limited use for three reasons.

First, the use of RIE, and in particular RCTs, is intrinsically very expensive because the data required 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 5). This means the cost per evaluation in the search algorithm is very high, which is the 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 5: Information requirements for various types of learning and their incremental costs and timing**

<table>
<thead>
<tr>
<th>Time</th>
<th>Y 1</th>
<th>Q1</th>
<th>---</th>
<th>---</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y 2</td>
<td>Q2</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Q4</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Y 3</th>
<th>Q1</th>
<th>---</th>
<th>---</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q2</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Q4</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Second, RIE (including RCT) on *outcomes* is slow—because usually the causal model of mapping from outputs to outcomes (e.g. micro-finance to income, new innovations to adoption, education to wages) is slow. Figure 5 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 to better identify “β” (the LATE) it precludes learning about the characteristics of the fitness landscape other than the evaluated project. 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. 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.

4.2 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 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.

For example, early evaluations of the Total Sanitation Campaign in India were hampered by lack of consistent evidence that the activities of the program even took place or were done particularly well. 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.

Another example involved social mobilization to support local governments in Karnataka, India. 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 form the real world of projects. What is more 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 randomization but is crucial to project outcomes nonetheless.

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 condition them out, this is explicated in Pritchett 2012.
25 Social Science and Medicine among many others.
That said, rigorous evaluation can be done of how different designs on implementation feed into project success at least in producing outputs. 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.

4.3 How organizations and systems learn

There is a puzzle about why RIE in general and RCTs in particular have not been more widely used in development projects. There has been no methodological or technological advance over the established practice of impact evaluation in the USA since the 1970s. The reason why there have not been more RCTs before, and what limits expansion of RCTs now, seems to be that governments, development funders, and implementing agencies have not particularly wanted them. But if someone comes with a tool that would allow implementing agencies to do better at what they are trying to do then why wouldn’t they, on hearing of this proverbial “better mousetrap” for learning about impacts beat a path to the door?

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 a perhaps 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.
4.3.1 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 allowed evaluations which focused on outputs was 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 (Hammer 1997) determines net usage. 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 5). 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.

4.3.2 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

5. Structured Experiential Learning: Introducing “e”

The value of information is related to the chance that a decision will be changed as a result of it. For data to be useful it must be regular, reliable and relevant to the project and its objectives. It must be disaggregated, de-mystified and ultimately facilitate better planning and implementation of the project. While development projects devote substantial resources to “M” (often much more than “E”), the data collected is often not linked to project management and implementation structures. The process of determining what data should be collected and why, often remains external accountability centered—reporting for compliance — with very little participation (if any) from the implementers despite their deep microscopic understanding of the reality on the ground. So for many implementers, monitoring is just another item to check off their long list of activities. In fact, the data collected is often stored in text documents or in report formats required by the project supporters but hardly ever analyzed. In addition, even when the program staff cares about these reports, the donor rarely provides them with useful feedback because it is also a check box for the donor (monitoring report received - check). The type, frequency, and analysis of monitoring data is often out of sync with the critical decision junctures in the project cycle so that managers, in spite of having substantial resources and units devoted to monitoring, are often operating on what they can see.

Our proposal is to explicitly add a new “e” to the standard formulation of “M&E”, so that development organizations (i.e. country government/agency, assistance/aid organization, foundation, NGO) have a learning strategy for its portfolio of projects. We define “e” 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 “evaluation”. “e”, is a way to bring the informal process of experiential learning, from project implementation, explicitly into the overall strategy of development organizations.

Essentially, (e)xperiential 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 over the course of the project cycle. The idea is to use the development project itself as a learning device and not as a single “experiment”. It is a way of searching for what works best (i.e. dynamically crawling the design space), by using variations in design within the project, to identify differentials in the efficacy of the project on the process of inputs to outputs (both measured at regular intervals), for real-time feedback into implementation, at key decision junctures.

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”.
Let us say that you are Ms. Eager Beaver\textsuperscript{31}, manager of a development project in the fictitious 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.

5.1 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 back to 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.

1(a). Begin with a clear definition of the problem you are trying to solve. Then state the goal as well as the magnitude of the desired impact. Often the magnitude of goals 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. 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. 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

\textsuperscript{31} Perhaps an older version of Ms. Speedy Analyst (Ravallion 1999).
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 need to be clearly articulated. 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 or hypothesis, at each link in the causal chain. The figure below illustrates her work.

**Figure 6: Reverse Engineer Goals to Instruments – Education Example**

<table>
<thead>
<tr>
<th>Assume teachers are conscientious</th>
<th>Assume teachers are physically present and teaching</th>
<th>Assume all children are in school and they learn in class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers are provided good training</td>
<td>Teachers are using the practical skills they were taught</td>
<td>Teachers are teaching effectively in class</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Goal: All children can read by Grade 3</td>
</tr>
</tbody>
</table>

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. The following is her causal chain.

**Figure 7: Reverse Engineer Goals to Instruments – Water Example**

<table>
<thead>
<tr>
<th>Capacity building for delivery of water</th>
<th>Delivery of water supply exists in villages</th>
<th>People have access to clean water in their villages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Goal: Increase water usage</td>
<td></td>
</tr>
</tbody>
</table>

Preliminary and Incomplete: For Comments Only
Page 30
June 22, 2012
Step 2: Design a Project

2(a). Design a project (let us call it P1) that will help you achieve your goals. It is important to note that deciding on a concrete project is creating an instance of the class of possible projects to achieve your goals (as discussed in section 4.1).

2(b). Specify the timing, magnitude and gain from the project for each link in the 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\(^{32}\), results based framework, a “complete, coherent, causal chain” or a theory of change,\(^{33}\) and is often required for project approval either within an organization, or by an outside funder. This is true for governments, large multilateral organizations like the World Bank, private foundations, as well as internally within NGOs.

Perhaps paradoxically, one of the biggest advantages of the increase in RIE 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. 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 evaluated, but “we expect the project to raise the score on this particular assessment instrument by 20 percent” is.

In her zealous way, 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 it for funding approval.

---


\(^{33}\) http://www.theoryofchange.org/background/basics.html
Step 3: Admit we do not know what combination of design parameters will work

This is the hardest step. The reality is that 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). Even though development project managers recognize and accept this, the context in which they work does not allow them to admit that they do not know what will work. That is, advocacy is
often solution driven and in order to attract donor funds you cannot admit this even if you know it is true.

Let us look at Eager Beaver's work. She made some implicit choices in designing her project in step 1. Specifically, she began by assuming that all children were in school (see figure 6). But, did she look at what percentage of children in Utopia go to school? Do schools exist (or does demand exceed supply)? What causes children to go to school? If they do not go to school, why do they not go? Is it expectations of returns to education? Without digging deeper, we just do not know. A study in the Dominican Republic found that students underestimated the returns to education because they relied heavily on information within their own community, which was biased downwards. However, when evaluators corrected this misperception on returns to education, enrollment increased (Jensen 2010). 

In addition, what contributes to children learning? Is it teacher training? Is it parent's education? Is it child motivation (e.g. a high stakes test)?

The key in designing a development project is balancing the complexity of the problem with the design of the project. The trade-off is that in order to get “authorization” or approval for a project, organizations want specificity at a level of detail that almost always exceeds the available, context relevant, evidence. This leads to excessive claims of confidence about what will work.

The solution is to create a process, which authorizes projects that are a structured crawl over the promising parts of the design space (see step 5 below). This balances giving project implementers the flexibility they need to find out what works with the accountability over use of resources that “authorizers” need.

We define the design space as the total of all possible combinations of your design elements (see tables 4 and 5). In projects, the form of the design space is often assumed (for simplicity) to be linear and that design elements in the space do not interact with each other. This is not true. In fact, the more design elements you have and the more options you have per element, the more complicated the dimensionality of your design space. As discussed in section 4.1, any development project is one instance in a class of projects or one element in the design space that is high dimensional and complex. Furthermore, the fitness function may be rugged (small changes in project design can
have big changes on outputs or outcomes) and is contextual in that the mapping itself from design space to impact differs from context to context.

There is usually a vector of outcomes of any intervention, not just the one you are measuring. “E” anticipates the design of the project ex ante and therefore you are locked in for 3-5 years with no dynamic feedback loop into project design. You also cannot make any changes to the project mid-stream because you will contaminate the sample rendering the data you have collected, useless. Say, for example, you were managing a project that tried to keep orphans and vulnerable children in the community by finding them foster homes. Your donor has requested that the project be (E)valuated for impact. One year later you find that several of the orphans placed by the project are complaining of abuse in their foster homes. You want to use this important information to make changes in your project (e.g. introduce random home visits every six months, or do background checks on the parents before placing children in their homes) which would directly affect the (E)valuation – what do you do?

Let us look at another example where mid-course corrections were made. 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 flexibility of reallocating project resources into another dimension of the design space (providing complementary services) the whole project would have failed.

Locking yourself into one project limits the potential learning from both the upside and the downside. What if there was a way for you to first learn about your project by creating variations in the design parameters before you lock into one arbitrary design that you think might work?

We propose (e)xperiential learning which uses design variations within a project to identify differentials in the efficacy of the project on inputs and outputs for real time feedback into project implementation.

**Step 4: Identify the key dimensions of the design space**

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 figure 10. Let the project P1, selected in step 2 be D1 (A, α, I). Recognize that what you selected was one project in a class of many other similar projects.

**Figure 10: Design Space for Teacher Training**

<table>
<thead>
<tr>
<th>Design Parameters</th>
<th>Design Space (DS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D1=P1 D2 D3 D4 D5 D6 D7 D8</td>
</tr>
<tr>
<td>Location (A,B)</td>
<td>A   A   A   A   B   B   B   B</td>
</tr>
<tr>
<td>Content (α,β)</td>
<td>α   α   β   β   α   α   β   β</td>
</tr>
<tr>
<td>Follow-up (I, II)</td>
<td>I   II  I   II  I   II  I   II</td>
</tr>
</tbody>
</table>

**Step 5: Specify the timing, magnitude and potential gain for each of the possible project variants in your design space and select new alternate projects**

5(a). Specify the timing, magnitude and potential gain for each of these possible project variants D2 through D8. Essentially, you are creating new projects by varying design parameters and with a clear theory of change for each but with the same desired outcome.

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 finds the magnitudes of the project variants, where P1 = D1 (see figure 11).
5(b). Select two (or more) new alternate projects based on the highest uncertainty and upside potential.

When deciding on project variants you want to look at both the magnitude (mean/average) as well as the upside potential (variance). Therefore, looking at figure 11 above, D2 has both a higher magnitude and upside potential than P1 and could be considered as P2. D3, D4, D6, D7 and D8 all have lower magnitudes and lower upside potential than P1 and therefore may not be worth considering. While D5 has a lower magnitude, it has a huge upside potential (variance) – higher than P1 – and would be worth considering as P3.
Below is an example of how Fraker and Shah (2011) used a similar approach to search the design space and filter out the best options to pilot. Figure 13 highlights that projects 4, 7, 10 and 11 look much better than the rest.

Source: Fraker and Shah (2011)
5(c). Repeat step 2(c) for each of the new projects (i.e. determine indicators for P2 and P3). It should be noted that the outcome you determined in Step 2 will and should not change. What may change per project will be the inputs and the outputs depending on the design parameters. Therefore P2 and P3 could serve as internal counterfactuals for P1. This is what we refer to as strategically crawl the design space.

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

As discussed earlier, all development projects collect (M)onitoring 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.

In addition, the findings of the (E)valuation often come too late – after the project has closed or ended. So neither “M” nor “E” is perceived as a useful exercise for project managers and implementers.

**Figure 14: The problem with only doing “M” and “E”**

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 to create a legitimate space for failure.
Introducing (e)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 15: Sequential crawl for teacher training project**

**Step 7: Implement the approved sequential crawl and learn**

7(a). Pilot the projects P1, P2 and P3 for the duration of time that you determined.
7(b). Collect all input and output indicators for all three projects (“M” data).
7(c). If changes need to be made to your project design, repeat step 4 (search the design space).

Throughout the pilot, analyze the “M” data you collected for all three projects (including additional data introduced on the basis of feedback from ground personnel) and use it to help make decisions in the pre-specified sequential crawl (see figure 15). 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. Figure 14 and 16 illustrate the difference between using M&E alone and adding “e” to the approach.
7(d). Go back to the system with new options. At this point you have crawled the design space and have a better sense of what key design parameters your project should have to help you achieve your goal or desired outcome. Some options you could consider after crawling the design space are:

- Scale up the most successful design and identify what priority questions and elements you want to focus on for the (E)valuation,
- Re-configure into a different project and keep crawling the design space, or
- Shut down the project.
5.2 The advantages of adding “e” to the mix of learning approaches

In section 5.1 above, we described how a manager of a development project could use our seven step approach to develop a learning strategy that achieves more than monitoring, is cheaper than impact evaluations and has timely dynamic feedback loops built into the project.

There are several advantages to using “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 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 have to be considered?

Third, when you split the project into links, you amortize large projects into small projects, which is more cost effective. The key insight is using the internal variation in project design (i.e. P1, P2 and P3) to measure the effectiveness. This is enormously more cost effective than impact (E)valuation. As illustrated in Figure 3 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 5 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, “e” is helpful and is in the interest of both the implementing organization as well as the project implementers. Experiential learning is about doing what the organization
wants to do, better, while independent Evaluations could be viewed as a double edged sword. One of the common issues in “E” is the disinterest, if not outright hostility, of the project implementation team to the Evaluation team.

During a recent visit to a dynamic and effective NGO, the director introduced the person responsible for the RCT as “this is the person who makes sure we don’t benefit any children”. This is a good illustration of the trade-off between what the NGO wants—which is real time, experiential learning about how to accomplish their objectives—and what impact (E)valuations require, which is to protect the “integrity of the experiment”.

Fifth, “e” 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 (see figure 18).

**Figure 18: The vicious cycle of “M” data**

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.

It is important to point out that we are not dogmatic about our MeE approach. It is clear that “M&E” alone, as currently practiced, is insufficient as a learning tool and there is a dire need to explicitly learn from development projects in real time, with built in dynamic feedback loops. Ours is but one approach as a number of others have been grappling with the limitations of M&RIE and have come to similar conclusions about the need to move on. 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) note that interventions never work indefinitely, in the same way and in all circumstances, or for all people. They argue that projects are embedded in social systems; require active engagement of individuals; and are open systems that cannot be isolated or kept constant. This is what we refer to, as a “rugged” fitness environment—“what works” cannot be easily extrapolated. They define “Realist Evaluation” as one that asks not, ‘What works?’ but instead asks, ‘What works, for whom, in what circumstance, in what respects, and how?’ Our take on MeE is similar to this “Realist Evaluation” a way of designing a learning and evaluation process consistent with both rugged fitness environments but also rigor in evaluation.

Szekely (2011) talks about the conflict of interest between evaluators and implementers, he 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.” He argues that development is a moving target and therefore more integrated approaches are needed to institutionalize learning. He suggests Results-Based Social Policy Design and Implementation systems, which could look like the diagnosis-design-implementation-evaluation-analysis-finetuning-implementation described by Greenberg (1968).

Other such approaches include Khagram et al (2009) who suggest diagnostic, contextual approaches to experimentation and innovation for development in the twenty-first century – Impact Planning Assessment Reporting and Learning Systems (IPARLS). The Impact Evaluation For Development (IE4D) Group released a document entitled, “Principles for Action”, in 2011, which 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.”

6. Ideal mix of MeE within an organization

In section 5, we explored what an experiential learning strategy would look like for a development project manager. In this section we examine what a learning strategy would look like for a manager of a portfolio of development projects. This could either be someone responsible for a sector, a country or an organization. Essentially, what would an organizational strategy for learning look like?

We propose a new learning strategy for development organizations that can produce more effective development projects. The learning strategy consists of a project specific mix of:

- (M)onitoring 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.

- **Impact (E)valuation** provides the most rigorous estimates, of the causal impact of projects on outcomes possible, given the nature of the project.

Making MeE work for you requires change management. A change in the culture of project design and implementation, to allow learning from experience to explicitly feed back into design changes, will ultimately improving the chances of success for your development project.

This learning strategy is the project level counter-part of an emerging 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. 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, 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 problem becomes the solution and leads to “capability traps” in administrative capability (Pritchett and Woolcock, 2004; Andrews, Pritchett and Woolcock, 2011; Andrews, 2011; Filmer et al 2000, 2002).

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 compared 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.

**Figure 19: Evaluation with “multi-sector” funding**

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.

1. **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.\(^{35}\)

Therefore, the optimal MeE strategy for a government might look like table 3.

**Table 6: Long on “M”**

<table>
<thead>
<tr>
<th>Projects</th>
<th>MeE</th>
<th>Portfolio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routine</td>
<td>“M” only</td>
<td>80%</td>
</tr>
<tr>
<td>Innovation</td>
<td>“M” + “e”</td>
<td>10%</td>
</tr>
<tr>
<td>Flagship</td>
<td>“M” + “E”</td>
<td>10%</td>
</tr>
</tbody>
</table>

2. **Large aid organization**

Large aid organizations like the World Bank operate in 6 regions; have 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.\(^{36}\) All World Bank projects have funds for “M&E” and all collect “M” data (even though most of it is not used). It is simply not feasible, desirable or cost effective to conduct an impact evaluation of each and every one of these projects. Therefore it is necessary to develop a differentiated MeE strategy for the portfolio. This could be done for a sector or a country or a region.

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” projects like PROGRESA (in Mexico), SSA (in India) and KDP (in Indonesia). On these large, scalable and novel projects, it would be important to conduct “E” to determine a rigorous estimate of the causal impact. Finally, 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 projects are on the rise. In fact, USAID recently launched Development Innovation Ventures (DIV) to invest in projects that will produce development outcomes more effectively and cost-efficiently while managing risk and obtaining leverage.

The optimal MeE strategy for a large aid organization could therefore have a majority proportion on doing “M”.

\(^{35}\) 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.

\(^{36}\) World Bank data as of 2006.
Table 7: Long on “M”

<table>
<thead>
<tr>
<th>Projects</th>
<th>MeE strategy</th>
<th>Portfolio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routine</td>
<td>“M” only</td>
<td>60%</td>
</tr>
<tr>
<td>Innovation</td>
<td>“M” + “e”</td>
<td>10%</td>
</tr>
<tr>
<td>Flagship</td>
<td>“M” + “E”</td>
<td>30%</td>
</tr>
</tbody>
</table>

3. 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.

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.

Table 8: Long on “e”

<table>
<thead>
<tr>
<th>Projects</th>
<th>MeE strategy</th>
<th>Portfolio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routine</td>
<td>“M” only</td>
<td>30%</td>
</tr>
<tr>
<td>Innovation</td>
<td>“M” + “e”</td>
<td>50%</td>
</tr>
<tr>
<td>Flagship</td>
<td>“M” + “E”</td>
<td>20%</td>
</tr>
</tbody>
</table>

4. Single sector 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.
7. Conclusion

The weaknesses of the traditional approach of M&E are widely recognized. “E” has made significant progress in becoming more rigorous, is now widely accepted and is viewed as being attractive. This is truly commendable. However, as we have seen, simply adding rigor to “E” alone cannot be a learning strategy. For some large organizations like the World Bank, it is not feasible or cost effective to conduct impact evaluations for every single project. DFID recently commissioned research on “developing a broader range of rigorous designs and methods for impact evaluations”, after acknowledging that a wide range of development activities are inappropriate or just not viable for RCTs. They estimated that only 5% of development projects are suitable for RCTs. Their terms of reference stated, “the primary purpose of this research is to establish and promote a credible and robust expanded set of designs and methods that are suitable for assessing the impact of complex development programmes.”

In this paper, we introduced “e” or experiential learning into the mix of M&E, which achieves more than monitoring, is cheaper than impact evaluations and has timely dynamic feedback loops built into the project design.

Our MeE approach is a learning strategy for development organizations. As a new learning strategy it hopes to improve upon existing development practice in several ways.

First, by having monitoring data feed directly into management decisions during project implementation in a structured way, MeE hopes to improve the quality of “M”, moving from merely tracking inputs and their uses for fiduciary accountability (a necessary, but limited, element of M), to an analytic use of “M” data to learn during the project cycle.

Second, MeE takes the insights from the RIE movement, that has emphasized the need for the use of a counter-factual to estimate project impacts, and brings them into project design and implementation, so that within project variation can be used to find the best context specific project design. Monitoring can be designed to make assessments of impact but requires measurement of outcomes (borrowing if micro credit, production and acreage if agricultural, attendance at meetings if accountability) outside of the project. Again, this may be most relevant to governments and their statistical offices but as others begin to collect more data, they can measure such things, too. These can then be linked to variations in implementation from purely internal data.

Third, MeE raises the profile and makes the learning that happens during project implementation, explicit. Currently it is very difficult to explicitly use the learning that happens during project supervision or implementation, as it is not considered “knowledge”. Therefore, this enormous body of “knowledge” often does not diffuse inside development organizations or even persist across projects.

Fourth, the use of MeE must be premised on changing the nature of what a development organization authorizes as a “development project.” The current practice in many mainstream development organizations is to specify the project too completely
beforehand hence the scope for learning is too limited by *ex ante* project design. With the explicit recognition of the value of “e”, development organizations can authorize a process of “crawling the design space” in which project implementers are empowered to learn and change during the project period. This balances the needs for fiduciary control with the goals of learning and development outcomes.

The MeE approach is both a project and development organizational learning strategy, which complements an “experimentation” approach to development. This approach acknowledges that there are common principles but emphasize heterogeneous ways of implementing those principles. It tries to strike a balance between “one size fits all”—which works well for organizational control but does not give sufficient context flexibility—and “anything goes”—that provides too much scope for continued dysfunction. MeE intends to be a process that allows development organizations to make the search for “best fit” an operational strategy.
References


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


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


