Searching for the Devil in the Details: Learning about Development Program Design

Sara Nadel and Lant Pritchett

Abstract

Motivated by our experience in designing a particular social program, skill set signaling for new entrants to the labor market in Peru, we articulate the need for, and explore the empirical consequences of, alternative learning approaches to the design of development projects. Using a simulation, we demonstrate that even with only modest dimensioned design space and even modest "ruggedness" of the outcome with respect to design a naive iterative approach of "crawling the design space" dominates an RCT learning strategy. We suggest that the empirical results of RCTs to date are consistent with social programs having high dimensional design space and outcomes sensitive to design and hence project/program/policy design must depend on more robust learning strategies than the attempt to directly apply results from "systematic reviews" or move prematurely to an RCT.

JEL Codes: O12, O22, C93, D04

Keywords: research, program design, RCT, development

Center for Global Development
www.cgdev.org

Working Paper 434
August 2016

Searching for the Devil in the Details: Learning about Development Program Design

Sara Nadel
Harvard Kennedy School of Government

Lant Pritchett
Harvard Kennedy School of Government

We are grateful to our many colleagues at Harvard Kennedy School for ongoing discussions on these issues, particularly Rohini Pande, Dani Rodrik, Ricardo Hausmann, Michael Callen, Asim Khwaja and Salimah Samji and at Center for Global Development, particularly Justin Sandefur. In addition we have had helpful discussions on the issues of learning about projects with Andrew Fraker, Neil Shah, Vijarendra Rao, Howard White, and Michael Woolcock. We also thank two anonymous reviewers of the CGD working paper. The paper represents the views of the authors alone and not necessarily those of their affiliated organizations.

The Center for Global Development is grateful for contributions from its funders and Board of Directors in support of this work.

Sara Nadel and Lant Pritchett. 2016. "Searching for the Devil in the Details: Learning about Development Program Design." CGD Working Paper 434. Washington, DC: Center for Global Development.
http://www.cgdev.org/publication/searching-devil-details-learning-about-development-program-design-working-paper-434

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

Contents

1 Introduction

Upon inspection, each project turns out to represent a unique constellation of experiences and consequences, of direct and indirect effects.
Albert O. Hirschmann, *Development Projects Observed*

The advantages of using random assignment to estimate the causal impacts of social programs have been well known for many decades.[1] Recently both academic development economics and development practice (donors and NGOs) have experienced a boom the use of randomized control trial (RCT) approaches in impact evaluation and field experiments.[2] Some of the enthusiasm for vastly expanding the use of randomization was the notion that greater use of RCTs in independent impact evaluations of development programs/projects/policies would produce sufficient number of "rigorous" studies for a "systematic review" about "what works" that would guide available resources into effective interventions. That a lack of "external validity" of RCT evidence–rigorous estimates of causal impact from one context may not constitute useful evidence for another context–might limit the gains from the "systematic review" approach has been debated extensively ((Deaton, 2009), (Heckman and Urzua, 2010), (Ravallion, 2009), (Pritchett and Sandefur,

[1] Randomization has been used to evaluate social programs in the USA at least since the 1960s, with prominent experiments on "income maintenance" (four large scale experiments between 1968 and 1982), health insurance (the RAND Health Insurance Experiment began in 1971), policing (Kansas City preventing policing experiment began in 1972), job training (randomized evaluation of the Job Training Partnership Act (JPTA) authorized in 1982) and the capability to implement randomized evaluations of social programs has been long available in organizations like MDRC, Abt Associates, Mathematica, Rand Corporation and others.

[2] Vivalt's (2016) database records results of over 600 evaluations 80 percent of which are RCTs and Shah, Wang, Fraker and Gastfriend (2015) estimate that there are now over 2,500 development RCTs.

2014)).

We raise a new concern with learning "what works" either from existing RCTs or even by using an RCT. We argue nearly all existing evidence about development programs lacks *construct validity*. Any specific program/project/policy.[3] is an *instance* of a *class* of possible programs defined by a *design space* That is, any specific "teacher training" or "job placement agency" or "microfinance" or "livelihood" or "export promotion" or "industrial policy" or "labor mobility" program potentially has many *design elements* and potentially many possible choices within each of those design elements. Simple combinatorics imply that the number of possible program designs can easily become very large (if there are 10 design elements with 3 choices each there are 1000 possible program designs). If outcomes are very different depending on program design then we say the fitness function/response surface/objective function is *rugged*.[4] Standard questions about "what works?" like "What is the magnitude of the impact of teacher training on student learning?" or "How much sustained increase in incomes results from participation in a livelihoods programs?" or "Will a job placement program help qualified new labor market entrants secure jobs?" are ill posed. They are ill posed because (a) *names* like "teacher training" or "livelihoods" or "microfinance" are names of *classes* of programs and only describe the *design space* and design spaces are high dimensional hence have very large numbers of possible *instances* of program design and (b) the fitness functions is sufficiently rugged over the typical development program design space that even rigorous evidence about the impact of one program design class contains little information about any other program design. Even if there

[3]Henceforth we use "program" but, unless otherwise specified, our argument applies to all three.

[4]We use the term "fitness function" rather than "response surface" or "objective function" although the former is the common term in the evaluation and economics literature and the latter the common term in the computer science literature. We use "fitness function" because our idea for simulating learning strategies on parametrically more or less rugged surfaces over high dimensional design spaces was influenced by the idea of "tunably rugged" fitness spaces using NK models a la Kaufmann and Weinberger (1989) and by a recent simulation paper in the medical literature using "clinical fitness"(Eppstein et al., 2012).

were "external validity" and the impact of exactly the same program were the same in all contexts, the question "Will *this particular* job placement agency succeed in placing applicants into formal sector jobs?" is not answered by the literature about job placement agencies, even a systematic review of rigorous studies about such agencies, unless either (1) the fitness function is "smooth" (impact is roughly invariant with respect to design features) or (2) the proposed particular job placement is exactly like the job placement agencies studied.

The devil may really be in the details. A "community development" program in Indonesia was widely regarded as successful and a "community development" program in Kenya was a spectacular disaster–and "community development" as a class of programs "community development" can produce pretty much anything ((Mansuri, 2013)). "Teacher training" can be both an integral component of program success or a nightmarish experience–and as a class of programs can produce pretty much anything. Even a replication of a rigorous evaluation ((, n.d.*a*) of a program that showed improved student learning from reducing class size with contract teachers *in the same country* produced completely different results when implemented by the government and not an NGO ((Bold, Kimenyi and Sandefur, 2013)).

Supposing the devil may be in the details and a new program is being designed. The existing literature–particularly the existing "systematic reviews"–are of granularity to guide program design. What is the right learning strategy? One possibility, the most common in practice (if a bit caricatured), is to design a program based on rigorous evidence, hunches, expertise, advice, intuition, experience, organizational fads, available funding, an informal assessment of the locally possible, and perhaps "piloting" and then implement that program with little or no scope for learning about its impact (much less providing

for rigorous learning). In response to program designs with very weak feedback or evaluation mechanisms there has been increasing pressure from funders (both large development agencies and foundations) to build a rigorous impact evaluation into program design.The Bank's own Independent Evaluation Group reported in 2012 that over 80 percent of the impact evaluations starting in 2007-10 used randomization, as compared with 57 percent in 2005-06 and only 19 percent in prior years. However, this pressure for rigorous impact evaluation often comes very early in the overall process of learning about program design. The question of the right *sequencing* of learning approaches for program design with a high dimensional design space and rugged fitness function (and possibly rugged *and* contextual) has yet to be addressed. There is a potential tradeoff between learning strategies that emphasize rigorous estimates on *impact* but which require early lock-in on program design and learning strategies that emphasize speed and flexibility in altering program design early on which full-blown RCT impact evaluations following a potentially extended period of exploration of the design space with quicker feedback loops.

We start with a concrete example of the trade-off in learning strategies from the experience of designing and implementing a job placement agency in Peru.

We then use a simulation model to compare two alternative learning strategies: (1) the standard randomized controlled trial for impact evaluation (RCT-IE) and (2) "crawling the design space" (CDS) (Pritchett, Samji and Hammer, 2013). The simulation attempts to capture the features of the job placement agency design and allows both the number of possible program designs and the ruggedness of the fitness space can be parametrically varied. The key difference is that over the learning period the RCT-IE explores relatively few program design options or "treatment arms" whereas the CDS strategy moves very quickly across the design space. This simulation shows that when the number of program

design options is even moderately large and the fitness space even modestly rugged the CDS learning strategy significantly outperforms the RCT-IE learning strategy as the "true" (in the simulated world) impact of the program design that emerges from CDS is typically much higher than the impact of the program design that emerges from the RCT-IE strategy. Moreover, the *variability* of the impact of the RCT-IE learning strategy program design (from repeated applications of the RCT-IE learning strategy to the same fitness function) is very high. These results are intuitive: if one evaluates the height of only a few local places starting from an arbitrary location in a rugged mountain range, the odds an evaluated height will be near the highest is small and repeated attempts will find both very low and very high heights.

We then argue that the existing results that have emerged from the application of the RCT-IE plus "systematic review" approach to social program design in developing countries suggest that, at least in many instances, the program design space is very high dimensional and the fitness function is very rugged (and contextual).

2 The Solution is the Problem?

One of us had an experience as a doctoral student that serves as an example, and was the motivation for, the two types of learning strategies we explore. As a doctoral student, Sara Nadel felt well-prepared with the ideal field research process:

1. Identify a problem. Be clear about what an ideal world looks like. Identify the anomaly in the market that prevents the ideal world from obtaining.
2. Write a causal model about the relationship between problem and a market failure that may be causing that problem.

3. Identify an intervention that could resolve the identified market failure.
4. Review existing literature to learn the evidence from previously-tested interventions.
5. Implement the program design with treatment and control groups. Compare the outcomes between the treatment and control populations.
6. Write the paper.
7. If results are positive, recommend scaling and replication of the program design.

While the above process is the recommended process for dissertation-writing in her field, it turns out it is very similar to the recommended processes for intervention design and evaluation of development programs in multi-lateral organizations, government, and non-governmental organizations. The best approach for resolving problems is to identify the problem, write a model, review (rigorous?) evidence about alternative interventions for resolving the problem, test, and, if successful, scale and replicate.

2.1 The Solution in Practice

Nadel set about to implement this process.

Identify a Problem: Through her experience living and working in Peru, expanded upon by followup field visits during her studies, she identified a problem: Peruvian youth from marginalized households despite their investments in higher education, were not securing formal sector jobs. This was in spite of rapid economic growth and the fact that formal sector firms complained they struggled to find appropriate talent.

Write a Causal Model: She hypothesized that the signal of higher education as discussed by Michael Spence (Spence, 1973) had broken down during Peru's massive expansion of

higher education offerings. According to the *Ministerio de Trabajo de Peru*, The number of people with higher education increased by 98 percent from 2001 - 2012, while the number of formal jobs increased by 38 percent. Talented youth from marginalized households had no effective mechanism to prove their skills and secure one of these increasingly competitive jobs even with some higher education.

2.1.1 Spence Model

[5] Spence identifies two groups with differing marginal products both in work and education:

Table 1: **Spence Model**

Group	Marginal Product	Proportion of population	Cost of education level y
1	1	q_1	y
2	2	$1 - q_1$	$y/2$

- Group 1: Education is costlier in terms of effort and productivity is lower.
- Group 2: Education requires half as much effort as Group 1, and productivity is twice as high.

In a world without a signal, an employer will presume that the productivity of each employee is the average of both, and pay accordingly:

$$q = q_1 * y_1 + (1 - q_1) * 2 * y_1 \tag{1}$$

[5] This table appears in Spence, (Spence, 1973)

However, employers may identify some optimal level of education, y^*, such that if $y < y^*$, the employer will know that productivity is 1 with probability 1, and if $y \geq y*$, the employer will know that productivity is 2 with probability 1. In this case, Group 1 will get $y = 0$, and Group 2 will get $y = y^*$.

2.1.2 Model of Signaling and Education in Peru

The hypothesis Nadel wished to study applied to how this model is complicated in the following conditions:

- Education is granular, not on a spectrum. Individuals either have a college degree or not.
- Credit constraints limit access to college education.
- There are more providers of college education and a decrease in the quality (non-financial cost) of education. In Peru, the number of college-age people pursuing a higher degree increased by 98 percent between 2002 - 2012.

In this environment, individuals make the following optimization decision:

$$Max \frac{q(y)}{\delta} - y - c, \; s.t. \; c \leq C \tag{2}$$

where c is the cost of higher education, and C is the maximum cost that an individual can pay. In this environment, she proposed to research the following hypothesis: *A reliable skill set signal* $\rightarrow$ *increased formal labor opportunities for people with* $C < c$. When $C \perp q$, the value of higher education as a signal of talent becomes nil.

Identify an Intervention: Identifying the model was easier than identifying the intervention. Spence's model identifies higher education as a skill set signal. If higher education no longer plays that role, the ideal intervention would offer a better skill set signal for applicants to firms. The existing research was useful only to a degree. Eventually, with the support of psychometricians experienced in our target population, we developed a test that would evaluate skill sets and preference sets consistent with dedication to work. We were eager to test its applicability.

Review the evidence: While there is a growing body of research about the characteristics of young adults from low-income backgrounds who succeed professionally (Rubinstein, Heckman et al., 2001), it was not clear that research focusing on the urban poor in the US would apply to the rural poor in Peru.

Table 2: **LogFrame of skill set Signaling to Improve Job Placements**

Activities take place inside the organization			Outside organization	
Inputs →	**Activities →**	**Outputs →**	**Outcome →**	**Impact**
	skill set Test	Signal	Firms hire differently	Increased productivity Increased youth employment
	Applicants must take test		*Firms must use test results*	height

Implement the Intervention: Nadel started a firm, *Farolito* ("little lantern" in Spanish) whose value proposition was to provide a more reliable signal about worker quality, allowing firms to hire higher-quality workers and pay them accordingly. However, the Farolito signal only becomes useful if the employer is able to attract high-quality workers. What the model above fails to consider is that the quality of jobs also varies, and job-seekers seek signals of the quality of a job when they choose where to apply, and try to optimize over the probability of landing a job and the long-term rents of having that job. This led to another model, that of the optimization of an applicant.

2.1.3 Application: Encouraging applicant turnout

Assume that each worker expects to work in perpetuity upon securing a job. This is unrealistic given the high level of turnover, but does not change the equilibrium decision-making as long as the length of time that a job-seeker expects to work does not vary by job. Job-seekers apply to jobs sequentially. Everything is priced at 1 other than wage and revenues. All characteristics are unique for each job, j, and the hiring entity must adjust the perception of the job j to make it more favorable to potential applicants. Variables in decision-making regarding application to job j:

P_j - Probability of successfully landing landing job j

W_j - Monthly wage at job j

e_j - Enjoyment of job j, which could include treatment of employees, cleanliness of facilities, likelihood of working late, etc.

S_j - Financial and time cost of applying to job j, including travel, printing resumes, childcare, etc.

$\bar{u}$ - outside option for predicted wage, time horizon and application costs of applicant. This could be an alternative job that the applicant has yet to secure, or a current job either outside the home or an informal family business.

The Job-seeker's optimization equations are:

$Max_j(P_j \cdot \frac{(W_j+e_j)}{\delta} - S_j)$, s.t

$P_j \cdot \frac{(W_j+e_j)}{\delta} - S_j \geq \bar{u}$

$P_j \cdot \frac{(W_j+e_j)}{\delta} - S_j \geq 0$

Given applicants' optimization process, the levers available to improve applicant turnout is to increase perceived P_j or e_j, or to decrease S_j. Farolito iterations tried to do all of these in addition to publicizing the job opportunity to more people in order to find more people for which the optimization equation support applying for the position in question.

Table 3 reviews adjustments made during the launch of Farolito with the goal of receiving more (high-quality) applicants and encouraging those applicants to complete the application process.

Small adjustments that were designed to lower the cost of applying by asking less of an applicant (planning ahead is a big non-financial cost for our talent pool) such as the amount of time between calling the applicant and the date of their test or interview or receiving applications by text message had big effects on turnout. Using the name of the client of Farolito (the ultimate emploer) instead of the name "Farolito" in publicity increased the perceived P_j or e_j because the job was perceived as a serious opportunity because our clients had better name recognition.

Other adjustments such as paying for Facebook advertisements, improved the applicant turnout and the quality of the candidates ultimately recommended for the job in some cases but not in others. For example, for the position in Chimbote, despite the extensive publicity, we received a total of 20 applicants. However, the same combination of publicity turned out many more candidates in Huanuco. This differential effect could be related to the population in each city (preferences and professional alternatives), suggesting that population preferences and outside options are a dimension that should be considered in the design space.

Table 3: **Learning at Farolito**

City	**Piura**	**Piura**	**Piura**	**Piura**	**Chiclayo**	**Chimbote**	**Arequipa**	**Huanuco**
Date of request	Aug 2012	Oct 2012	Feb 2013	Mar 2013	Mar 2013	Mar 2013	Mar 2013	May 2013
Announcement Mechanisms								
Newspaper	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Computrabajo*	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Bumeran*	No	NT	No	No	No	No	No	No
Facebook Page	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Paid Facebook Ad	No	No	No	No	Yes	Yes	No	Yes
Flyer	No	NT	No	No	Yes	Yes	No	Yes
University Career Counseling Centers	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other Operations								
Used Company Name (instead of Farolito)	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Invitation to test sent $<$ 24 hrs. beforehand	No	No	No	No	Yes	Yes	Yes	Yes
SMS reminder 12 hours before Test	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SMS applications accepted	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Schedule test online?	Yes	Yes	No	No	No	No	No	No
Success Measures								
Position viewings	NT	NT	783	128	236	20	119	NT
Filled out first filter	NT	NT	539	97	121	20	54	243
Met basic requirements	NT	NT	418	54	72	12	27	188
Finished Application	NT	NT	372	52	63	12	21	188
Took Test Invited/showed up	NT	NT	197	31	43	NT	18	152
Recommended (when based on test)	NT	NT	61	NT	11	NT	4	NT

*Common online job boards in Latin America
NT: Not Tracked

This type of learning was crucial to the Farolito product, although it had nothing to do with the original model that Nadel built or the problem she aimed to solve. If she had written a paper about the impact of providing better skill set signals to matching in the labor market, this background research would not have been included. Had the paper concluded that there is an opportunity to improve matching in the labor market through better skill set signaling without mentioning all of this background learning, practitioners seeking to replicate success would have to redo this (or simular) learning process over again.

Alternatively, had she not engaged in this learning process at all, she would have found limited impact of better signaling because she would have not had enough high-quality applicants with my partner organization to filter acceptably. Our efforts to attract users on both sides highlighted two characteristics about real-world implementation that my model was not prepared to incorporate:

- *Granularity and High Dimensionality* In trying to encourage users to adopt the test, it became necessary to revise aspects of the program that would be considered "insignificant" in the research context. Small iterations involved designing a better logo, creating a fancy information sheet, updating our website, and others. But while it was easy to see how "reliability of the signal" fit into the causal model, it was harder to see where specific adjustments fit into a model. How could she quantify a better logo or changing the logo color from blue to orange? The granularity of the interventions complicated their role in her model.

- *Ruggedness* Adjusting small characteristics of the program highlighted the number of small adjustments that can generate big changes in outcomes. There were hundreds of things to consider in an actual program, and the combination of those things generated drastically different responses.

The importance of such small characteristics that were seemingly irrelevant and hence not included in discussions either of the model or in the existing "evidence" but highly consequential to the implementation highlighted holes in the *construct validity* of the approach. That is, models and evidence often presumed that useful discussions could be had about *classes* of programs/projects/policies roughly independent of the consideration of the granularity in design that distinguished *instances* of the class. That is, I supposed that one could consider theoretically and empirically the *class* of "skill set signaling job placement programs" along the relative few dimensions identified in the theory (e.g. the reliability of the signal) and the relatively few instances of the class empirically examined. This experience with a failure of construct validity about classes of programs due to dimensionality and ruggedness was not unique, it is increasingly discussed in a variety of domains.

In a world where the design space is high-dimensional and where seemingly small design changes can have big impacts on outcomes, the step from *writing a causal model about the relationship between an outcome and the market failure that causes it* to *implementing a program to resolve that market failure based on existing evidence* is a much more complex process than the recommended steps for program design currently recognize. Hence the desire to build into program design a mechanism for learning about program efficacy.

3 Simulating the performance of alternative learning strategies

We are going to address the performance of different learning strategies in cases like Farolito in which we consider that the fitness function is rugged over a high dimensional design space. We do this by building an artificial world that abstractly represents some key features of the learning problem and examine results in this artificial world. The advantages of using simulation to explore learning stratgies are that the "truth" is known and we can

parametrically alter the world to examine how it affects the relative performance of the alternative learning strategies.

The simulation contrasts the performance of two alternative approaches to learning when facing a given fitness function.

The RCT-IE learning strategy:

- Starts at a random point in the design space.
- Implements that chosen program design as the base case and a local alternatives ("treatment arm") in one design space dimension over relatively long periods (say a "year").
- At the end of a year does statistical calculations and moves from the base case to the best alternative if the best alternative is statistically significantly better.
- In the second year the base case and a local alternative in a different design element direction are evaluated and again the RCT-IE moves to the new program design if the treatment arm is statistically significantly better than the treatment arm.
- The result of the RCT-IE learning strategy is the final program design.

The CDS ("crawling design space") learning strategy:

- Starts at the same random point in the design space as the RCT-IE strategy.
- Implements that program design and one other chosen randomly (with replacement) from the entire design space.

- At the end of one short period ("month"[6]) compares the outcomes and moves to the better of the two alternatives (with no account of statistical significance).
- Repeats this procedure for 24 months (two years).
- The result of the CDS learning is the final program design.

We can compare at the end of the two year period the performance of the RCT-IE and CDS final program designs to the best possible program design for the given fitness function.

3.1 Simulation: Design Space and Fitness Function for Farolito

The Job Placement Agency (JPA) objective is to find people who are a two sided match to JPA's contracting firm's available jobs. Each person has a job specific productivity or "aptness" for the particular job and one element of match is that the productivity is high enough the firm wants to retain the person. Each person also has a "hedonic" match to the particular job such that, if hired, he or she would choose to stay on the job. We define success for the JPA as placing people with the contracting firm who are above a threshold in both aptness and hedonic match.

Each month, we assume a pool of P people potentially interested in a job. The JPA program design problem is to *attract* people to its services and then use a *filter* to identify the two sided matches.

In order to be able to visualize the fitness function over the design space as a 3D graph we limit the JPA program design space to two dimensions: a communications (C) strategy to

[6] We say "month" and "year" in quotes as these are just arbitrary periods and any mapping from elements of the software code to elements of the world is allegorical but, to avoid pedantry, we will use terms like month and year to describe our artificial world without scare quotes each time.

attract applicants from P possible people and a filter (F) element that creates the signal of aptness for an employer. Each of C and F have N possible options and hence the design space has N squared elements.

3.1.1 Communications Strategy

This is simulated as a draw of NP people from a random distribution on aptness and a random draw on hedonics where we can control the correlation coefficient of aptness and hedonics.

C represents how we communicate with applicants in order to attract applicants. For instance, the variables that we adjusted, in order from lowest to highest density in terms of price, are:

1. Email & online only
2. Receive applications online, communicate by email and text message
3. Email, Text message + 1 phone call to invite to test
4. Email, Text message, phone call to invite to test, plus reminder text message

We represent a C design as a triplet: (c_1, c_2, c_3).

A person j from the P people applies to the JPA if:

$C_j = c_1 + c_2 * a_j + c_3 * h_j + \epsilon_{c,j} > C_{threshold}$

This is simple and intuitive. Communications strategies can either try to attract more people to apply (an increase in c_1) or try to induce high-ability people apply (an increase

in c_2) or try to induce people with a good hedonic match for the job to apply (an increase in c_3). It is obvious that there is a trade-off of different types of errors. Suppose the communications strategy, in a communications attempt to attract only high quality applicants, discouraged people whose aptness was in fact above the threshold. Then there are potential successes who are never seen and their exclusion from the interview process reduces the total number of successful placements. Conversely, if the communications strategy attracts many on the basis of hedonic match who do not meet the requirements, then for a given filter applied by the JPA, more "bad hires" would be made—people hired but were a mistake because they were not in fact highly apt or productive and hence would not increase success.

The application of the C strategy results in some proportion of the pool applying to the JPA.

3.1.2 Filter Strategy

F strategy represents how we filter applicants. Computerization adds a level of difficulty among our job applicant base. As such, in order from lowest to highest density, the variables are:

1. Group interview
2. Handwritten test
3. Online filter which confirms applicant meets basic job requirements, completed at home
4. Computerized test given in a supervised environment

The second element of the JPA program design is the application of the filter, which is two sided. On one side, the application of some assessment produces an estimate of the aptness and hence filters out of the application process those applicants below the estimated aptness threshold but it is also the case that applicants may choose to drop out of the recruitment process as a result of the experience of the filter. Hence the F strategy is (f_1, f_2).

A person is considered hired if:

$A_j = f_1 * a_j + \epsilon_{a,j} > A_{threshold}$ (the person is estimated by the filter to be above the critical threshold on aptness)

and

$H_j = f_2 * h_j + \epsilon_{h,j} > H_{threshold}$ (the person, even after exposure to the JPA filter process, still wants the job).

The application of the filter F to the applicants produces some number (perhaps zero) of the P pool of possible applicants in a given month who are hired.

The (expected) number of successes in a given month from a given strategy is the number of hires whose actual aptness and hedonic match are above the threshold, as this implies they will be hires who both the firms desires to stay and who desire to stay.

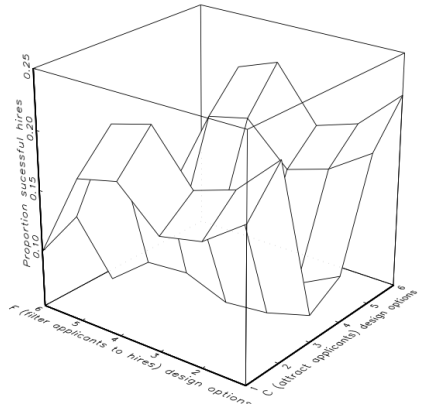
3.1.3 Fitness Function

We design our artificial world to have a fitness function over the design space which is *rugged* in a manner that is parametrically controlled. The ruggedness has three elements illustrated in the examples in Figure 1. First, the fitness function is not linear (or quadratic) in one strategy conditional on the other. Second, the fitness function is interactive, the

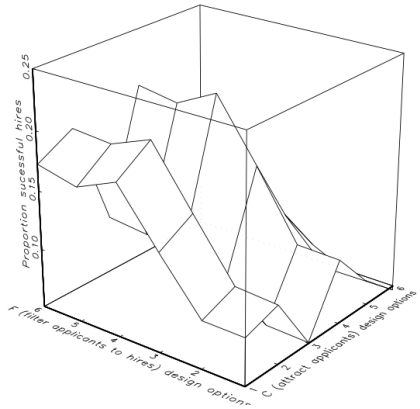
relative outcome of F_j versus F_k for C_j is not (necessarily) the same as for C_k. Third, these outcomes differences across (C,F) strategies can be parametrically made bigger or smaller in the relevant fitness metric even across alternatives that are local in the design space.[7] There is an optimal strategy of communications and filter (C^*, F^*) but proximity in the design space to the (C^*, F^*) combination does not ensure proximity to the optimal outcome.

[7]The notion of "local" in the design space is problematic as many program design options are discrete alternatives rather than alternatives that can be cardinally ranked. With multiple mutually exclusive discrete alternatives there is no natural metric or definition of "local" in the design space.

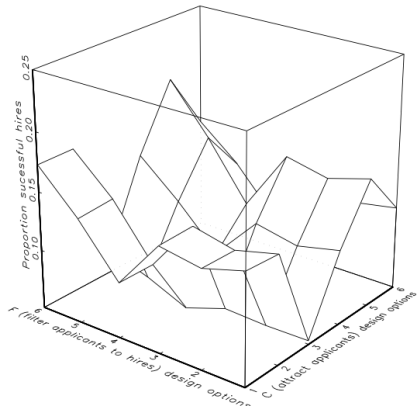
Figure 1: **Three Examples of Rugged Fitness Function**



(a) Fitness Function Example 1



(b) Fitness Function Example 2



(c) Fitness Function Example 3

A fitness function is completely described by the parameters of the communications and filter strategies.

For C the first strategy was always $C_1 = (0.5, 0, 0)$ which was the default strategy that attracted applicants uncorrelated with either aptness ($c_2 = 0$) or match ($c_3 = 0$). For the remaining strategies the elements were chosen randomly from the possibility set [-0.5, 0, 0.2, 0.5 and 1]. For example, a C design triplet (0.2, -0.5, 1) would attract fewer default applicants than the base strategy but attract applicants in inverse correlation with aptness but positively selected on hedonic match. This 3 by N communications strategy matrix was then inflated or deflated by a ruggedness parameter.

The elements of an F design were chosen from the possible values 0, 0.5,1,1.5 or 2 randomly for each element for each of the N strategies for C and hence the F parameters are different for each C strategy. For instance, an F design of (1.5, 0) implies the filter process would strongly select on aptness and dropout from the job application process would be uncorrelated with match at the filter stage.

These choices fix the fitness space as they determine of the population who is attracted to apply for the job via the communications strategy, who is offered the job as a result of the filter and who takes the job as a result of hedonic match.

Success is determined by whether those who are offered and take the job are *in fact* good job matches. That is, our measure of JPA success is the fraction of the population who are place and are truly above the aptness and hedonic match thresholds. That is, the filter on aptness can be "good" or "bad" in selecting on true aptness. Communications and filter strategies have differential performance because they can make mistakes of various types–they can attract too few applicants but successfully filter those who do apply but result in too few matches relative to the optimal, they can attract lots of applicants but

of the wrong sort (e.g. the communications process can attract ill-suited applicants on aptness or hedonic match) and then, for a given filter to many of the wrong people will be fired.

3.2 Simulation: Learning Strategies in the Artificial World

We have built this simulation to examine questions about the dynamic process of learning about program design. Suppose we were facing a rugged fitness function over a high dimensional design where the fitness function is contextual (or at the very least we do not know *ex ante* if it is contextual or not). That is, the efficacy of a particular (C,F) design might be different in Detroit versus New Delhi versus Cairo and hence cannot be assumed to apply to Peru (or to different cities in Peru). Moreover, existing evidence about the success of job placement agency programs may lack *construct validity* as it lacks information about all of the elements of the design space. What is an appropriate learning strategy as a sequence of actions and feedback loops from those actions that would be likely to lead over a fixed period to a good program design—a combination of C and F that produces a high (if not optimal) outcome? We assume the fitness function over the design space is fixed over the period of the simulation but unknown for a "context" (where context can include country, region, implementing organization, availability of other alternatives, etc.). We are going to simulate a period of 24 months with observed feedback on job placement outcomes at the end of each month.[8] and apply two different learning strategies (CDS and RCT-IE). Each of the two learning strategies starts at the same point in the design space and then, relying on the feedback from outcomes, dynamically alters

[8]This involves a modest elision on the actual dynamics as we assume at the end of each month we observe not just the actual placements made by the (C,F) strategy but also the *successful* placements. This essentially assumes that the "true" aptness and hedonic match are revealed instantaneously after the job has begun. In reality it would take some time for this to be revealed.

the (C, F) strategy being pursued. We then compare the strategies at the end of the period to see which was better at learning, on average, over a variety of possible fitness spaces.

3.2.1 Learning Strategy: CDS (Crawl Design Space)

Both CDS and RCT start at a given program design, (C_0, F_0) (where 0 indexes time not strategy number).

In the CDS learning strategy in each period two strategies are implemented: the current best and one alternative. The alternative is chosen each period from all other strategies besides the current best. Sampling on alternative strategies is with replacement so previously tried strategies can be tried again. At the end of each period t the outcomes of successful hires from the two strategies are compared as simple averages (which is the same as a count of successful hires, since the number of potentials is fixed at NP) and:

If:

$$FF(C_{CurrentBest,t}, F_{CurrentBest,t} > FF(C_{Alternative,t}, F_{Alternative,t})$$

then the current best program is retained:

$$(C_{t+1}, F_{t+1}) = (C_{CurrentBest,t}, F_{CurrentBest,t})$$

if the alternative is better then it is adopted and:

$$(C_{t+1}, F_{t+1}) = (C_{Alternative,t}, F_{Alternative,t})$$

At the end 24 months the resulting program design is:

$$(C_{CDS,T}, F_{CDS,T})$$

and hence the outcome of implementing that program design is:

$FF(C_{CDS,T}, F_{CDS,T})$.

We acknowledge this is an extremely simplistic learning strategy. There is no optimal choosing of the starting point, no attempting to guess what a good "alternative" strategy might be, no use of "statistical significance" to choose whether to switch, no memory to the learning process (e.g. if a program design is outperformed in one month it is replaced even if it has worked for months against other alternatives). We intend this to be a simple search/learning strategy as we are not searching for the "optimal" learning strategy.[9] rather we are attempting to articulate a simple learning strategy that pretty much any organization could implement.

3.2.2 Learning Strategy: RCT-IE

The learning strategy for RCT-IE in this artificial world is intended to mimic the standard RCT which chooses a "treatment" and perhaps a few alternative "treatment arms" and then holds the treatment fixed for a period sufficient to generate statistically significant results and hence shifts treatments relatively infrequently.

We model this by having the RCT-IE start from exactly the same starting point as CDS. This default or "current best" program design is tested against one other alternative, which is a local alternative[10] that alters the strategy in just one dimension. In our simulation we first search in the C dimension and then in the F dimension.

[9] We are reasonably confident from the large literature on optimization there is no generically "optimal" algorithm for finding the optimum of an arbitrary rugged fitness function.

[10] Since we do not assume the design space alternatives are ordered in any way we treat "locality" as a cycle so that alternative 1 is "local" to both alternative 2 and alternative N, where N is the number of possible choices for either C or F designs.

The principal difference is that program modifications occur only at the end of the year and not each month. This is to (as generally claimed) preserve the "statistical power" and to "maintain the integrity of the experiment." At the end of 12 periods the "current best" (which is the initial at the end of period 12) is compared to the alternative. The alternative is adopted only if it is statistically significantly better than the "current best." Then, having adopted the new strategy for period $t + 13$ and on the alternative is chosen as a local alternative in the F dimension.

Then, at the end of 12 more periods (and hence the end of the first two years of implementation) the current best is compared to the alternative and the alternative is adopted if it is statistically significantly better than the current best.

The result is an RCT-IE strategy and outcome:

$(C_{RCT,T}, F_{RCT,T})$

and hence the outcome of implementing that program design is:

$FF(C_{RCT,T}, F_{RCT,T})$.

The key differences between the learning strategies:

1. CDS learning updates the default program design upon any superior outcome, while RCT-IE learning updates only on a statistically superior outcome.
2. CDS updates once a month compared to RCT-IE updating once a year. The RCT-IE measurements are more precise, with lower variance and less likely to be influenced by noise.
3. CDS learning chooses a program design variant from the universe of strategies. RCT-IE learning only adjusts in only one design element direction, first C then F.

To the objection that this simulation is cooked in favor of the CDS over the RCT-IE strategy there is a three-fold response. First, we think that the description of program design changing at best once a year once the RCT has been initiated is not a complete caricature. We personally have been acquainted and/or directly involved with a several RCT-IE studies in which a treatment arm was obviously badly failing but the experimenters insisted on continuing with implementation "as is."[11] Second, only two treatment arms is not uncommon as Vivalt (2016) finds an average of 1.74 treatments per evaluation. The combination of cost, statistical power with potentially noisy measurement and modest anticipated impact size tend to limit treatment arms to a small integer. Third, given the attention and resources RCT-IE have received recently as a method for learning about "what works" in development, if it is really so easy to created simulations in which they under perform as a learning tool this is in itself revealing.

3.3 Mechanics of the Simulation

The basic structure of the simulation is:

Step I: A fitness function is created as a random choice over the possibilities for the five parameters (c_1, c_2, c_3) and (f_1, f_2)) and a ruggedness parameter.

Step II: Each month a universe of NP (where NP is 1000) individuals are exposed to the C strategy and choose whether to apply to the JPA. The JPA then filters applicants to create job candidates. Afterwards, the candidates, having experienced the filter process, choose whether to take the job.

[11] One of my (Lant Pritchett) first trips to an RCT that was a collaboration of an NGO and an academic partner the head of the NGO introduced me to junior worker from the RCT partner saying "This is Dan, his job is to make sure we don't help any children" as the academic partner was encouraging them to stick to a treatment arm the implementing NGO had recognised as flawed and wished to abandon.

Step III: The number of successful hires at the end of each month is known.

Step IV: Based on the results the (C,F) strategy is updated according to the rules of CDS (monthly) or RCT (yearly).

Step V: At the end of the T periods (where T is 24 months) the results of the final strategy of the two learning strategies are compared for the given fitness function (that is, the success of $(C_{CDS,T}, F_{CDS,T})$ and $(C_{RCT,T}, F_{RCT,T})$ are compared when applied over data for all 24 periods as an approximation of the "true" results as the monthly results depend on the random variances in the various selection functions).

Step VI: After this process is repeated for I iterations, the results for having applied the CDS and RCT strategies over a large number of different fitness functions of given ruggedness is known and the average and variance of outcomes of the strategies can be computed.

4 Results of the Simulation

The purpose of the simulation exercise is to ask: "In an artificial world with a large design space and an unknown but potentially rugged fitness function, what are the implications of various learning strategies for program design?" The results show the RCT-IE strategy is low mean (the learning gain is smaller) and high variance (the risk of getting really poor results is larger) compared to CDS.

4.1 Baseline Results

These simulations produce two main results, illustrated in Table 4 using design spaces from 5 to 10 options for each of the two design elements C and F and hence a design space of only 25 to 100 program designs.

First, crawling the design space (CDS) typically reaches a substantially better program design than RCT-IE.

The second column of Table 4 shows the average over 1000 iterations on different fitness functions and different starting points of the excess performance of CDS over RCT-IE scaled as the ratio of the gap between the "best" and the "average" for each fitness function. Since the absolute success rates are more or less arbitrary, (e.g. we can produce more or less average success by varying the thresholds) we feel this is a natural metric for the gain from learning: *how much of the distance between having just having picked a program design at random (which would produce on average the average result) and having reached the best possible result was closed by the learning one did?* For 6 options for each design element (which is our default) the CDS program design produces 49 percent better outcome than RCT-IE tahn the best program design to average program design gap.

To illustrate the intuition with absolute numbers, in the baseline parameters and 6 design options for each element the average success is 10.2 percent. The best result in each fitness function averaged over 1000 fitness functions is that 15.7 percent successful hires. The average CDS result is 14.9 percent, so it nearly reaches the best result of 15.7 and makes substantial improvement over its starting point which, since it is randomly chosen is the average result of 10.2 percent. CDS learning improves success on average from from 10.2 percent to 14.9 percent. The average successful hire rate for the RCT-IE learning strategy

across the 1000 different fitness functions is 12.2 percent. Hence the gain of CDS over RCT-IE is: (14.9-12.2)/(15.7-10.2)=0.49.

Interestingly, the learning gain of CDS relative to RCT gets somewhat smaller as the number of options (hence the dimensionality of the design space) gets larger. This is because the RCT strategy does about the same in gain but the gain of the CDS as a proportion of the possible gain (best over average) gets lower as the ratio of the 24 trials to the total design space gets smaller.

Table 4: **Simulation Results**

(1) Number of options	(2) Gain CDS over RCT to max over average possible*	(3) Percent excess of RCT standard deviation**
5	0.516	1.57
6	0.490	1.64
7	0.487	1.65
10	0.467	1.74

*$((C_{CDS,T}, F_{CDS,T}) - (C_{RCT,T}, F_{RCT,T}))/(BestPossibleSuccessfulHires - MeanSuccessfulHires)$
**(Standard Deviation of RCT Results / Standard Deviation of CDS Results)

The second main result is that CDS has a lower variance across alternative fitness functions than does the RCT-IE. Column 3 of Table 4 shows that the ratio of the RCT to CDS standard deviation is 1.57 times higher with a design space of 25 options and 1.74 times higher when there are 100 options. The intuition in the simulation is that since the RCT has fewer moves across the surface of the fitness function if it happens to get started with a bad program design (by random chance) this happens to be in a bad neighborhood of the fitness function since it only has two possible local moves it could end up with a very poor outcome. At the same time the random initial program design could be close to a

good one. For instance, in a run of 1000 simulations of the baseline parameters and six options, the 10th percentile result for the RCT-IE was 10.6 percent–which is substantially worse than even the average result.

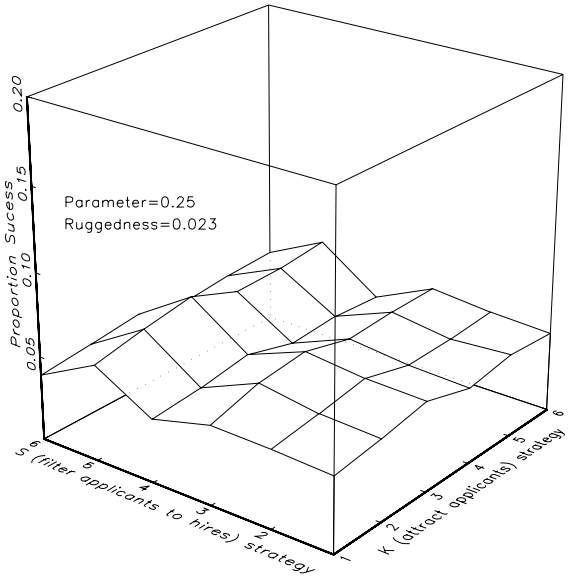
4.2 Performance of Learning Strategies across Degrees of Ruggedness

In our artificial world we can parametrically alter the ruggedness of the fitness function. Figure 2 shows typical fitness spaces when, relative to the baseline ruggedness parameter of 1 (which produces the graphs in Figure 1) the ruggedness parameter is either 0.25 (producing the smoother surface) or 4 (producing a more rugged surface). We can compute the actual ruggedness as the average absolute deviation of fitness at each element of the design space compared to its eight local neighbors.[12]

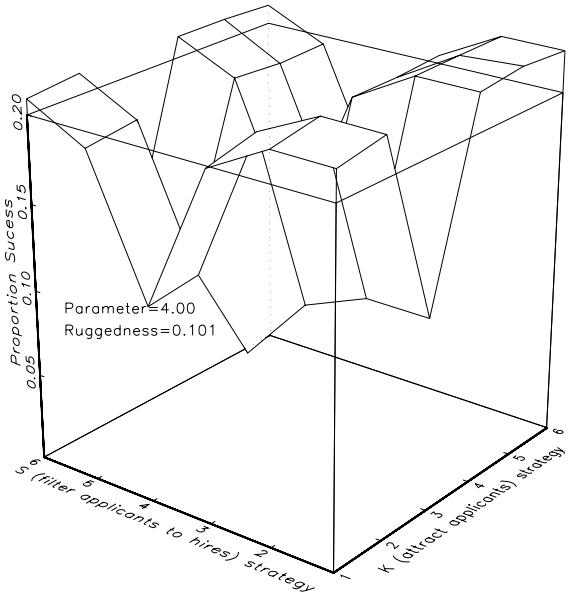
Table 5 shows the results of holding all our parameters as in Table 4 for six options and then varying only the ruggedness parameter. As designed, from the lowest to highest values the ruggedness increases by a factor of five from .020 to .103. The striking, even if expected, result is that the ratio of the standard deviation of the RCT strategy to the CDS strategy increases from roughly 1 (they do about the same) for the smoothest fitness functions to 4.25 for the most rugged. Practically (if a simulation result can be said to be practical) this says: if the fitness function is very rugged then design matters a great deal and hence the losses from not crawling the design space can be vary large and the results of an RCT-IE can be good (if one happens to start at or near a good alternative) or very bad (if one happens to start in a bad program design neighborhood).

[12]This again assumes that since the design space isn't ordered the space "wraps around" and hence the local program design results can be computed by tiling the design spaces side by side. Hence the eight local neighbors of design space point (1,1) are (clockwise from northwest): (6,6),(6,1),(6,2),(1,2), (2,2), (2,1),(2,6),(1,6).

Figure 2: **Comparing smoother and more rugged fitness functions**



(a) Smooth Fitness Space



(b) Rugged Fitness Space

Table 5: **Learning results varied across ruggedness of the fitness space**

(1) Ruggedness parameter	(2) Ruggedness (absolute difference)	(3) Gain CDS over RCT (ratio to max less average)	(4) Percent excess of RCT over CDS standard deviation
.25	.020	.319	1.04
.5	.042	.445	1.19
1 (base case)	.074	.489	1.64
2	.094	.461	2.36
4	.103	.412	4.25

Interestingly, the average gain of CDS over RCT-IE as a ratio to best less average first grows with ruggedness and then declines.

4.3 Other Variations on the Base Case

The results presented in Table 4 just varied the simulation across the number of program design options but kept all other parameters of the simulation constant. It is possible that the two main results are fragile with respect to minor variants in the artificial world. In this section we vary three elements of the simulation to examine the robustness.

Correlation of aptness and hedonic fit. In the base case we assumed that there was no correlation between a person's aptness in a job and their preferences to be in the job. Since the program designs often affect the two characteristics of the pool of people differently (e.g. communications strategies may attract the apt but deter well-matched or vice versa) this could affect the results. In Table 6, the first row are the results for the base case parameters with 6 options for each element of program design (36 total designs). In the second row the correlation of aptness and match was increased to 0.8. This produced roughly similar results with modestly higher average success and even more difference in the RCT-IE and CDS variation.

More noise in decision rules. One feature of the simulation is the extent to which the program design versus randomness (including the inability of program design to effectively target the "right" applicants through communications or filter the most apt applicants with an instrument) affects outcomes. We add more noise (via the individual specific ϵ's in the equations) to the process. The result is that the learning advantage of the CDS over RCT-IE persists but is lower and the relative variability of RCT-IE versus CDS final program design outcomes declines. This is intuitive as it increase the uncertainty of identifying

precisely the program design versus random elements month to month and as the CDS can crawl away from a good design due to noise.

Higher thresholds of success. In the base case the thresholds for aptness and hedonic match ($A_{threshold}$ and $H_{threshold}$) that define successful placement were set to zero. If we increase that to 1 (since these are Normal distributions the threshold rises from 50th to 16th percentile of population) this reduces the success rate from .102 to .025 on average. This produces roughly the same learning gain from CDS over RCT-IE but also dramatically increases the RCT variability as it increases the gap between successful and failing program designs which increases the risk the RCT-IE ends at a relatively low performance final design.

Table 6: **Variations on the Base Case**

(1) Number of options	(2) Gain CDS over RCT to max over average* possible	(3) Percent excess of RCT standard deviation**	(4) Average Success***	Description of parameter changes
6	0.490	1.64	0.102	Base case parameters
6	0.504	1.84	0.163	Correlation of aptness and match 0.8 (instead of zero)
6	0.371	1.07	0.057	Variance of the noise in decision rules increased
6	0.457	4.99	0.025	Higher threshold in ability and hedonic match for success

*$((C_{CDS,T}, F_{CDS,T}) - (C_{RCT,T}, F_{RCT,T}))/(BestPossibleSuccessfulHires - MeanSuccessfulHires)$
**(Standard Deviation of RCT Results / Standard Deviation of CDS Results)
***(% successful hires from a population of 1000

5 Is the Fitness Function for Social Programs Rugged? Evidence about What "Evidence" Means

We show that in an artificial world in which the design space has even a moderate number of program designs and outcomes are sensitive to program design (rugged fitness function) a fast turnaround learning procedure dominates a more precise and statistically reliable but slower and more local learning process in both mean gain and in the variance of expected outcomes across contexts. This finding about an artificial world is obviously of little interest unless it captures at least features of the "real" world of development program/policy/projects.

The next two sections argue that: (a) the available evidence from the cumulative RCTs suggest that not only are there the well-known issues of *external validity* (that the fitness function may vary across contexts) but that there are also major issues of *construct validity* in the *classes* of policy/program/projects at which there are attempts to summarize what the evidence says about "what works" are insufficiently granular to be of use to practitioners and (b) to the extent that the new concerns about behavioral economics are important, construct validity in development projects/policies/programs becomes nearly impossible.

5.1 Heterogeneity in Estimated Impacts: External Validity and Construct Validity

The simulation result that the RCT-IE learning strategy has a variance increasing in the ruggedness of the fitness function highlights the potential role of "construct validity" in assessing empirical evidence. In nearly all domains in which there have been sufficient evaluations (e.g. microfinance, business training, income generation/livelihoods, community development) and even in sub-domains (e.g. within primary education the impact on

learning of "teaching training" or "class size reduction" or "ICT") one finds large differences in the estimated impacts across the evidence. While it is impossible disentangle the two, many of these differences appear to be construct validity, not purely external validity issues due to "context."[13]

5.2 Examples of program ruggedness from impact evaluations

5.2.1 *Impacts of programs on student learning*

The domain in which there have been perhaps the most RCTs has been examining impacts of various projects/programs/policies on child learning of reading and arithmetic and hence all of the issues are now illustrated.

(Muralidharan and Glewwe, 2015) show there are now (at least) four well-identified estimates of the impact of providing additional textbooks in low income, low textbook availability settings. All four studies show zero impact of the provision of additional textbooks on the learning performance of the typical child. A mechanically implemented "systematic review" of the "high quality" would justifiably conclude that there was no evidence that textbook availability mattered for child learning. But, as Muralidharan and Glewwe (2015) emphasize, each of the studies point to elements of the design space that *interact* with the impact of textbooks (note these are nearly all detected *after* the study is completed and the puzzling empirical finding of no impact emerges, not before). (Glewwe, Kremer and Moulin, 2009) conclude the limited impact on the average fourth grade Kenyan student is because the textbooks were too hard, an element of the program design space that was not

[13]Where "context" is itself under-specified as people think of "country" or "place" as a catch-all for "context" but "context" could well be (and has been shown to be in some applications) regional, organizational, personal, path-dependence on history, existing alternatives to suppliers, etc.).

varied during the RCT-IE. [14] Other studies have other *ex post* evidence supported explanations of the lack of impact: teachers didn't actually open the textbooks but stored them (Sabarwal, Evans and Marshak, 2014); when the provision of textbooks was anticipated it mostly crowded out household purchase (Das et al., 2011); or the textbooks only had positive impact when interacted with teacher performance pay (Mbiti, Muralidharan and Schipper, 2015). In even a seemingly straightforward class of programs called "provision of textbooks" design matters as elements of program design are interactive.

McEwan (McEwan, 2015) reviews the rigorous evidence about impacts on learning in primary schools of developing countries and compares the average impact across 11 classes of interventions like "Computers or technology" or "Instructional materials" or "Deworming drugs." The average effect size (gains as ratio to student standard deviation) 0.072, and the standard deviation across these 11 classes of interventions is 0.05 and the range is 0.161 (from 0.15 to -0.011) and the highest average impact size is for "computers or technology" with an average effect size of 0.15. However, if one looks across the 32 "computer or technology" interventions the range is more than 1 full standard deviation from positive 0.45 to *negative* 0.58–the "within class" range of instances within the class is six times the range across averages of classes of programs. While this might be chalked up to "external validity" or cross-contextual heterogeneity in impacts, some of the largest differences are across treatment arms in the same context. So for instance, the same study in India found that "after school" computer-assisted instruction had an effect size near 0.30 whereas "in school" computer assisted instruction *reduced* learning by 0.58 standard deviations–an across treatment arm range of 0.88–more than 5 times the range across all classes of interventions. This suggests a very rugged fitness function for "computer or tech-

[14] As our paper is about the *dynamics* of learning about prorgram design it is worth pointing out this Kenya study was initiated in 1995, with textbooks provided in 1996 and 1997. The working paper (NBER) version appeared in 2007 and the published version in 2009.

nology" learning interventions and makes a comparison of mean effects of "computers or technology" at 0.15 versus "contract or volunteer teachers" at 0.101 seems inconsequential relative to the within class impacts of differential design.

Evans and Popova (2015) review six "systematic reviews" of the literature on how to improve learning in basic education in developing countries and show that the "systematic reviews" of the "rigorous evidence"–even of the same topic (and two were by the same organization) often come to very different conclusions. This is not surprising when the within class of intervention variance is high relative to the across class mean differences and in that case small changes in the methods and filters to search for and include studies can lead to different results as the heterogeneity in impact estimates is so large across studies.

In very recent work Evans, Popova and Arancibia (2016) examine the evidence on the impact of "teacher training" and find large variation in impacts across programs. However, when they went to dig deeper and ask "what works" on program design within the class of "teacher training" they found that the published studies to a large extent did not report in sufficient granularity to investigate this. That is, they created a design space characterization of 43 elements (with indicators for instance for who implemented the training, did the training have professional implications, what was training in (e.g. content or pedagogy), what were the activities undertaken in the training) which, even at 43 elements, is rudimentary. They then looked at 23 rigorous evaluations of teacher training and found that the typical study only had 50 percent of the indicators. With follow up with the authors they could find the information on design and even with relatively few observations (26 distinct programs in 23 studies) they could show the reported program impact of "teacher training" was associated with these design characteristics–but in ways impossible for any

standard "systematic review" to have uncovered.

5.2.2 *Livelihoods*

There are a variety of programs that transfer assets to poor individuals in an attempt to achieve sustained increases in incomes. In India, the central government has supported states in launching livelihoods programs that create and support women's self-help groups as an instrument to women's empowerment and income gains. An evaluation of the Jeevika program in Bihar India in the early implementation phase showed phenomenally large improvements measures of women's autonomy of action in Jeevika villages. However the randomized evaluation of the program scaled statewide showed impacts on these same measures that were up to an order of magnitude smaller, often not statistically significant. In this case even the same program design failed to produce the same results when the intensity and integrity of implementation was lessened.

Recently an experimental evaluation (Banerjee et al., 2015) showed success in an income generation program in five of six countries, suggesting that it is possible to achieve similar results across contexts. Yet in some respects this study demonstrates just how hard it is to achieve similar results as at all sites the intervention had the same six elements (an asset transfer (e.g. livestock), training on managing the asset, food or cash support, frequent coaching visits, health education/access, a savings account and at all sites and was implemented by exactly the same NGO. This was the scaling-across-contexts of an approach that had devoted effort to learning the need and how to, combined with the various dimensions of the design space before moving to an RCT and controlled the implementation afterward.

5.2.3 *Electricity Consumption*

Allcott and Mullainathan (2012) review a program designed to encourage households to consume less electricity by providing them with information about how their electricity consumption compares with that of their neighbors. The program was implemented by one organization, OPower, in partnership with 14 different utilities across the US. The authors find significant heterogeneity in treatment impact – with a cost-effectiveness of the program ranging from 1.66 cents per kilowatt-hour to 5.82 cents per kilowatt-hour – across implementing utilities. Some utility characteristics, such as whether the utility is investor owned, are correlated both with treatment effects *and* with takeup of treatment. Thus, the documented examples of differing external validity, such as those reviewed in systematic reviews, do not capture the entire range of differential impact of programs. The authors note that the OPower implementation was a particularly attractive one for research about external validity because of the minimal variation in program design across utilities.

Even with minimal program design differences, the program construct was not identical: the frequency of treatment varied between monthly, bi-monthly, and quarterly across implementing agencies; more frequent treatment is significantly correlated with a bigger treatment effect. The paper demonstrates the danger of failing to consider both construct and external validity in learning.

5.2.4 General Reviews of RCTs

Eva Vivalt (Vivalt, 2016) founded an NGO, AidGrade, that has systematically collected results of over 600 RCTs and has attempted to use that data to ask almost exactly the empirical counter-part of our simulation exercise: how much do RCT results for the same class

of intervention differ across studies? This combines all sources of variability across studies: external validity, construct validity, sample variability, and others. She first calculates a standardized impact estimate:

$$SMD = (\mu_{treatment} - \mu_{control})/\sigma_{pooled}$$

She matches intervention and outcome pairs (e.g. impact of a Conditional Cash Transfer on the enrollment rate) and calculates for each study the intervention impact on outcome. She can then calculate across a class of interventions two measures of the reliability of the impacts: the coefficient of variation and the I^2. The I^2 measure for meta-analysis was introduced by (Higgins and Thompson, 2002) and "describes the proportion of the total variation in study estimates that is due to heterogeneity." Table 7, adapted from (Vivalt, 2016) shows the results for three intervention-outcome pairs with relatively large numbers of studies and the median across all the 51 intervention-outcome pairs with sufficient studies for these estimate to be computed.

The basic finding is that the typical (median) coefficient of variation of impact estimates is 1.77. This implies that if, suppose, a program class of RCT studies showed, on average, a massive impact of an effect size of 0.5 then the standard deviation across studies would be 0.885 and hence the once standard deviation confidence interval of the impact estimated in the next study would range from 1.38 to *negative* 0.385. In other words, if the assumption was the impact of the intervention was to be positive the existing results would be roughly uninformative about the likely result: the plausible range based on the mean and variance of RCT evidence roughly spans the entire plausible range of outcomes as it ranges from very negative to implausibly large. (Vivalt, 2016) reports the typical CV in the medical literature is 0.05 to 0.5. (Higgins and Thompson, 2002) suggests that for a meta-analysis an I^2 of 0.25 is low and 0.75 is high and obviously 1.0 is the maximum so the data suggest

the heterogeneity in development program RCTs is massive.

Moreover, much of this variation in estimated impacts was "within papers"–that is across different groups or treatment arms (inclusive of differences due to sampling variability) of the same paper estimating the same impact in the same intervention-outcome pair. This is at least suggestive that the traditional interpretation of heterogeneity due to "external validity" caused by "context" is incomplete.

We would argue that these degrees of heterogeneity suggest that the classes of interventions that are often discussed in reviews of evidence about policies/programs/projects exist in a world that is sufficiently "rugged" that they lack construct and external validity hence are roughly meaningless–it is not clear an "HIV/AIDS Education project" (or most of the other intervention-outcome pairs Vivalt considers), without further specification of the program design and context, could be the object of meaningful empirical discussion as the evidence does not predict future outcomes. And yet, Vivalt (2016) reports that one in five studies failed to even report who was responsible for implementation.

Table 7: **Variability across RCT studies for intervention-outcome pairs**

(1) Intervention	(2) Outcome	(3) CV(SMD_i)	(4) Within paper CV	(5) I^2	(6) Number studies
Conditional Cash Transfers	Enrollment Rate	0.83	0.968	1.00	37
HIV/AIDS Education	Use of contraception	3.12	6.97	0.51	10
Micronutrients	Hemoglobin	1.44	0.731	1.00	46
Median (51 intervention/outcome pairs)		**1.77**		**0.99**	**7 (per pair)**

Source: (Vivalt, 2016), Appendix C, Table 12.

The fact of construct validity due to variability in program efficacy due to details of program design causes a further problem through the "systematic review" of classes of programs. Systematic Reviews weight different studies and come to general conclusions, which is even more problematic than one study with external validity. This is particularly true because (a) there often isn't a coherent and common way of classifying treatments, (b) key elements of program design are not often even mentioned and (c) the process of coming to the particular treatment arms which are part of RCT-IE is often completely undocumented. That is, often researchers will do extensive informal "piloting" or work with an organization informally as part of "field work" to choose a project design. There is no ex post comparability of RCT-IE results if two studies differed dramatically in their pre program design stage. This leads to problems in which the reported impact of the RCT-IE is actually the result of the *combined* process of getting to program design, program design optimization, plus RCT-IE finding.

5.3 Behavioral approaches and ruggedness

As the advent of the increased use of RCTs in development and the increased popularity of behavioral economics came roughly at the same time it is worth pointing out that one of the key features of behavioral economics is that it often suggests large impacts across small differences in design where standard theories suggest little or no impact. Perhaps the classic behavioral finding is that whether the default is that a person is signed up for retirement deductions and must opt out or the person has to actively opt in to retirement deductions has a huge impact on participation and hence the "power of suggestion" (Madrian and Shea, 2001) created by the program design feature of the default selection has a much larger role on behavior than standard economic theory would have suggested. Another

cited example from RCTs is large changes in demand for health and education services at zero monetary cost pricing (Holla and Kremer, 2009) which is inconsistent with standard economic theory that zero monetary cost is just another point on the demand curve.

Many of these behavioral models imply that the fitness space over program design is rugged and in ways that are contextually hard to predict *ex ante*. For example, Gino et al. (Gino, Ayal and Ariely, 2009) demonstrate that Carnegie Mellon students are more likely to cheat at a task when someone observed cheating at the task is wearing a plain shirt and thus assumed to also be a Carnegie Mellon student, but when the cheater is wearing a University of Pittsburgh t-shirt (Carnegie Mellon's cross-town rival), it does not affect cheating by other students. This study suggests that cheating decisions are norm-driven and norm-driven behavior is driven by identity affinity to the norm violater. But how would one translate that into policy about cheating on property taxes in Pakistan or teacher attendance in Kenya? In another example, (Bertrand et al., 2010) find that when an advertising mailer includes a photo of someone of the same race the effect of the mailer on loan take-up is nearly twice as large. Precisely as the authors say: "Although it was difficult to predict ex ante which specific advertising features would matter most in this context, the features that do matter have large effects."

6 Emerging Learning Mechanisms for Development Projects/Policies/Programs

RCTs embedded in "independent impact evaluation" (IIE) pair implementing organizations, both NGOs and governments (often financed by donors) with "independent" academics/think tanks/consultants and a randomized intervention/treatment. This approach produces some summary statistics of causal impact of the specific program in its specific context with internal validity. The apparent hope was a proliferation of RCT-IIE stud-

ies plus "systematic reviews" would produce knowledge about "what works" that would produce a superior development practice.[15]

As with the application of any idea to a new domain there have been significant adaptations as the idea of RCT for learning has been adopted:

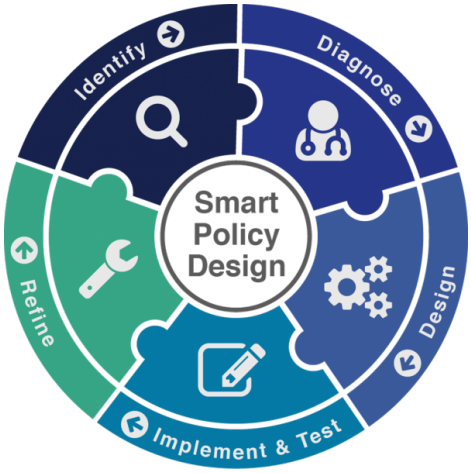
First, there are in development practice few truly "independent" impact evaluations. In most instances, those doing the RCT study work with the implementing organization and are engaged in producing the program design that is implemented or are, as with "field experiments", both implementer and evaluator. Here we are thinking about RCT-IE not RCT-IIE.

Second, the sequence of first program design then design of an evaluation of that program design that uses randomized assignment of program exposure had given way to increasing use of randomization to learn about design dynamically. This often means implementers use randomization of features of program design not just to learn about final impacts but also to explore efficacy in channeling inputs into activities or activities into outputs. Many projects/programs/policies fail not because the causal model of outputs to outcomes fails but because the organization fails to produce activities from inputs or outputs from activities. There are many examples of increased use of randomization to explore program design and provide real time feedback:

- MeE (Pritchett, Samji and Hammer, 2013) (Monitoring, experiential learning, impact Evaluation) is a learning approach embedded into a larger strategy for building organizational capability called PDIA (Problem Driven Iterative Adaptation).

[15]There are many elements of this causal chain that many have argued are implausible. The claims about the pathway from knowledge of what works to development impact has never been grounded in a persuasive positive model of the development process (Rodrik, 2009) or organizational demand for learning (Pritchett, 2002) or "policy maker" adoption (Pritchett, 2009) or the buidling of capability within organizations.

Figure 3: **SMART approach to policy design**



- The group IDinsight (Shah et al., 2013) makes the distinction between KFE (Knowledge-Focused Evaluation) and DFE (Decision-Focused Evaluation) which focuses on using learning to inform decisions about program design during earlier stage implementation than a typical RCT-IE.

- The SMART approach promoted by Evidence for Policy Design (EPoD) (EPoD, 2015) at Center for International Development also emphasizes embedding the feedback loops into the policy design process.

- JPAL and IPA stress engaged learning with partners in a process of program design and its modification rather than just independent impact evaluation and

- The World Bank's research group has supported a "Social Observatory" which engages in not only impact evaluation but also in real-time feedback on projects. This is particularly used on social dimensions of project design such as women's empowerment(Sanyal, Rao and Majumdar, 2015), deliberative democracy (Besley, Pande and

Rao, 2005) and developed methods for this feedback (Bamgartner, Woolcock and Rao, 2010).

6.1 Similar learning approaches in other domains

Advocacy for the use of RCTs in generating evidence for development often appeals to the analogy of the use of double blind control trials in medicine. However, even in medicine there have always been many channels for learning and there are increasing turns away from the standard approach.

In a widely cited article (Paul et al., 2010) discuss the decline productivity and increasing cost of bringing new drugs to market in the traditional paradigm of pharmaceutical RD which relies on expensive Phase II and Phase III RCT testing. Instead they propose a new approach called "quick win, fast fail" which moves more action into the "proof of concept" phase where costs are lower. By shifting costs from later phases to earlier phases this research can generate more early entities and by providing rapid early feedback can raise the probability of success in the later stages. So, while the RCT is the Gold Standard for phase II and III these stages occupy less of the total research budget and learning process.

Berwick (Berwick, 1998) proposes a process called Plan-Do-Study-Act (PDSA), whereby physicians (or groups of medical-delivery professionals in the form of clinics or hospitals), can plan and implement a change to their process as they see fit, study the success of that change, and either adopt or reject that change upon reviewing the effects. Eppstein et. al (2012) suggest a similar process for learning, Quality Improvement Collaboratives (QIC) also in a medical setting. They propose that a number of agents (hospitals) implement Berwick's PDSA proposal, and share their discovered best practices on an ongoing basis, each adopting recommended practices of the others. Eventually, through several si-

multaneous PDSA mechanisms, they will converge on an "optimal" program design, whose outputs are a local maximum (adjusting the program design will reduce the desired output, although it is possible that a more effective project design exists elsewhere in the design space). Eppstein et al (2012) simulate the QIC learning strategy compared with a standard RCT whereby program alterations are made only when they are proven to be significantly more effective and show in simulations that QIC results in a more effective program than a typical RCT in nearly all cases. RCT is superior only in the highly idealized scenario that the design space is non-rugged and the number of observations in each iteration is quite high. So while the RCT is the Gold Standard for drug approvals it is not necessarily the best learning stategy for improving medical practices in complex organizational settings.

The concept of a Realist Evaluation has appeared largely in the public health literature (J Health Surv Res Policy, Vol 10 Suppl 1 July 2005). A Realist Evaluation identifies three key variables: the Context (C) in which a program is implemented, and the Mechanism (M) through which the program has the desired Outcome (O). As Pawson et al. state, the question in a Realist Evaluation becomes *What is it about this program that works for whom in what circumstances?* thus limiting the question and identifying the relationship between the implementation and the outcome (Pawson et al., 2005). A realist evaluation of a leadership development program in Ghana ((Kwamie, van Dijk and Agyepong, 2014)) relied on an explanatory case study of the program. In addition to H1, the hypothesis they seek to prove or disprove, the authors detail H0: a parallel, alternative hypothesis, which would exist should the proposed hypothesis be rejected, and looked for both characteristics of H1 and of H0 in their analysis. Their research uses a combination of collected data, observation, document review, and semi-structured interviews. Upon finding significant instances that support H0, they rejected the alternative hypothesis they sought to evaluate.

Through the Lean Startup Methodology and the slew of imitators it has generated, the ongoing optimization process has been organized into a popular learning methodology in the startup community. Eric Reiss, author of *The Lean Startup* describes the methodology as "Ideas - Code - Data" where a concept is determined, implemented, tested, and then improved upon to start the circle all over again. Reiss argues that the benefits of learning about your product, how it is used, and whether it meets the needs of your target customers outweigh the costs of going to market too quickly.

The Lean Startup Methodology, and similar processes recommend that entrepreneurs follow a specific model of carefully identifying the problem they aim to solve and characteristics of that market, and then design small experiments to determine whether their hypothesis that their product will solve that problem is correct. At the earliest stage, the small experiment could be speaking with people on the street. In later stages, it might be releasing a beta version of the product and determining ahead of time the number of users after 24 hours that would demonstrate that the product is on the right track. It is notable that entrepreneurs are advised to determine their decision-making rule before running their study, and adjusting some aspect of the product design should the study not obtain the desired results. There is no room for explaining away the results and continuing on the same path. The concept of applying the Lean Startup Methodology to social programs has already gained traction. Acumen+ offers a course called "Lean Startup Principles for Social Impact" (website, 8/20/2015), and Lean Impact for Social Good organizes summits and learning opportunities about applying the Lean Startup Principles to social programs.

7 Conclusion

If the world of development social programs/projects/policies (a) has high dimensional design spaces, (b) is characterized by rugged fitness functions (response surfaces/objective functions) over those design spaces and (c) has a fitness function that is contextual (in many and perhaps unknown ways) then the standard approach of "use theory, design a program, implement with an impact evaluation, and learn from the resulting rigorous evidence" is unlikely to be the optimal approach. Rather, the process has to involve a (potentially extended) period in which learning comes from rapid feedback on movements across the design space to reach effective (if not optimal) program designs. This approach to learning and improvement of program/project/policy design and learning itself has to be built into organizational practices and capabilities and encouraged by the authorizing environment for the devil to be discovered–and exorcised–from the details.

Bibliography

n.d.*a*.

n.d.*b*.

Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2013. "Escaping capability traps through problem driven iterative adaptation (PDIA)." *World Development*, 51: 234–244.

Atkinson, Richard C, and Saul Geiser. 2009. "Reflections on a century of college admissions tests." *educational Researcher*, 38(9): 665–676.

Bamgartner, Michael, Michael Woolcock, and Vijayendra Rao. 2010. "Using Mixed Methods In Monitoring And Evaluation : Experiences From International Development." *World Bank Policy Research Working Paper No. 5245*.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." *Science*, 348.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *The Quarterly Journal of Economics*, 125(1): 263–306.

Bertrand, Marianne, Dean S Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2009. "What's advertising content worth? Evidence from a con-

sumer credit marketing field experiment." *Yale University Economic Growth Center Discussion Paper*, , (968).

Berwick, Donald M. 1998. "Developing and testing changes in delivery of care." *Annals of Internal Medicine*, 128(8): 651–656.

Besley, Timothy, Rohini Pande, and Vijayendra Rao. 2005. "PARTICIPATORY DEMOCRACY IN ACTION: SURVEY EVIDENCE FROM SOUTH INDIA." *Journal of the European Economic Association*, 3(2-3): 648–657.

Bierman, Karen L, Robert L Nix, Jerry J Maples, and Susan A Murphy. 2006. "Examining clinical judgment in an adaptive intervention design: The fast track program." *Journal of Consulting and Clinical Psychology*, 74(3): 468.

Blattman, Chris. 2008. "Impact Evaluation 2.0." *Presentation to the Department for International Development (DFID), London.*

Bold, Tessa, Mwangi S Kimenyi, and Justin Sandefur. 2013. "Public and Private Provision of education in Kenya." *Journal of African Economies*, 22(suppl 2): ii39–ii56.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. 2011. "School Inputs, Household Substitution, and Test Scores." , (16830).

Deaton, Angus S. 2009. "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development." *National Bureau of Economic Research*, January(14690).

Duflo, Esther, Abhijit Banerjee, Rachel Glennerster, and Cynthia G Kinnan. 2013. "The miracle of microfinance? Evidence from a randomized evaluation." National Bureau of Economic Research.

EPoD. 2015.

Eppstein, Margaret J, Jeffrey D Horbar, Jeffrey S Buzas, and Stuart A Kauffman. 2012. "Searching the clinical fitness landscape." *PLOS ONE*, 7(11).

Evans, David, and Anna Popova. 2015. "What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews." *An Analysis of Divergent Findings in Systematic Reviews (February 26, 2015). World Bank Policy Research Working Paper*, , (7203).

Ganco, Martin, and Glenn Hoetker. 2009. "NK modeling methodology in the strategy literature: bounded search on a rugged landscape." *Research methodology in strategy and management*, 5(2009): 237–268.

Ganco, Martin, and Rajshree Agarwal. 2009. "Performance differentials between diversifying entrants and entrepreneurial start-ups: A complexity approach." *Academy of Management Review*, 34(2): 228–252.

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2009. "Contagion and differentiation in unethical behavior the effect of one bad apple on the barrel." *Psychological science*, 20(3): 393–398.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, 1(1): 112–35.

Hanna, Rema, Sarah Bishop, Sara Nadel, Gabe Scheffler, and Katherine Durlacher. 2011. "The effectiveness of anti-corruption policy: What has worked, what hasn't, and what we don't know." EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Heckman, James J., and Sergio Urzua. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics*, 156(1): 27–37.

Higgins, Julian P. T., and Simon G. Thompson. 2002. "Quantifying heterogeneity in meta-analysis." *Statistics in Medicine*, 21: 1539–1558.

Holla, Alaka, and Michael Kremer. 2009. "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." *Center for Global Development Working Paper*, 158(January).

Hoxby, Caroline M, and Christopher Avery. 2012. "The missing" one-offs": The hidden supply of high-achieving, low income students." National Bureau of Economic Research.

Iansiti, Mark. 1995. "Shooting the Rapids: Managing Product Development in Turbulent Environments." *California Management Review*, 38(1).

Imbens, Guido. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*, 48(2): 399–423.

Kauffman, Stuart, and Simon Levin. 1987. "Towards a general theory of adaptive walks on rugged landscapes." *Journal of theoretical Biology*, 128(1): 11–45.

Khagram, Sanjeev, Craig Thomas, Catrina Lucero, and Subarna Mathes. 2009. "Evidence for development effectiveness." *Journal of Development Effectiveness*, 1(3): 247–270.

Kline, Brendan, and Elie Tamer. 2011. "Using observational vs. randomized controlled trial data to learn about treatment effects." *Department of Economics, Northwestern University, Evanston.(Available from http://dx. doi. org/10.2139/ssrn. 1810114.)*.

Kwamie, Aku, Han van Dijk, and Irene Akua Agyepong. 2014. "Advancing the application of systems thinking in health: realist evaluation of the Leadership Development Programme for district manager decision-making in Ghana." *Health Res Policy Syst*, 12(29): 10–1186.

Madrian, B, and D Shea. 2001. "The power of suggestion." *Quarterly Journal of Economics*, 18–116.

Madrian, Brigitte C., and Dennis F Shea. n.d.. "The Power Of Suggestion: Inertia In 401(k) Participation And Savings Behavior." *Quarterly Journal of Economics*, 116(4): 1149–1187.

Mansuri, Ghazala; Rao, Vijayendra. 2013. *Localizing Development : Does Participation Work?* Policy Research Report;. Washington, DC: World Bank.

Mbiti, Isaac, Karthik Muralidharan, and Youdi Schipper. 2015. "Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania." *in process.*

McEwan, Patrick. 2015. "Improving Learning in Primary Schools of Developing Countries." *Review of Educational Research*, 85(3): 353–394.

McKEachie, Wilbert J. 1987. "Higher Education's Choices: A Balanced Look at the Problems and Possibilities." *Change*, 19(1): 50–52.

Muralidharan, Karthik, and Paul Glewwe. 2015. "Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications."

Nahum-Shani, Inbal, Min Qian, Daniel Almirall, William E Pelham, Beth Gnagy, Gregory A Fabiano, James G Waxmonsky, Jihnhee Yu, and Susan A

Murphy. 2012. "Experimental design and primary data analysis methods for comparing adaptive interventions." *Psychological methods*, 17(4): 457.

Paul, Steven M., Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. 2010. "How to improve RD productivity: the pharmaceutical industry's grand challenge." *Nature Reviews Drug Discovery*, 9: 203–214.

Pawson, Ray, Trisha Greenhalgh, Gill Harvey, and Kieran Walshe. 2005. "Realist review–a new method of systematic review designed for complex policy interventions." *Journal of health services research & policy*, 10(suppl 1): 21–34.

Perkins, Linda M. 2001. "Meritocracy, Equal Opportunity, and the SAT, Review." *History of Education Quarterly*, 41(1): 89–95.

Prichett, Lant, and Justin Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review*, 105(5): 471–75.

Pritchett, L. 2011. "Development As Experimentation: (and how Experiments Can Play Some Role."

Pritchett, Lant. 2002. "It pays to be ignorant: A simple political economy of rigorous program evaluation." *The Journal of Policy Reform*, 5(4): 251–269.

Pritchett, Lant. 2009. "The Policy Irrelevance of the Economics of Education: Is "Normative as Positive" Just Useless, or Worse?" In *What Works in Development?: Thinking Big and Thinking Small.* , ed. Jessica Cohen and William Easterly. Brookings Institution Press.

Pritchett, Lant, and Amanda Beatty. 2012. "The negative consequences of overambi-

tious curricula in developing countries." *Center for Global Development Working Paper*, , (293).

Pritchett, Lant, and Justin Sandefur. 2014. "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix." *Journal of Globalization and Development*, 4(2): 161–197.

Pritchett, Lant, Salimah Samji, and Jeffrey S Hammer. 2013. "It's all about MeE: Using Structured Experiential Learning ('e') to crawl the design space." *Center for Global Development Working Paper*, , (322).

Ravallion, Martin. 2009. "Should Randomistas Rule?" *Economists Voice*, 6(2): 1–5.

Rodrik, Dani. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In *What Works in Development?: Thinking Big and Thinking Small.* , ed. Jessica Cohen and William Easterly. Brookings Institution Press.

Rubinstein, Yona, James J Heckman, et al. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review*, 91(2): 145–149.

Sabarwal, Shwetlena, David K. Evans, and Anastasia Marshak. 2014. "The permanent input hypothesis : the case of textbooks and (no) student learning in Sierra Leone." *World Bank Policy Research Working Paper*, , (WPS 7021).

Sanyal, Paromita, Vijayendra Rao, and Shruti Majumdar. 2015. "Recasting Culture to Undo Gender: A Sociological Analysis of Jeevika in Rural Bihar, India." *World Bank Policy Research Working Paper No. 7411.*

Shah, Neil Buddy, Paul Wang, Fraker Andrew, and Daniel Gastfriend. 2013.

"Evaluations with impact Decision-focused impact evaluation as a practical policymaking tool." International Initiative for Impact Evaluation Working Paper 25.

Spence, Michael. 1973. "Job market signaling." *The quarterly journal of Economics*, 355–374.

Vivalt, Eva. 2016. "How much can we generalize from impact evaluation results?"

Woolcock, Michael. 2009. "Towards a Plurality of Methods in Project Evaluation: A Contextualised Approach to Understanding Impact Trajectories and Efficacy." *Journal of Development Effectiveness*, 1(1): 1–14.