Learning about Program Design With Rugged Fitness Spaces

Sara Nadel and Lant Pritchett

Abstract

We propose that we live in a world characterized by a *hyperdimensional design space* with a *rugged fitness function.* All aspects of our environment interact to change the outcomes and impact of social programs (hyperdimensionality). Small adjustments to program designs can cause big changes to outcomes and impact (rugged fitness function). In this world, we benefit from learning about social programs through testing more points along the design space, even when our sample sizes are compromised. We run a simulation comparing the program impact when learning is done through crawling the design space (CDS) vs. through randomized control trials and find that CDS generates greater impact. We motivate our study through our experiences designing a skill set-signaling program for new entrants to the labor market.

1 Introduction

The world of development and particularly academic development economics has experienced an absolute boom in "impact evaluations" using randomized control trials (RCT) methods. The method of randomly assigning units to "treatment" and "no treatment" groups avoids the difficult, if not insoluble, problems of inferring causation from observational data. This advantage has led to a veritable explosion in the use of RCTs both by

NGOs and others in evaluating impact in on-going or new programs as well as by academics either working with others or designing their own "interventions." Shah, Wang, Fraker and Gastfriend (2015) estimate that there are now over 2,500 studies using an RCT. Vivalt (2015) has a data base recorded the result of over 600 studies.

This movement was premised on the notion that using an RCT embedded into an impact evaluation was a promising way to learn about development. However, now is a good time to examine that premise. In particular, "randomization" as a technique is flexible and can be embedded into very different learning strategies than an "impact evaluation" that attempts to trace the impact through the entire causal chain (or "theory of change" or "log-frame") from inputs to outcomes. The question is the sequencing of learning and whether the RCT-IE approach is a fruitful technique for learning about program design.

We draw on our experience in attempting to use the standard RCT-IE approach to the design and implementation of a social enterprise of a job placement agency (JPA) to show that the usual approach fails if applied to early in the process. This motivates a simulation analysis that compares two alternative learning strategies, the RCT as impact evaluation approach and a "crawl the design space" (CDS) approach (Pritchett, Samji and Hammer 2013) which provides more flexibility in exploring program design options. We show in a simulation that the RCT approach as a learning strategy results in both worse and more variable outcomes in the performance of the resulting program design than even a naive but flexible CDS learning approach.

2 The Solution is the Problem?

As a doctoral student designing my first solo field research, I was well-prepared with the ideal 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. Review existing literature to learn about previously-tested interventions and those results as part of this process.
4. Implement the program in randomly-identified half of a target population, and compare the outcomes between the two populations. Improved outcomes in the treatment group suggest that the proposed market failure did exist, and the program becomes a proposed policy for resolving this problem moving forward.
5. Write paper.
6. If results are positive, expand and replicate elsewhere.

As it turns out, the above process is the recommended process for dissertation-writing in my field, but is very similar to the recommended processes for intervention design and evaluation in multi-lateral organizations, government, and non-governmental organizations. The best approach for resolving global problems is to identify the problem, write a model, review interventions for resolving the problem, test, and replicate. This is a social science version of the scientific method.

2.1 The Solution in Practice

Identify a Problem: Through my experiences living and working Peru, expanded upon by followup field visits during my studies, I identified a problem and a hypothesis about what caused it: Peruvian youth from marginalized households, despite rapid economic growth in the country and their ongoing investment in higher education, were not securing jobs. Firms complained that they struggled to find talent. *Write a Causal Model:* I hypothesized that the signal of higher education as discussed by Michael Spence (Spence, 1973) had broken down as Peru experienced an expansion of higher education offerings. According to the *Ministerio de Trabajo de Peru*, The number of people with higher education increased by 98% from 2001 - 2012, while the number of formal jobs increased by 38%. Talented youth from marginalized households had no effective mechanism to prove their skills and secure one of these increasingly competitive jobs although they now possessed some higher education.

2.1.1 Spence Model

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:

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

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 I wished to study in application applied to how this model is corrupted 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 % 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, I proposed to research the following hypothesis: *A reliable skill set signal → incresed 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 more simple than identifying the intervention. Spence's model identifies higher education as the reliable skill set signal. If higher education no longer plays that role, the ideal intervention would offer a better skill set signal. Designing the intervention was more challenging than identifying the model. While there is a growing body of research about the characteristics of young adults from low income backgrounds who succeed professionally (Heckman, others?), it was not clear that research focusing on the urban poor in the US would apply to the rural poor in Peru. The applicability of 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 it's applicability. *Implement the Intervention:* Farolito's value proposition was to provide a more reliable signal about worker quality, allowing firms to hire higher-quality workers and pay them

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*	

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 worker expects to work does not vary by job. Workers 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 mprove 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 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 great candidates in Huanuco. This differential effect could be related to the population in each city (preferences and professional alternatives), and thus a population preferences and outside options are a dimension that should be considered in the design space.

Table 3: Learning at Farolito

City Date of request	**Piura** Aug 2012	**Piura** Oct 2012	**Piura** Feb 2013	**Piura** Mar 2013	**Chiclayo** Mar 2013	**Chimbote** Mar 2013	**Arequipa** Mar 2013	**Huanuco** 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 I built or the problem I aimed to solve. If I 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 including. 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 my success would begin having to rerun this learning process over again. Alternatively, had I not engaged in this learning process at all, I would have found limited impact of better signalling because I 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:* In trying to encourage users to adopt the test, it became necessary to revise aspects of the program that I would have deemed "insignificant." 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 my causal model, it was harder to see where specific adjustments fit into a model. How could I quantify a better logo or changing the logo color from blue to orange? The granularity of the interventions complicated their role in my model.
- *Hyper-dimensionality:* Adjusting small characteristics of the program highlighted the number of small adjustments that can generate big changes in outcomes. There are hundreds of things to consider in an actual program, and the combination of those things generated drastically different responses as well.

The importance of such small characteristics that were seemingly irrelevant to the model but highly consequential to the implementation underlined holes in the construct validity of my model. This experience was not unique; and is increasingly discussed in literature. Muralidharan (Muralidharan, 2015) demonstrates the hyper-dimensionality of program implementation in his reference to four unique textbooks-in-schools project, all of which failed. In each project, the failure is attributed to different reasons: The textbooks were stored away from the children; the textbooks simply crowded out learning materials that households ordinarily purchased; the textbooks were worthless for illiterate students; and the textbooks were useless on their own but quite positive when interacted with teacher performance pay. In short: environmental factors in each of these implementations prevented the benefit of additional learning materials from actually increasing students' learning. More broadly, the minimal applicability of the results of one project to an implementation of a similar project in a different environment has been well documented in papers like Pritchett and Sandefur (Pritchett and Sandefur, 2013). The role of small behavioral nudges in program design and take up, and the granularity of their results, is discussed by Bertrand et. al. in their evaluation of the role of small advertising tweaks and loan takeup (Bertrand et al., 2009). However, there is no mechanism for measuring the strength of a behavioral nudge. While some behavioral responses are known, such as the discontinuous response to price at 0, others are nearly impossible to anticipate. In a world where context matters for design, and where small 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* and *implementing a program that can resolve that market failure* is a big one, with several opportunities for error. In the next section, we discuss the characteristics of the design space that causes these challenges, and review program design / learning methods to optimize the design of the implemented program.

3 Design Space Characteristics and Interaction Optimization

We propose two characteristics of the design space and fitness function that make it difficult to systematically draw conclusions about the optimal program design from previous experimentation, and a learning mechanism to efficiently approximate the optimal intervention in such a design space.

3.1 The Design Space is Hyperdimensional

We propose that intervention impact is influenced by a nearly infinite number of factors: environment, population characteristics, political setting, implementing organization, and others. As these characteristics contribute to intervention impact, they adjust what the design of the optimal program would be. The best interventions optimize impact in the setting in which they are implemented. Because the design space is contingent upon so many factors, program learning must be done whenever a program is implemented in a new environment. Supposing otherwise is akin to making the tacit assumption that for all characteristics Y, Z, $F[X|Y = i, Z = j] = F[X|Y! = i, Z! = j]$. In other words, construct validity is influenced by the hyperdimensionality of the design space. Research supports this hypothesis. McEwan (McEwan, 2002) review the impact of several ICT programs in schools, finding that impact varied widely, even within the same country. Vivalt (Vivalt, 2014) reviews the outcomes of several programs across different environments, and concludes that past results of programs, when combined, can provide useful guidance on average, but that the results from one specific environment cannot be extrapolated to predict impact in a different environment. We characterize the influences on construct validity into two groups: internal and external factors. An *internal factor* is one related to the capacity of the implementing agency (Andrews, Pritchett and Woolcock, 2013). An *ex-*

ternal factor is a characteristic of the environment or population that reduces the impact of the output. Duflo et al. (Duflo et al., 2013) demonstrate that the benefits of microfinance in one randomized rollout were greater for businesses that were already profitable. By extension, we can expect that a microfinance rollout in a similar community without already-profitable businesses would have a smaller impact. The fact that there are a variety of social programs in existence supports the concept of a hyperdimensional design space. In a simple design space, policy would be identical globally. The design space is physics or chemistry is simple, and as a result, we can predict what a gravitational pull or the arc of a thrown ball will be nearly everywhere that we have the requisite information. In a world with perfect, infinite information, it would be possible to include all factors in a hyperdimensional design space in a model that predicts the impact of specific inputs. Factors could include infrastructure, government functionality, climate, literacy, and even measurements of social cohesion, among others. In many cases, systematic reviews reflect an attempt to do this. However, we do not see success in mapping an optimal design function for interventions, as more and more design factors arise. We do not discard the value of previous research. When our target is reducing poverty, we recognize that Microfinance, Conditional Cash Transfers, Tax Policy, and hundreds of other programs are tools we can use. Many have proposed ways to approach designing a program using inputs from existing evaluations in other environments that recognize the risks of construct validity. Imbens () suggests weighting previous studies by similarities between implementing environments (and by evaluation rigor). However, the possibility that a complete answer can be found in existing literature is nil. In a design space with nearly infinite variables contributing to intervention impact, it is impossible ex-ante to determine optimal intervention design. It is necessary to incorporate learning into each project.

3.2 The Fitness Function is Rugged

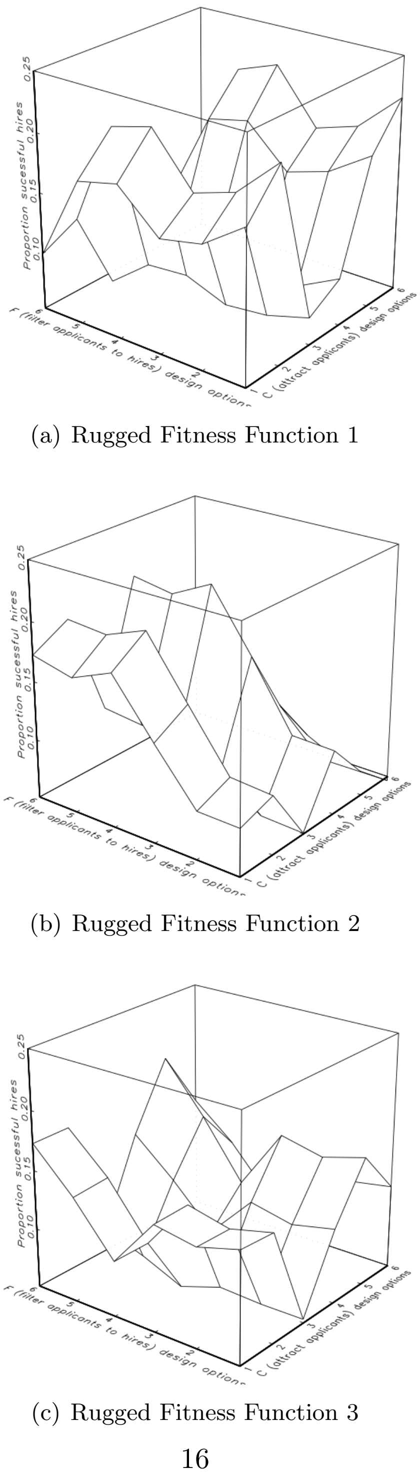
The relationship between variables and program impact is rugged: granularity of variables and interaction effects cause jumps in the fitness function. We define a rugged design space as being one in which $\exists C, F : lim_{c \to C, f \to F} F(c, f)! = F(C, F)$. The ruggedness of the design space reflects two characteristics:

1. There are *discontinuities* in impact at incremental adjustments. The most common example of this phenomenon is the demand discontinuity at positive pricing, illustrated in Kremer and Miguel (2007).

2. There is *granularity* in the strength of an input, causing related output granularity. This phenomenen is clearest in behavioral research. For example, Bertrand et. al. examine the advertising effects of affinity on sales in a loan offering. The find that when a loan advertisement mailer includes a photo that reflects someone of the same race as the potential mailer recipient, the effect on take-up is nearly twice as much as when the photo reflects someone of the same gender. However, measured as *strength* of the "Affinity Strategy", there is no real way to quantify how those two strategies relate to each other (other than the outcome the effect, which is circular). Similarly, Gino et al. (Gino, Ayal and Ariely, 2009) demonstrate that Carnegie Mellon students are more likely to cheat when another cheater in the room is wearing a plain shirt and thus assumed to also be a Carnegie Mellon student compared with when the cheater is wearing a University of Pittsburgh t-shirt and thus assumed to be different. While the finding is notable in that it demonstrates that cheating decisions are not purely cost-benefit, it is impossible to expand that finding to a policy proposal. What is the size of the "other" treatment if the confederate is wearing a t-shirt with a logo from a different university entirely? Or if the confederate is a different age from the other

students? There is no smooth increment of “other” and thus the fitness function is rugged as well.

Figure 1: Three Examples of Rugged Fitness Function



(a) Rugged Fitness Function 1

(b) Rugged Fitness Function 2

(c) Rugged Fitness Function 3

In practice, within a rugged design space, it is not always possible to infer the impact of increasing the strength of input X based on the previous impacts of increasing input X. The interaction of variables create an environment where the highest "peak" (biggest impact) exists on a part of the design space where implementers may not even be testing. Together, a hyperdimensional design space and rugged fitness function confound efforts to design an optimal-impact intervention from existing research. Because of these characteristics, implementers must go through a new learning process that considers a broad range of program inputs each time a program is implemented in a new setting or among a new population. This process has been recognized and applied in settings outside social policy. We review a few such applications below.

4 Existing Alternative Learning Mechanisms

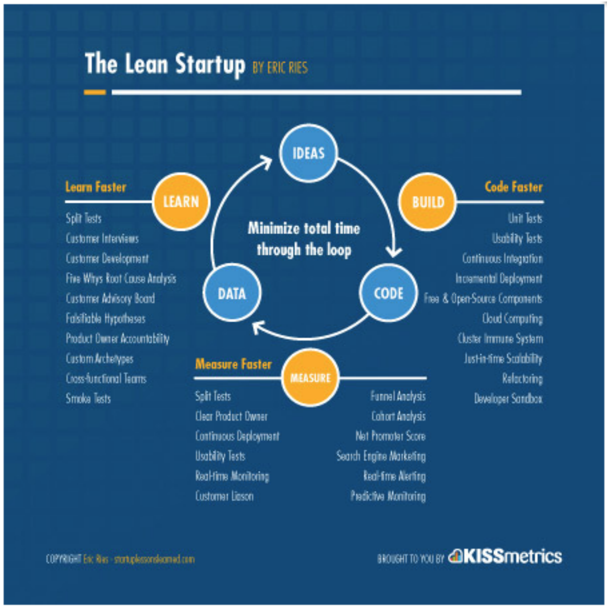
4.1 Medicine

Sequential testing and learning has been employed in a medical setting. 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 suggest a similar process for learning, 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 simultaneous PDSA mechanisms, they will converge on an "optimal" program design, whose outputs are a local maximum (adjusting the program design will have reduce the desired output, although it is possible that a more effective project design exists elsewhere in the design space). The authors simulate this design process compared with a standard RCT whereby program alterations are made only when they are proven to be significantly more effective, and determine that their proposal results in a more effective program than a typical RCT in all cases except in the highly idealized scenario that the design space is non-rugged and the number of observations in each iteration is quite high.

4.2 Lean Startup

Through the Lean Startup Methodology, 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.

Figure 2: Lean Startup Methodology



Accessed from theleanstartup.com website, 8/20/2015

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 in 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 leanring opportunities about applying the Lean Startup Principles to social programs.

4.3 Realist Evaluation

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 et al., “Advancing the application of systems thinking...”) 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 their original hypothesis.

5 Simulation

We are going to address the implications of adopting different learning strategies for situations like Farolito where we are facing a fitness function that is rugged over a hyperdimensional design space and where the fitness function is contextual. We are doing to do this by doing a simulation, which is building an artificial world that abstractly represents some key features of the problem and then seeing what we can learn in this artificial world. The advantage of a simulation approach is that one can understand completely the world and what is going on—in a way never possible in a world with human beings—and one can control the world to see how variations in the simulated world affect outcomes in the simulated world. The disadvantage of course is that a simulation isn't direct proof or evidence about any actual world.

In our simulation we care going to contrast the performance of two different approaches to learning. One learning approach is the RCT. The RCT starts at a given point in the design space, evaluates some alternatives ("treatment arms") and over relatively long periods (say a year) does statistical calculations and moves to the statistically proven best element of the design space. The other approach to learning crawls the design space (hence: CDS) by starting from a given element of the design space and then at the end of a short period evaluates the difference between the initial design space program design and a randomly chosen alternative and then moves to the one that produced the better outcome.

In order to be able to easily visualize this we limit the design space to two dimensions, a "C" element (for "communications") and an "F" element (for "filter"). The question is illustrated in Figure 1. Suppose we were facing a fitness function over a design space where each possible "program" is a combination of choice of C strategy and F strategy each of which had N possible options. Then there are N-squared possible designs of a Farolito-like job placement strategy. We measure "fitness" as the number of successful placements (the vertical dimension). Suppose the fitness function is rugged (as illustrated) and the number of possible programs is large. Also suppose the fitness function is "contextual" so the elements of a job matching program that might have worked in Detroit or in New Delhi cannot be assumed to apply to Peru. What is an appropriate learning strategy as a sequence of actions that would lead to a good program design—a combination of C and F that produces a high (of not optimal) outcome?

To describe the simulation we are first going to describe the artificial world then the two possible learning strategies on that world, RCT and CDS and then the results of simulating the application of these strategies to the world.

5.1 Simulation: (Artificial) World

We start with the idea that there a NP people in each period, call it a month, who are potentially interested in a job (though of course the descriptions of the objects of the computer world as "people" or periods a "month" is all just heuristic, the simulation just is the code). A firm wants to hire people for a job. Each person of the NP has an "aptness" for the particular job. This is not an abstract "ability" or "human capital" of the person but a job specific productivity in the job. Each person also has a "hedonic" match to the job such that, if hired, they would choose to stay on the job. Since the objective of a JPA is to place people into jobs with hiring firms (the client of the job match firm) where the firm wants to keep the hired person (so aptness is above a threshold) and the hired person wants to stay in the job (so hedonics are above a threshold).

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.

The JPA design problem is to apply to these NP people a choice of communications strategy

C to attract people to the come to and apply at the JPA and filter F so that those who pass the filter and hence are hire by the JPA's client are good matches.

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

S 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 interviews
2. Handwritten test
3. Online filter which confirms applicant meets basic job requirements
4. Computerized test given in a supervised environment

Relative strength of program design components is frequently, perhaps usually, nebulous. Behavioral studies are a useful example of this nebulousness. Bertrand et. al. (I'm referring to "What's advertising content worth?", 2009) examine the advertising effects of affinity on sales in a loan offering. They examine the effects of the photo in an advertising mailer representing someone of the same gender on loan take up, and they examen the effects of the photo in the same mailer representing someone of the same race on loan take up. We learn from the study that when the photo reflects someone of the same race, the effect on take-up is nearly twice as much as when the photo reflects someone of the same gender. However, measured as *strength* of the "Affinity Strategy", there is no real way to quantify how those two strategies relate to each other (other than the outcome the effect, which is circular).

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

A person j from the NP 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 and just attract more people to apply (an increase in c_1) or it try and 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 reduce the total successful placements. Conversely, if the communications strategy attracts many on the basis of hedonic match but do not meet the aptness then, for a given filter applied by the JPA, more "bad hires" would be made—people hired but were a mistake because there were not in fact high aptness or productivity.

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

The second element of the JPA program design is the application of the filter, which is two sided. That is, the application of some assessment produces an estimate of the aptness and hence filters out as hires those 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, wants the job).

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

The number of successes in a given month is the number of hires who are truly, based on their actual aptness and hedonic match with the job, above the threshold as this implies they will be hires who both the firms desires to stay and who desire to stay. Since the JPA client firm wants to fill the position with productive hires and low hiring cost per applicant this is the desired outcome.

5.2 Simulation: Learning strategies in the artificial world

Our artificial world is designed to be *rugged* as illustrated in Figure 1 meaning three things. First, conditional on any one strategy the fitness function is not linear (or quadratic) in the other strategy. Second, the fitness function is interactive so that the path along S strategies is not the same for all C strategies. The difference in outcomes between any two S strategies depends on the C strategy so that the experiment of doing S_j versus S_k is for C_j is not (necessarily) the same as C_k. Third, this differences across (C,F) strategies are "big" in the relevant fitness metric even across "local" alternatives. 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.

We are going to assume that the fitness function is fixed over time, but contextual and hence unknown for the given context (where context can include country, region, implementing organization, availability of other alternatives, etc.). We are going to simulate a period of 24 periods (think months) which feedback on outcomes at the end of each month and apply two different learning strategies (CDS and RCT). Each of the two learning strategies starts at the same point and then, relying on the feedback from outcomes, dynamically alters 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.

5.2.1 Learning Strategy: CDS (Crawl Design Space)

Both CDS and RCT start at a given program design, (C_0, F_0).

In the CDS learning strategy each period t two strategies are implemented: current best and an alternative. The alternative is chosen each period from all other strategies besides the current best with replacement. At the end of each period t the outcomes of successful hires from the two strategies are compared as simple counts (or averages as the number of potentials is fixed at NP) as the fitness function (FF) and:

If:

$FF(C_{currentbest,t}, F_{current_best,t} > FF(C_{alternative,t}, F_{alternative,t})$

then the 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 of the periods (24 in this simulation exercise) 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 mean it to be so as we are not searching for the ”optimal” learning strategy [1] rather we are attempting to articulate

[1] We are reasonably confident there is no generally ”optimal” algorithm for finding the optimum of an arbitrary and arbitrarily rugged fitness function.

a learning strategy that pretty much any organization could implement.

5.2.2 Learning Strategy: RCT

The learning strategy for RCT is intended to mimic in this artificial world 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 shifts treatments relatively infrequently.

We model this by having the RCT start from exactly the same starting point as CDS. This default or ”current best” program design is tested against one other alternative, which is chosen by choosing a local alternative [2] which alters the strategy in just one dimension. In our simulation we first search in the C dimension and then in the F dimension.

The principal difference is that, in the interests of ”statistical power” and to maintain the ”integrity of the experiment” rather than making modifications to the program design each month the program is changed only at the end of one year (12 periods/months).

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 alternative adopted if it is statistically significantly better than the current best.

[2] We treat ”locality” as a cycle so that alternative 1 is ”local” to both alternative 2 and alternative D+1, where D is the number of possible choices for either C or F designs.

The result is an RCT 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 simulation is built around the following key differences between the learning strategies:

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

While one might object that this simulation is "cooked" in favor of the CDS strategy our response is three-fold. One, we think that the description of program design changing at best once a year is not a complete caricature of the actual practice of RCTs. We personally have been acquainted and/or directly involved with a several RCT studies in which a treatment arm was obviously badly failing but the experimenters insisted it be maintained so that the study design was implemented [3] Two, we also feel it is not a terrible caricature of RCTs to imagine one "main" and one "alternative" treatment arms. The interests of

[3] One of my (Lant) 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 insisting they stick to a program design the NGO had recognised as flawed and wished to abandon.

statistical power with potentially noisy measurement and modest impact size tend to limit the number of treatment arms to a small integer. Many just implement one program and test it against the counter-factual of "no treatment," others have one alternative, some (but few) as many as four. Three, given the attention that "impact evaluation" style RCTs have recieved as a method for learning about "what works" in development have received if it is really so easy to cook the books against them as a learning tool this is itself revealing.

5.3 Mechanics of the Simulation

The simulations have three steps for each set of parameters of the simulation such as the number of program design options for each of C and F, the random noise in the filter and application process, correlation of aptness and hedonic match and the ruggedness of the fitness function.

Step I: Fix the design and fitness space by determining the values of c_1, c_2, and c_3 which determine a C design f_1 and f_2 that determine a F design for each of NA design alternatives.

For C the first strategy was always $C_1 = (.5, 0, 0)$ which was the "default" strategy that attracted applicants uncorrelated with either aptness or match. For the remaining NA-1 strategies the elements were chosen randomly from the possibilities: -.5,0,.2,.5 and 1. So for instance a C design triplet (.2,-.5,1) would attract fewer default applicants than the base strategy but attract applicants in inverse correlation with aptness but positive on match.

The elements of an F design were chosen from the possible values 0,.5,1,1.5 or 2 randomly for each element for each of NA alternatives. For instance an F design of (1.5, 0) would strongly select on aptness and dropout would be uncorrelated with match at the filter stage.

There is also noise introduced into the application, aptness and drop-out stages which is parametrically adjustable.

These choices fix the fitness space (at least on average) as these applied for the formula for thresholds produce from each set of NP possibles the probabilities of passing the thresholds and being hired while a true match. Step II: Iterate through 24 periods. Each period NP individuals (which in the base case is 1000) are the pool of potential applicants. The C and F program designs and produce a pool of applicants to the JPA which applies the assessment of aptness (which can also cause, for various reasons, applicants to drop out). Those who pass the the aptness (demand of firm) and match (willing supply of applicant) are hired.

At the end of each period the realized fitness for each of the NA$^2 programdesignsarecomputed.TheCDScompar$

Step III: Calculate performance at the final (end of 24 period) program designs each of the CDS and RCT learning strategies. This then returns to Step I where a new fitness function is drawn.

Our baseline results are based on 1000 iterations through the entire procedure so is the results of applying the two different learning strategies for 1000 different design spaces (of a fixed dimension) and fitness functions.

6 Results of the simulation

The purpose of the simulation exercise is to ask: "In this artificial world built to schematically capture features of the world one faces in attempting to design a program to accomplish an objective when facing a high dimensional design space and rugged fitness function what are the implications of various learning strategies?" The results show the RCT strategy is a low mean (the learning gain is smaller) and high variance (the risk of getting really poor results is larger) learning strategy compared to CDS.

6.1 Baseline results

These simulations produce two main results, illustrated in Table 4 using design spaces from 5 to 10 options for each of C and F (hence between 25 and 100 total possible program designs).

First, the CDS (crawl design space) learning strategy typically reaches a substantially better program design than the RCT learning strategy.

The second column of Table 4 shows the average over 1000 simulations (each of which was a different fitness function) of the excess performance of CDS over RCT scaled as a function of the gap between the "best" and the "average" for each fitness function. Since all of the absolute numbers are more or less arbitrary we feel this is a natural metric for the gain from learning: "how much of the distance between having just having picked a strategy at random (which would produce the average result) and having reached the best possible result was closed by the learning one did?" For 6 options (which is our default) the superiority of CDS over RCT is 47 percent of the total best versus average gap. Overall all options the gain is consistently 45 percent or more.

With the baseline parameters and 6 options the average of the best results across 1000 fitness functions is that .198 of the pool are successful hires, the average across all program designs is only .140 (these raw results are in Appendix Table 6). The average for the RCT learning strategy is .166, which is a rough 20 percent gain over the average (and the starting strategy for both RCT and CDS was randomly chosen so would be the average) but only 84 percent of the best (=.166/.198). In contrast the CDS learning strategy achieved an average of .193 which is 97.5 (=.193/.198) of the best possible outcome. Again, with six options it is not so surprising the CDS reaches near the best as in the simulation there are 24 possible trials.

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 gets lower as the ratio of the 24 trials to the total design space gets smaller.

Table 4: Simulation Results

(1)	(2)	(3)
Number of options	Gain CDS over RCT	Percent excess of RCT
	to max over average possible	standard deviation
5	0.517	2.303
6	0.473	2.397
7	0.457	2.431
8	0.461	2.648
9	0.464	2.790
10	0.447	2.835

The second main result of the simulations is that the CDS (crawl design space) learning strategy has a much lower variance in final program design outcomes across alternative fitness functions than does the RCT strategy. Column 3 of Table 4 shows that the ratio of the RCT to CDS standard deviation is between 2.3 times higher with a design space of 25 options and 2.8 times higher when there are 100 options. This is a very intuitive in the context of the simulation but also very important result when thinking about the world.

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 (though given ruggedness a bad 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 was 0.106–which is substantially worse than the average outcome of 0.14.

The result that the RCT learning strategy has high variance is important for two reasons.

One, this simulation result makes the existing findings on the lack of external validity of RCT results of the same general program type (e.g. ”job placement” ”micro-credit” ”ICT in classrooms” ”business training”) across contexts (and within contexts across treatment arms). Suppose the real world for programs really does have a high dimensional design space and the fitness function over the design space is both rugged in a given context and differs across contexts. Then as our simulations show, one would expect to find exactly what Vivalt (2015) or Pritchett and Sandefur (2015) do find, that the variability of impact of the same program type is enormous large both across contexts and program design variants (”treatment arms”) even within the same study.

Two, suppose one imagined that RCTs were generating ”knowledge” and suppose one imagined this empirical ”knowledge” or ”rigorous evidence” had both construct validity (that the particular program design was adequately representative of programs of its type) and external validity (the fitness function was similar across contexts) then one could go badly wrong. That is, suppose did one RCT to produce ”rigorous evidence” about the benefits of ”providing textbooks” or ”de-worming” or ”cameras in classrooms” and produced a specific empirical result. This is the impact of one program design option (or perhaps across a small number of variants) in one context [4]. This evidence, even if "rigorous" due to internal validity for causal inference will have little reliability in predicting program impact. One could either find a program impact so low one "failed to reject" program impact was zero or alternatively find a successful program design matched to contextual fitness function that exaggerated the impact of the "typical" program of its type.

6.2 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 match hedonics. In the base case we assumed that there was no correlation between a persons aptness in a job and their preferences. 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 5 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 .8. This produced roughly similar results.

Differences across program design. In the base case the elements of each dimension of design space (c_1, c_2, c_3) and (f_1, f_2) took on specific numeric values that determine their

[4] Where "context" is itself under-specified as people think of "country" as the context but "context" could well be (and has been shown to be in some applications) regional, organizational, personal, dependent on history, dependent on existing alternatives, etc.).

efficacy (e.g. f_1 as a more positive number made a better filter, c_2 as a negative number drew a worse pool of applicants). As a variant we just multiplied the program designs by 3, which increased the variability of the program impact. As can be seen in the third row this increased substantially the variability of the RCT results versus the CDS. This is intuitive as this should have increased the ruggedness of the space by making some program designs really awful and others better.

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, the individual specific ϵ's in the various equations, to the process. The result is that the learning advantage of the CDS over RCT persists but the relative variability of RCT versus CDS final program design outcomes declines. Again this is intuitive as it increase the uncertainty of identifying precisely the program design versus random elements month to month and hence reduces the ability of the CDS to identify a good program design.

Table 5: Variations on the Base Case

(1)	(2)	(3)	(4)
Number of options	Gain CDS over RCT to max over average possible	Percent excess of RCT standard deviation	Description
6	0.473	2.40	Base case parameters
6	0.446	2.80	Correlation of aptness and match 0.8 (instead of zero)
6	0.391	5.32	Design parameter sets c_1, c_2, c_3 and f_1, f_2 multiplied by 3 for all 6 options
6	0.485	1.57	Variance of the noise in decision rules increased

The results of a simulation of course prove nothing general (just as the empirical findings of an impact evaluation prove nothing general) but can be thought of as a ”proof by construction” or perhaps just a ”counter-example.” We have shown that we can construct an artificial world and that in that artificial world the strategy of learning about program impact by launching an RCT impact evaluation that holds program design fixed across relatively few treatment arms is strictly dominated by an unsophisticated learning strategy of just crawling the design space.

7 Conclusion

There was a burst of enthusiasm for the use of RCTs as a learning tool for development. A characterization (hopefully not caricature) of this approach was that many organizations, both NGOs and governments (often financed by donors), were implementing projects. These projects did not use prospective randomization of units into treatment and control and hence the attempts at evaluation of the impacts of these projects were hopelessly are producing any rigorous evidence about causal inference. The idea was that pairing these projects with evaluators (academic and others) would produce a body of rigorous evidence about ”what works” and from this body of evidence ”systematic reviews” could provide ”knowledge” (perhaps even ”scientific” knowledge) that would produce a superior development practice.

Our paper explains one possible why there has been a massive shift away from this vision. If the the actual world of development practice has fitness functions which are rugged–such that program design matters–over hyper-dimensional design spaces–program design has to make many, possibly interacting, choices, and fitness functions are contextual–so that what plays in New York doesn’t play in Peoria–then the RCT-IE approach will have three significant downsides. One, it will not be useful to organizations that are searching for the locally appropriate design. The involvement of actors with an interest in "impact evaluation" who want to protect the integrity of the experiment is at odds with those who want to alter the program design in response to real time feedback. Second, the granularity of the reality of program design (particularly as informed by behavioral economics) versus attempts to summarize evidence about broad classes of programs such as "micro-credit" or "business training" or "ICT in classrooms" or "job placement" will lead to inappropriate and/or inadequate advice as there is no construct validity of the class of programs. Third, with a lack of construct validity there will be a lack of "external validity" as the results of RCTs will not be more predictive of program impact. [5]

This is leading many people that are interested in more than just using RCTs as a tool to write academic papers to pursue learning strategies that may use randomization, but which do so in a way that is embedded with the organization, allows for more rapid feedback loops, and often focuses on a narrower part of the causal chain often looking just at producing outputs rather than full evaluation of "impact" on outcomes. There are many rapidly emerging approaches that share these themes. The MeE (Monitoring, experiential learning, impact Evaluation) is one (Pritchett, Samji and Hammer, 2013). The SMART approach to policy design promoted by EPoD is another (EPoD, 2015). The group IDinsight (Shah et al., 2013) is making the distinction between KFE (Knowledge-Focused Evaluation) and DFE (Decision-Focused Evaluation). Groups like JPAL and IPA are increasingly downplaying the role of "independent" impact evaluation and stressing much more a learning process in which those involved in evaluation are engaged in a process of program design

[5] Vivalt (2015) shows this lack of external validity across many programs, Bold, et al (2013) show lack of replicability of result when a program was scaled by a different organization, and Pritchett and Sandefur (2015) show that even the simple regression methods that are unsophisticated about causality can provide better predictions about program impact that "rigorous" evidence from other contexts, and Evans and Popova (2015) show "systematic reviews"–even of the same topic by the same organization often come to very different conclusions.

and its modification. All of these are very much like CDS learning strategy we explore and move away from the narrower interpretation of RCT-IE as just neutral evaluators doing science on (to) other organizations' projects and programs in the interests of codifiable "knowledge."

A Appendix

A.1 Additional Simulation Results

Table 6: Additional Simulation Results

(1) Number Options	(2) Max. Success	(3) Avg Strategy	(4) RCT Result	(5) CDS Result	(6) Ratio: Gain CDS over RCT over (Max-Avg.)
5	0.196	0.138	0.162	0.192	0.501
6	0.198	0.140	0.166	0.193	0.469
7	0.200	0.140	0.166	0.194	0.455
8	0.202	0.140	0.166	0.195	0.457
9	0.203	0.140	0.166	0.195	0.459
10	0.204	0.140	0.167	0.196	0.442

Bibliography

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.

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

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.

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.

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.

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.

McEwan, William. 2002. "Virtual machine technologies and their application in the delivery of ICT." *Proceedings of the 15th Annual NACCQ, Hamilton, New Zealand*, 2002.

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

Muralidharan, Karthik. 2015. "Comments at RISE Meeting."

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.

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.

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

Pritchett, Lant, and Amanda Beatty. 2012. "The negative consequences of overambitious curricula in developing countries." *Center for Global Development Working Paper*, , (293).

Pritchett, Lant, and Justin Sandefur. 2013. "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix." *Center for Global Development Working Paper*, 336(August).

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).

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. 2014. "How much can we generalize from impact evaluation results?"