Optimally Generate Policy-Based Evidence Before Scaling

John A. List
U. Chicago, ANU, and NBER

Feb. 2024

Forthcoming *Nature*

Preface

Social scientists have increasingly turned to the experimental method to understand human behavior. One critical issue that makes solving social problems difficult is "scaling" the idea from a small group to a larger group in more diverse situations. The urgency of scaling policies impacts us every day, whether it is protecting the health and safety of a community or enhancing the opportunities of future generations. Yet, a common result is that when we scale ideas most experience a "voltage drop": upon scaling, the benefit-cost profile depreciates considerably. To combat voltage drops, we must optimally generate *policy-based evidence*. Optimality requires answering two crucial questions: what information to generate and in what sequence. The economics underlying the science of scaling provides insights into these questions, which are in some cases at odds with conventional approaches. For example, there are important situations wherein I advocate flipping the traditional social science research model to an approach that, *from the beginning*, produces the type of policy-based evidence that the science of scaling demands. To do so, I propose augmenting efficacy trials by including relevant tests of scale in the original discovery process, which forces the scientist to naturally start with a recognition of the big picture: what information do I need to have scaling confidence?

Over the past four centuries, the experimental method has produced a steady stream of deep insights, helping us understand the world around us. From the work of Galileo, who tested his theories of falling bodies by using quantitative experiments in the early 17th century, to Rosalind Franklin's X-ray diffraction experiment, which was central in the construction of a theory of the chemical structure of DNA, the scientific approach has been a key engine to knowledge creation, economic growth, and enhanced societal well-being.

Social scientists have increasingly turned to the experimental method to understand human behavior and how the world around us affects that behavior. Field experimentation in the social sciences has uncovered insights from the theoretical to the

practical. For example, at odds with his more philosophical contemporaries, William McCall insisted on quantitative measures to test the validity of education programs[1]. For his efforts, McCall is credited as an early proponent of using randomization rather than matching to exclude rival hypothesis, and his work continues to influence field experiments conducted in education today. In political science, Harold Gosnell and Charles Merriam are oftentimes credited with conducting the first social "megaproject" when they explored techniques to enhance voter turnout. One example is Gosnell[2], who found that the use of cartoons and informational reminders increased both voter turnout and votes cast by roughly 10 percent. In a similar spirit, Ronald Fisher used the concept of randomization in his field experiments to understand the agricultural production function[3]. Likewise, Kurt Lewin[4] directly studied questions of where, how, why, and under what conditions an effect does or does not appear, anticipating contemporary discussions of generalizability and scaling.

Within economics, early experimental work took the form of laboratory experiments[5]. Over the past few decades, however, economists have begun to depart systematically from these roots; they are recruiting subjects in the field rather than in the classroom, using field context rather than abstract terminology in the instructions, and striving to carry out randomization in naturally-occurring settings, often without the research subjects being aware that they are part of an experiment, which has been denoted a "natural field experiment"[6,7].

Such studies have spanned broad areas of our economy, lending insights into the underpinnings of markets and how we can achieve more efficient and equitable exchange[8], crucial factors driving unemployment[9], how simple social science principles can improve public health interventions[10], and what are the roots of generosity, reciprocity, altruism, and how can we use the private provision of public goods to overcome market failures and collective action to solve the free-rider problem[11–15]? Each of these areas, and several others, have been advanced by modern field experimentation in economics[16].

One crucial question for the experimental research agenda in the social sciences relates to the scale-up problem: can this idea work at scale? In the physical sciences, where it is

assumed that the same physical laws prevail everywhere, Galileo could fluidly extrapolate his experimental results to the heavenly bodies. Yet, when experimenting with humans, the scale-up problem takes center stage. In its simplest form, this question relates to the proliferation of an idea or policy from a small group—students at a certain school, for example—to a larger group in more diverse situations[17,18]. The urgency of scaling important policies impacts us every day, whether it is protecting the health and safety of a community, improving the sustainability of development, or enhancing the educational opportunities of future generations. Scaling was once viewed as important only in the domain of business startups, yet it underlies all social and technological progress, since the innovations that have the best chance to change the world are those that reach the largest number of people. Innovation is crucial, but diffusion is its perfect complement.

While its import is undeniable, the scaling process is not simple, with pitfalls present every step of the way, running from the seed of an idea to well after policy launch. One consequence of having multiple fault lines is that ideas tend to experience a "voltage drop": the benefit-cost profile depreciates considerably when moving from the small to the large[19–22]. Indeed, across disciplines ranging from software development to medicine to education and beyond, between 50 to 90 percent of programs lose voltage at scale, with many of the scaled programs having weak or no effect at scale even though they showed great promise in early trials[18,23]. Elegant ideas to fight climate change work in a Petri dish but fall apart at scale[24]. An educational reform works in pilots but when scaled to the whole country effect sizes considerably diminish[25]. In my own work, I have found that certain ideas are predictably scalable, while others are predictably unscalable, and the reasons for each are visible through a scientific lens[18].

Today, policymakers focus on developing "evidence-based policies," which in most circles means advancing public policies, programs, and practices that are grounded in some type of empirical evidence, with little attention paid to how far removed that evidence is from what, for whom, where, and how they want to scale. To combat voltage drops, I argue in this Perspective that we must optimally generate *policy-based evidence* before the decisionmaker makes the decision to scale (and even throughout the life cycle

of the policy). Optimality requires a recognition of two key aspects of the problem: what information to generate and in what sequence. The economics of data generation and the behavioral economics of the decisionmaker and scientist provide insights into these two crucial areas.

To provide insights into these two key aspects, I argue that the nature in which policy-based evidence is generated demands an understanding of a scaling continuum: there are some instances when the researcher should adopt a learning by doing approach, whereby first tests are purely for efficacy and future explorations tinker with populations of people, treatment types, and testing boundary conditions incrementally to produce a scalable policy. While this approach is the current status quo, there are other important cases wherein the researcher should *start* by imagining what a successful, fully implemented intervention looks like, applied to the *entire* population with their varying situations, sustained over a long period of time. I denote this class of ideas as "HFIDs": high fixed cost with impatient decisionmakers.

In the case of HFIDs, I advocate flipping the traditional research and policy-development models. To accomplish this goal, original experimental designs and prototyping must address potential stumbling blocks from the beginning: we must engage in backward induction and understand the big picture rather than focus on the immediate incentives we face as scholars. In the traditional model (learning by doing), scholars only consider (if they do at all) how to scale ideas and interventions after they are shown to work in efficacy tests. In the experimental community, such efficacy tests are commonly denoted as "A/B tests."

This leads to a natural question, if classic A/B tests are insufficient in such cases, what data should be generated? I argue that we must begin with an understanding of the factors that cause voltage drops and a discernment of the mechanisms underlying why the policy works. This includes identifying important mediators, heterogeneity, and causal moderators as well as whether the idea remains promising in the face of crucial real-world constraints it will face at scale. Accordingly, it is often economically and scientifically efficient to not only do the efficacy test, but *also* relevant tests of scale and mechanisms *within* the original discovery process.

As a mnemonic, I denote this approach as "Option C thinking," which effectively asks: if I want to scale this idea what extra information do I need beyond an A/B test? To answer this question, I argue that we must include a scalable version of the studied program alongside the "best case" A/B test. Leveraging Option C thinking in our initial designs is not simply adding a new treatment arm; rather, it augments the traditional approach with an experimental enhancement that produces the type of policy-based evidence that the science of scaling demands from the *beginning*.

What Data to Generate? *Policy-Based Evidence*

A hallmark of public policy decision-making is a comparison of the benefits and costs associated with proposed programs or regulations. Likewise, firms, both for profit and non-profits, along with many individuals follow a basic rule of comparing benefits and costs of a proposed action. In this spirit, as knowledge creators interested in scaling, we must answer the following question: after a program has been claimed to pass a benefit-cost test in an initial study, what is the probability that it will pass a benefit-cost test in the target setting of interest?

A crucial piece of information that is necessary to answer this question relates to the transportable of information from the experimental setting to the scaled setting. The difficulty arises because the object of interest in social science experiments is humans; thus, in most (all?) cases attempting to develop behavioral laws that parallel those from the natural sciences is a fool's errand. Humans are creatures of habit, until we are not. We are animals that respond to stimuli in predictable ways, until something imperceptible to the scientist in the background changes and causes a seemingly irrational reaction. In short, behavioral principles observed in one environment are not (always) shared broadly. Humans, who are inherently maddening experimental subjects, make the social sciences the "harder sciences".

In contrast, for physical laws and processes, evidence to date supports the idea that what happens in the lab is equally valid in the broader world. Shapley[26], for instance, noted that "as far as we can tell, the same physical laws prevail everywhere (p.43)." Likewise, Newton[27] scribed that "the qualities … which are found to belong to all bodies within the

reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever (p. 398)." With humans, the heterogeneity in populations and situations yields different experimental results quantitatively, and even sometimes qualitatively across scaling domains.

Within economics and the broader social sciences generalizability models naturally arise from Mill's assumption of the lawfulness of nature in that they are based on the "distance" between time, space, population, and the decision environment between the study setting and the setting of ultimate interest[28-31]. Generalization revolves around the degree of "stickiness" of nature: the models assume that as two decisions become closer in distance, the congruency in response effects will heighten because the two environments tend to follow similar laws. While theoretically this might be pleasing, it leaves open the crucial question: what key factors underlie the voltage effect?

Leveraging a set of economic models and the voluminous empirical literature, in List[18] I outline five threats that can cause voltage drops and prevent an idea from having its promised impact when scaled. I denote these as the 5Vital signs, and I summarize them in Exhibit 1. In effect, these five threats inform us of what information is necessary to generate policy-based evidence.

Vital Sign #1: False Positives

An active debate has emerged in the social sciences that claims there is a "credibility crisis," whereby the foundation of the experimental approach and the credibility of the received results are called into question[32-35]. The debate has evolved into several streams of inquiry, but the channel connecting them is false positives, with lack of replication often carrying the water. False positives fall into one of three buckets: statistical error (alpha), human error (how we generate/evaluate/interpret data), and fraud (less rare than we hope[36]). The literature has addressed these concerns by changing both the incentives around creating replications and examining data to control the false positive rate[37-40].

To put this inferential issue into perspective, an often-misunderstood consideration is that there is a crucial distinction between the probability that a reported significant finding in the literature represents a real relationship and the probability that an individual

experiment has uncovered a real relationship[41,42]. For example, even if a fully powered experiment reveals a program works at conventional significant levels in an experiment, if the result was a surprise the likelihood of the program working again, especially at scale, remains a long shot[43-45]. The power of replication to overcome false positives becomes abundantly clear to build scientific knowledge around scaling[18,42,46].

Vital Sign #2: Representativeness of the Population

The second Vital sign is representativeness of the experimental sample. In this case, the voltage effect arises when the decisionmaker assumes that the small subset of people who are affected by the policy in the Petri dish are more representative of the general population than they are at scale. There are two key aspects of the sample and the population that must be considered before scaling. The first concerns selection into the experiment: after choosing the population, is the experimental sample truly representative of *that* population? The second question relates to the population itself: is the experimental population representative of the target population to which the program will be scaled?

The first question concerns whether inference drawn from an experiment can be extended to non-participants from the same population. Before taking part in the experiment, individuals must decide whether to participate in the experiment. An example of the exploration of this notion is[47], which examines participation in a laboratory experiment. Their results show that some of the characteristics of participants versus non-participants are significantly different. In particular, the experimental participants have significantly less income, more leisure time, are more likely to major in economics, and are more pro-social. They even find some support for key outcome differences.

This insight is not germane to only economic experiments, as decades ago it was noted that volunteers in human research "typically have more education, higher occupational status, earlier birth position, lower chronological age, higher need for approval and lower authoritarianism than non-volunteers."[48] Indeed,[49] concluded that social experimentation is largely the science of "punctual college sophomore" volunteers and have further argued that subjects are more likely to be "scientific do-gooders," interested in the

research, or students who readily cooperate with the experimenter and seek social approval (see also[50]).

One method to create a representative sample from the experimental subsample is to use propensity score methods to adjust the sample for the population features. This approach is effective if selection into the experiment occurs only on observables that the analyst can use in the statistical analysis, or if the assumption of conditional independence is met. Another approach to achieving representativeness is to conduct a natural field experiment[6,7,16], which handles the selection decision by covertness: subjects neither know that they are being randomized into treatment nor that their behavior is being scrutinized. Of course, this might not be applicable in all situations.

Concerning the second question—is the experimental population representative of the target population?—this area has been well traveled in the extant social science literature[51-55]. One line of thought is that with multiple potential locations, if the researcher chooses locations at random in an initial stage of the experimental design, then this will lead to generalizable results across all potential locations. This approach only helps to tackle our inferential problem if subjects themselves are randomly assigned across locations, which of course is dubious[56].

A more credible identification approach is to gather information on covariates in the experimental sample and estimate conditional average treatment effects (CATEs). Then, with a good assortment of individuals in the experimental sample, the researcher can average these CATEs over the distribution of covariates in the target population, "adjusting" treatment effects accordingly. This approach is often used when the researcher constructs a random sample that is useful for interrogating heterogeneous treatment effects by using multi-site trials and then relies on probability re-weighting and effect estimation using functional form assumptions to transport insights[56-62]. Once again, this approach identifies the relevant treatment effect if selection *into the experiment* occurs only on observable characteristics. To avoid such selection issues, conducting a natural field experiment might be necessary[6,7,42].

Vital Sign #3: Spillovers

As the above discussion intimates, there are factors beyond the sampled individuals that can affect scaling. The third vital sign is an understanding that the implementation of the policy can have unintended consequences, or spillovers, that work against (or in favor of) the desired outcome. Such "general equilibrium" effects, or changes to the overall market or system outside a program or policy evaluation lurk everywhere, whether from labor market policies to "hot spot" policing in criminology to effects of housing and education voucher programs. This is because if a policy alters the value of a certain program outcome, the benefit an individual gains from that program may be relatively small if many more people are also benefiting from the program. For example, more people benefiting from certain educational credentials may decrease the individual benefit of that credential if a program leads to many more people earning that credential. Much research has been completed that measure such spillovers[63,64].

Vital Sign #4: Supply Side

The fourth vital sign is the "supply side" of scaling—if a policy has diseconomies of scale as it expands it becomes costlier to sustain. The supply side of scaling a policy is well-trodden theoretically and empirically. Economies and diseconomies of scale have been important to economists since the beginning, playing a central role in Adam Smith's Wealth of Nations[65]. Below we will learn that the supply side of research can also play a crucial role in optimally sequencing our information acquisition.

Vital Sign #5: Representativeness of the Situation

Our last key component of scaling is representativeness of the situation. This is a broad and rich category that includes any situational feature that impacts the science around scaling. For example, interventions that worked well to accelerate COVID vaccination when vaccines were initially rolled out were less effective later when that initial high demand was largely quenched; and, it seems to be situational features rather than representativeness of the population that caused such voltage drops[66]. Likewise, the Becoming a Man Program[67] worked well initially but when the program was expanded, it lost voltage, and situational features likely played a role, as discussed in[68]. Finally, an early study showed that FAFSA simplification worked quite well[69], however, when the program was scaled using a variation on the original idea it failed[70].

As these examples make clear, this final vital sign demands that information across a rich assortment of situational features must be generated before having scaling confidence. In its most basic sense, such information can then be used by the researcher to average these types of CATEs over the distribution of situational covariates in the scaled setting, "adjusting" treatment effects accordingly. This approach can be done by using multi-site trials and relying on spatial variation to re-weight in a similar spirit to that done with individual covariates[56-62]. For example, an intervention designed for schools should sample schools that vary in location/income, teacher quality, scholastic outcomes, size, and other variables that are known to be related to key educational outcomes the scholar is attempting to affect. This is because sometimes programs do not scale well due to changes in such details or even who implements the program[18]. Work across the social sciences has grappled with providing *meaningfully* representative samples and situations for decades[18,28-30,71-75].

A general lesson from the literature in this area is that understanding whether a program works with the constraints it will face at scale is invaluable before scaling. The examples show that most discussions focus on ensuring that the *effects of causes* measured in the initial study will remain at scale. A complementary exercise that importantly informs the scaling exercise is to understand the *causes of effects*[76,77]. In this spirit, understanding the mediation path(s) and creating a path that will exist in environments where scaling will occur is vital.

Consider recent work that explores the effect that changing parental beliefs has on early childhood investment and ultimately on childhood outcomes[78]. For this idea to be scalable, it is important to recognize that their approach relies on i) the treatment changing parental beliefs, ii) parents, whose beliefs change, have adequate resources to invest, iii) those investing parents understand how to invest, iv) the parents invest accordingly, and v) those investments move the child outcome of interest within the experimental time-frame. If this entire chain holds in the target setting of interest with the fluidity that it held in the experimental setting, then the program passes the mediation path test.

Exhibit 2 illustrates this idea in a dramatic manner. In this case, the scientist is interested in a simple question: how can I get children to read more books? The top panel of Exhibit 2 provides a crisp sense of what is necessary for a particularly complex solution to scale: the idea will work in school districts that have creative vending machines, a random awaiting unused balance, Title IX rules that provide women's soccer, an ingenious biologist on staff studying geese, a music budget that has not been slashed, city hunting and gun laws that permit such hunting chicanery, and a bibliophile squirrel who happens to have a book lodged in her nest. Except for the names and a few other changes, we can all think of similar examples from the literature and the policy world.

The "machine" that I have created in the top panel of Exhibit 2 also provides another key consideration: great caution should be taken when drawing conclusions from a localized experiment about a policy implemented at scale. In many quarters, splitting a complex problem into several distinct smaller parts has been celebrated. While this approach certainly has merits - I have done the same in some of my field experimental work[79] - care must be taken to understand the complete set of mechanisms before generalizing the insights from a part of the puzzle to the whole. Indeed, without a keen sense of how the whole puzzle fits together (understanding mediators and moderators), one should take great care in generalizing and scaling ideas. The overarching message is that creating ideas that have simple mediation paths is crucial. Greater trust should be given to a policy that is well understood theoretically and one that has a simple mediation path.

My machine provides an ocular depiction of that truth: simplification is useful, but not at the expense of duping oneself to scale the predictable unscalable. The message is that while the academy often rewards complexity, creating ideas that have simple paths are crucial. The simple mediation path in my machine is to have books everywhere, even on trees, and children will read them. If it grows in the garden, it will grow in the wilderness is oftentimes a solution that has great efficacy, even if it might not attract the attention of journal editors.

What Data Generation Sequence is *Optimal*?

Once the 5Vital signs to create policy-based evidence are understood, the next piece of the puzzle is how to generate such evidence in an optimal manner, where optimality in this case is defined as arriving at the correct conclusion using the fewest resources. The traditional research approach to scaling is one of learning by doing: first run a pilot under idealized (not scalable) conditions, and then, once we find promise under those conditions, run scaling trials under more realistic conditions with real-world scaling complexities increasingly introduced.

Beyond medical trials, one area where this approach has been carried out is social scientists exploring the effects of growth mindset and social belonging via multi-site trials[58,60,74,80]. For example, the growth mindset work advanced from early studies that showed effects of growth mindset interventions in certain settings[81] and evolved to develop more scalable insights by systematically testing interventions and then disseminating treatments widely only after those conditions were rigorously tested[58,60]. Within economics, this approach can be found in the learning by doing experiments conducted by Pratham, an Indian NGO. To augment human capital accumulation among its students, Pratham designed a program called "teaching at the right level" ("LEVEL"). The basic idea of LEVEL is to group children according to their knowledge rather than their age. The partnership between researchers and Pratham started with a trial field experiment in the cities of Vadodara and Mumbai[82,83]. Over time, Pratham tinkered with aspects of the program and eventually scaled it to over 33 million children.

While the learning by doing approach has merits, there is an important class of problems wherein we should flip the current approach. Under this new model, scholars and policymakers should *start* by imagining what a successful, fully implemented intervention looks like, applied to the *entire* subject population with their varying situations, sustained over a long period. In such cases, we must engage in backward induction rather than focus on efficacy tests in the beginning.

My reasoning for flipping the research agenda follows from basic economics, which point to three-interrelated drivers. First, short-term incentives encourage researchers to create a Petri dish that provides results that are overly optimistic: consumers of scientific journals demand studies that report important insights[21]. They reward journals via the

purchase of subscriptions and by citing a journal's papers. In turn, a well-documented bias emerges: professional academics, potential funders, and laypeople all regard reporting of large treatment effects as more noteworthy than smaller effects[84]. The result is that we are essentially performing efficacy tests on steroids without telling outsiders. Thus, the first insights given to decisionmakers about an idea's prospects will be of an overly optimistic nature.

The second reason relates to the behavioral economics of the problem: impatient decisionmakers. In the business world, this time-horizon mismatch problem has been denoted as 'short-termism,' and various estimates point to the issue being wide-spread and important. For example, 78% of executives surveyed self-report sacrificing long-term value to smooth quarterly earnings[85], and the qualitative evidence suggests that the social costs of short-termism is up to 20% of potential output[86]. Various unresolved issues and proposals to combat short-termism have been advanced[87,88]. A corollary of short-termism that has been discussed related to policymakers revolves around choices made that involve discounting. The tension between social and private discount rates, particularly for long term problems such as climate change, and especially how that interacts with uncertainty in eventual damages reveals the import of time discounting[89-92].

Likewise, there are recent policy choices where impatience has played a key role. Consider the case of anemia, which affects 1.6 billion people worldwide. To provide solutions to the problem, researchers in India ran pilot studies measuring the benefits of consuming iron-fortified salt on anemia (double-fortified salt (DFS); salt fortified with iron and iodine). While early pilots showed some efficacy, surprisingly, in 2012 Indian officials did a nationwide scale-up of DFS despite the lack of large-scale trials. It turned out that the fortified salt had no effect on the policy goal of reducing general anemia with the broader population[93].

Why did this happen? The original studies had specifically sought out adolescent women. While their unique physiology benefited, these health gains did not manifest at scale. Similar failures at scale have occurred elsewhere with initiatives to decrease rates of STD transmission and promote safe sex, such as providing condoms to a community. Such

practices have produced weak results after expansion due to variations in community mores related to sex[18].

In sum, the first two sequencing considerations suggest that to achieve optimality in some cases we must "tie" the researchers' and decisionmakers' hands. In effect, by committing the researcher to provide relevant information on what the program will involve at scale, the decisionmaker will find it more difficult to ignore that such problems might exist before scaling. This approach has some antecedents in medical trials: a new drug cannot be taken to market until the relevant sign-posts have been passed.

The third reason for reconsidering sequencing of data generation stems from the economics of the idea testing itself: for high start-up cost interventions, backward inducting from the beginning makes economic sense. Consider a parochial example that takes a minor step in this direction. I led the creation and launch of the Chicago Heights Early Childhood Center (CHECC) in 2008. From 2007-2010 I directed the building of two separate pre-schools (from scratch) in Chicago Heights, IL, a suburb south of Chicago. Building the programs/schools represented large start-up costs, as securing relevant licenses, contracts, curricula, buildings, buses, lunches, teachers, administrators, community and schoolboard buy-in, etc., represented a considerable resource cost.

We opened the schools in 2010 and conducted the field experiment from 2010-2014, including nearly 1,500 students annually. One key feature of our field experiment was a full-day pre-school program that emphasized noncognitive skills, such as socialization, active listening, and delaying gratification compared to a program that focused on cognitive skill development. In addition, the experiment included a set of treatments that included parent academies, where the parents were taught rather than their children[64,94-99].

At this point, one might ask: if I want to create a program that scales, how should I design the CHECC experiment at the outset? Upon some introspection, using the traditional learning by doing model for CHECC does not exactly fit. This is because if the researcher had to continuously tinker over time with new programs, ideas, teachers, administrators, populations of people, treatment types, and testing boundary conditions incrementally the cost to build each of the new pre-schools to generate the necessary data would be exorbitant. This is because the fixed cost to begin data collection differentiates

this idea from a drug trial, the "wise interventions," and Pratham's tutoring program. In the CHECC case, even continuing data collection at the same site might feature a high fixed cost if considerable programmatic changes are made. When time costs are added the tinkering model becomes untenable.

We decided to take a different route: introduce Option C to the standard A/B testing approach. I turn to this idea next.

Introducing Option C Thinking to A/B Testing

The cornerstone of the experimental approach in the social sciences is A/B testing. For example, if an early education scholar is interested in testing whether an early education program works, she will gather a group of children and split them into two groups: a control "A" group that does not receive the intervention and the "B" group that receives the early education program. If the experiment satisfies the classical identification assumptions[42], then a causal effect can be recovered. The causal effect is traditionally measured by comparing sample means of particular outcome variables across the two groups, as displayed in Exhibit 3. In the A/B experimental test summarized in the top two arms of Exhibit 3, the program is found to triple Kindergarten Readiness: from 17% to 51%!

One might view this result as extraordinary, and immediately want to scale the program. To understand why that choice is not prudent, consider what exactly we have learned from this research. If it is a typical social science experiment, then it has likely been conducted as an *efficacy test*: the "best case" test of the program is arm B versus the control, arm A. To understand why more information is necessary, we must consider the incentives that the researchers faced. Those incentives are set up to create a petri dish that provides results that gives the intervention its "best shot," or likewise the largest treatment effects.

In this manner, we are answering the wrong question if we are attempting to provide policy advice. We are asking: can this idea work in the petri dish under the best-case situation rather than will this idea work at scale? At this point the astute critic might respond: of course, this is just a 'phase 1' test rather than a 'phase 3' test, it is meant for

efficacy not generalizability. While that reasoning makes sense for medical trials, in most program explorations in the social sciences there is considerable fixed cost to start-up. Accordingly, it is economically and scientifically efficient to not only do the efficacy test, but *also* relevant tests of scale within the original discovery process. The economics of many situations demand such an approach.

I refer to this new approach as Option C thinking. This is not because it represents a simple treatment arm to augment standard A/B testing. Rather, leveraging Option C thinking in our initial designs flips the traditional research model from efficacy trials to an approach that produces the type of policy-based evidence that the science of scaling demands. The Option C analogy is meant to take the initial discovery process from one of focusing purely on the details of an efficacy test to engaging in a bigger picture view, including questions such as: what constraints will the idea face at scale, what key factors can impact scaling, how can theory help me understand mediation paths and moderators in place at scale? Such evidence should be generated in the *original* design alongside the efficacy test. In this vein, theory plays an even greater role in design than usually imagined by experimentalists.

To complete our thought experiment in its simplest possible terms, in Exhibit 3, when following this approach we add treatment arm C, which is a program that includes the constraints or warts that the idea will face at scale. After doing so, we find that the program that will be scaled minimally increases the Kindergarten Readiness—from 17% to 18%—a result that is not statistically significant relative to control and certainly will not pass a benefit-cost test.

Let us now return to our CHECC example. In this case, we augmented traditional A/B testing by using Option C thinking as follows. If we backward-induct from the reality that, at scale in thousands of schools, our program would not have its dream budget or a dream applicant pool of teachers to choose from, then several potential warts emerge. So, we designed our experiment to examine if our curriculum could work with teachers who have varying abilities. That is, we employed teachers who would typically come and work in a school district like Chicago Heights. As we prepared to open CHECC, we hired our 30 teachers and administrators the same way the Chicago Heights public

schools would, from the same candidate pool and with the same salary caps. This choice provided the A/B efficacy test because we had several stellar teachers, but we populated Option C too, which ensured that the situation was representative, at least on the dimension of teacher quality. While it may sound counterintuitive, or even idiotic, not to search out and hire only the best talent in the early stages of an endeavor, it does provide insights necessary to scale from the outset.

Our approach provided a teacher pool that would allow us heterogeneity testing on teacher value-added metrics. Of course, this example focuses on teachers to understand a kindergarten readiness program but beyond teachers, a wealth of other considerations could be examined, as discussed above with the 5Vital signs. For example, at scale we also must convince many principals and superintendents to implement our experiment with our new way of doing things. Fidelity concerns might yield voltage effects at scale[18]. The larger point here is that there are key cases where the researcher should backward induct and explore crucial elements and constraints that the idea will face at scale. But, how can we identify those cases?

Introducing HFIDs

Combining the three-interrelated drivers discussed above—researcher incentives, impatient decisionmakers, and program fixed costs—provides a strong impetus for flipping the scaling research approach for certain idea types. Specifically, uniting the notions of high fixed cost programs and impatient decisionmakers creates the class of potentially scalable ideas that I denote as HFIDs: high fixed cost with impatient decisionmakers. I provide Exhibit 4 to categorize such ideas in a digestible format.

For example, in Region I of Exhibit 4 are true HFIDs. These are programs that are characterized as high fixed cost programs to generate information and are in topic areas where decisionmakers are impatient. Region II of Exhibit 4 relaxes the impatient requirement but continues with high fixed cost programs. Regions III and IV include low fixed cost programs, such as many of those discussed above that use the learning by doing approach.

CHECC falls into Region I or II, and of course it is possible to fall into both. Making my point anecdotally, we can examine how practitioners have used CHECC results. In one instance, policymakers scaled the parental component of CHECC in London without incentivizing parents, ignoring that parent incentives were used in CHECC[18]. While we did not have evidence that showed conclusively that parental incentives were necessary, we advised local officials that they were likely quite important. Their program did not scale well. Had we shown experimentally that a non-negotiable for the program to work was parental incentives, perhaps the officials would have been less likely to scale the program without them. Alternatively, in another distinct case, scaling the full-day pre-school CHECC program has worked well in Bangladesh[100]. This is congruent with the fact that CHECC worked well with a variety of teachers in the initial experiment, thus we had scaling confidence in that input variety.

My perspective is that for those cases that fall into Region I of Exhibit 4, an approach of flipping the traditional research and policy-development models should be used. That is, in the original experimental designs and prototyping, Option C thinking should be included to understand potential stumbling blocks from the beginning. In this sense, using backward induction to create policy-based evidence begins by identifying features that can cause voltage drops and then testing those factors. If the idea falls into Regions II or III, I view the prudent action as considering the benefits and costs of flipping the traditional research and policy-development models. In many such instances, introducing Option C thinking in the beginning will make sense.

While in theory Region IV might yield the least benefits of introducing Option C thinking, I argue that the researcher should still consider doing so if there are economies of scale in data collection. For instance, one question that falls into Quadrant IV relates to non-profits: what is the best email that a fundraiser can send to maximize charitable contributions? With adequate sample sizes, it is quite simple and cost-effective to explore mechanisms and relevant tests of scale *within* the original discovery process when answering this question.

A Path Forward and its Potholes

I suspect that most academics and practitioners agree with my policy-based evidence argument. The key is determining whether, and to what extent, the 5Vital signs curb our scaling enthusiasm. Yet, there is likely more consternation around my sequencing proposal, which remains largely a theoretical construct. This is because it is difficult to find social science experiments that imagine what a successful, fully implemented intervention looks like, applied to the *entire* subject population with their varying situations, sustained over a long period of time and test accordingly at the outset. In this manner, the CHECC example takes a small step using backward induction from the beginning, but many key pieces were not tested because of resource constraints, such as whether incentives are necessary in the parent academy or the effect of administrators.

A useful path forward is to develop an understanding of where and when high fixed start-up cost programs that are delivering insights to impatient decisionmakers arise. Those are ripe cases to not only do the efficacy test, but *also* relevant tests of scale within the original discovery process by adding Option C Thinking. Likewise, even when an idea falls into Regions II, III, or IV of Exhibit 4, it is important to explore the economics of the research process to determine if Option C Thinking should be tested alongside the "best case" program. Policymakers deserve such an approach, and the economics of many program and idea evaluations that have a chance to scale demand such an approach. Exhibit 5 provides a tree-based approach to generating policy-based evidence in an optimal sequential fashion (see also Stuart Buck's excellent discussion[40] of policy-based evidence).

My perspective on timing, however, is not without warts. First, there are potentially wasted resources —exploring whether an idea works in a compromised situation might not be worthwhile if it cannot even succeed in its best light. Alternatively, we may also want to avoid the risk of dropping valuable ideas too quickly that could work in some settings, only because our initial effort at implementing them in certain settings was sub-par. In other words, there are good reasons why we might take the easy piloting steps first, and slowly grow to tackle the more challenging situations. A useful trade-off thus emerges: if an idea is in Region I of Exhibit 4, do we gamble by putting efficacy

estimates in impatient hands when we know it is questionable whether it will work in realistic states of the world?

This consideration leads to a related second issue. Some academics might view my scaling concerns as missing the real point, which in their opinion is that academics' research is *not used for policy purposes enough*. My view is if that statement is true, then it might have arisen because our ideas have experienced such dramatic voltage effects in the past, leading policymakers to lose confidence in our results. To regain trust, we must be sure to advance scalable policies that satisfy the 5Vital signs. Under my approach, we are not slowing down progress and putting up new barriers for scientists, we are creating a systematic research agenda that in the long run will lead to more science being used by policymakers because it is more likely to work at scale.

A third area of concern is potential inequities. Demanding greater information from researchers in early stages might lead to an inequitable distribution of scientific discoveries, journal pages, and subsequent grant funding should this approach only be possible by well-funded researchers. This concern represents a key resource call to funders. Such a call could be answered in many forms. For example, it might yield critical investments in shared R&D infrastructure to move ideas from Regions I and II in Exhibit 4 to Regions III and IV. By eliminating the need for researchers to produce their own CHECCs by allowing them to share infrastructure can considerably change the economics of knowledge creation. Alternatively, funders, such as the NSF, NIH, and private foundations could recognize that current data generation is not optimal to produce scaling insights. They could create funds that recognize HFIDs, and the need for them to generate data via scaling treatment arms from the beginning. Likewise, research organizations such as J-PAL, Resources for the Future, and other well-funded organizations might have the ability to do this within their strategic agenda.

At this point, a useful consideration is where medical trials fit in Exhibit 4. I trust that impatience served as a key impetus for why regulators created safeguards to avoid decisionmakers releasing new drugs too early. Thus, by design, regulators have restricted decisionmakers in a manner that has been absent for policymakers enacting social rulemakings (although benefit cost analysis is completed on every major economic

rulemaking in the U.S.[101]). In addition, for drug testing, once the high fixed costs have been absorbed in the innovation stage to create novel compound recognition, generating new data to explore dosage levels, frequencies, or treatment heterogeneities in the diffusion stage does not have high fixed cost. Thus, by regulatory fiat and the economics of the problem, one can argue that the current landscape dictates that the modal medical trial is in Region I, where my approach has the least bite.

What my perspective makes clear is that the promise of the scientific method within the social sciences will not be realized until we understand that grasping the interrelationships of factors in field settings is only the beginning. We must then seek to understand the mechanisms underlying those relationships, and whether more distant phenomena have the same underlying structure. Together, such information provides insights on whether the causal impacts of treatments implemented in one environment transfer to other environments, be them spatially, temporally, or scale differentiated. Until policy-based evidence is gathered in an optimal manner, we will not reap the true rewards of experimentation.

Acknowledgements

Many thanks to Mary Elizabeth Sutherland for guidance, inspiration, and helpful comments on earlier versions of this Perspective. Three reporters provided excellent comments that markedly improved the message of this study. Faith Fatchen and David Franks provided able research assistance. I have no financial or non-financial competing interests.

Cause of Voltage Drop	*Definition*	*Example*
False Positives	It didn't actually work in the first place.	Initial tests of D.A.R.E. in 1985 showed great efficacy, leading it to be scaled broadly; later analyses showed that D.A.R.E. was a false positive[18].
Representativeness of the Sampled Population	It worked for the sampled population, but that population is too different from the population at scale.	First tests of energy conservation programs work well but as program expands the initial findings are overly optimistic because the early experimental subjects are more responsive to treatment than later participants[84].
Spillovers	If an intervention affects groups other than those sampled, then the impact at scale will not be like the impact in the initial test.	Uber attempts to raise driver pay with base fare increases or by adding tipping but spillovers cause each effort to be thwarted[102,103]
Supply Side	Even if benefits persist at scale, if expansion of the program causes costs to rise disproportionately, then "diseconomies of scale" cause a voltage drop.	As expansion occurs at a cheese packing plant, management teams get spread too thin and duplication of tasks occurs, leading to average total cost increases[104].
Representativeness of the Sampled Situation	The intervention worked in a particular situation too different from the world at scale.	A "special moment in time": nudges for vaccine uptake work well early in campaigns, but the efficacy decays[66].

Exhibit 1: Causes of Voltage Drops: The 5Vital Signs

To create policy-based evidence we must understand the various threats to scaling. These are factors that can cause "voltage drops"—when the idea is scaled the benefit-cost promise is not realized. There are five key threats that can cause voltage drops. I denote these as the 5Vital signs, and they include false positives, representativeness of the population, spillovers, the supply side, and representativeness of the situation. Any one of the five threats, or a combination of them, can frustrate scaling, causing our ideas not to have the impact that was promised from the original evidence.

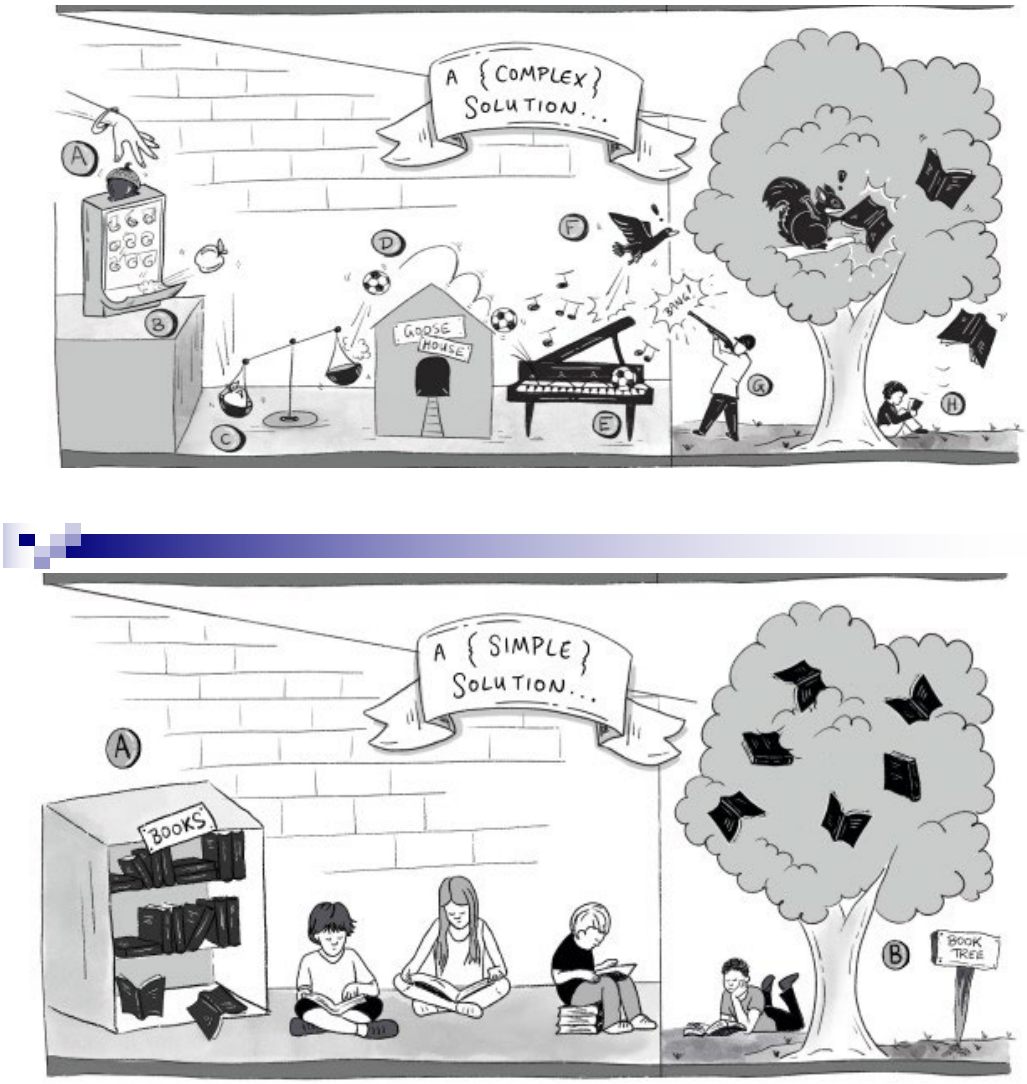

Exhibit 2: It is about the Mediation Path!

The top panel of Exhibit 2 provides a crisp sense of what is necessary for this idea to scale: find a school that has a vending machine that accepts acorns (A) and dispenses apples (B), after which the apple lands on a balance (C), causing a woman's soccer ball to fly through the air over a goose house (D) and onto an organ, which subsequently plays beautiful music (E), which causes the goose to fly out of the classroom window (F), where a hunter notices the goose and unloads his file with no success (G), startling a squirrel who jumps from her nest, causing a book to fall gently into the awaiting arms of a student reader (H). The bottom panel provides an illustration of an idea with a simple mediation path.

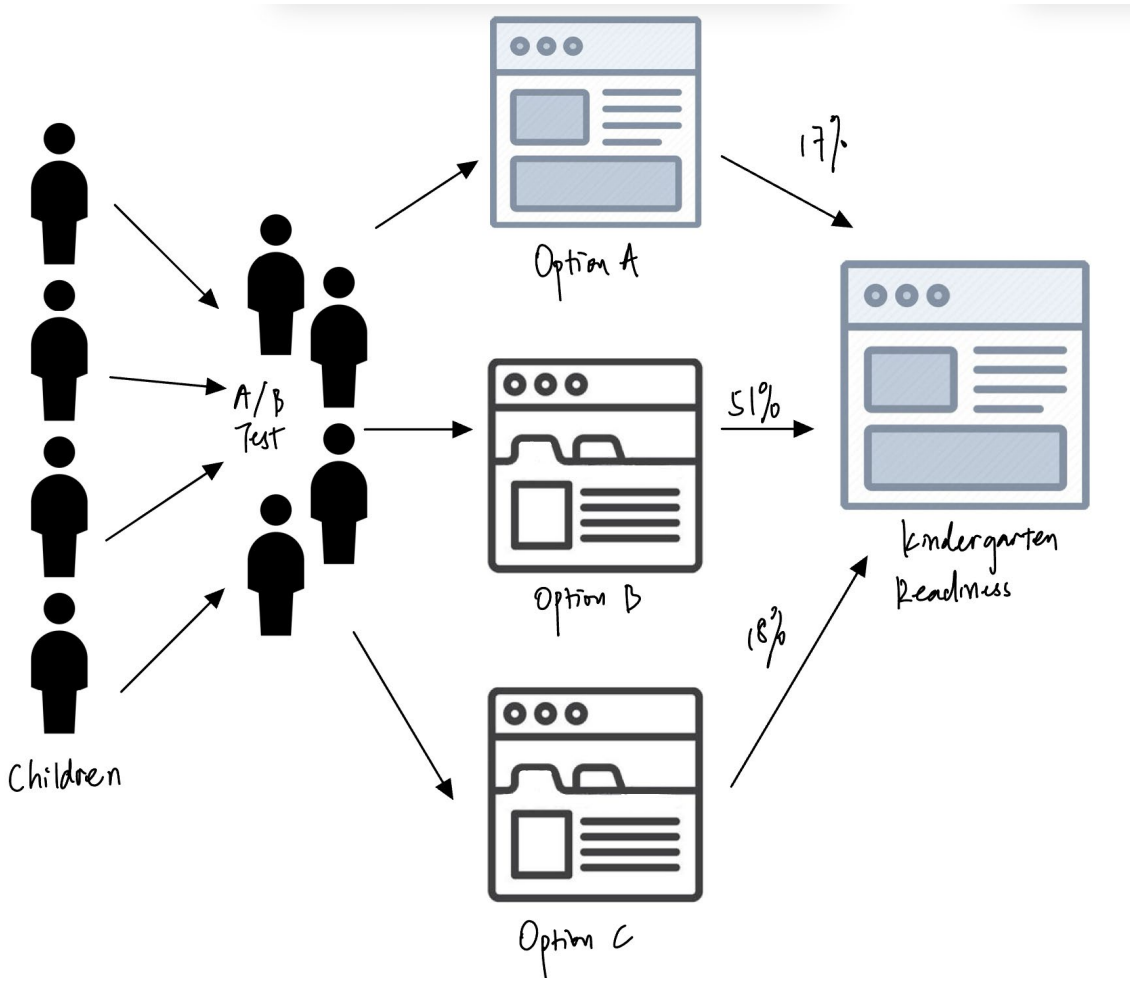

Exhibit 3: Introducing Option C Thinking to the Classic A/B Testing Approach

The cornerstone of the experimental approach in the social sciences is A/B testing. For example, to test whether an education program works, the researcher takes a group of children and splits them into two groups: a control "A" group that does not receive the intervention and the "B" group that receives the early education program. This represents the top two arms of the Exhibit. If the experiment satisfies the classical identification assumptions[42], then a causal effect can be recovered. In this example, the program moves Kindergarten Readiness from 17% (Option A: control) to 51% (Option B: treatment).

If this is a typical social science experiment, however, then the A/B test has likely been conducted as an *efficacy test*. In this manner, we are answering the wrong question if we are attempting to give policy advice with the A/B test. In such a case, we are asking: can this idea work in the best-case situation rather than will this idea work at scale? There are many situations where the economics and science demand a new approach. I refer to this approach as Option C Thinking: create the evidence that provides greater scaling confidence in the original design *alongside* the efficacy test. To complete our thought experiment, in this case, after doing so we find that the program that we would scale raises Kindergarten Readiness to 18%--only a 1 percentage point lift relative to control. Accordingly, when the program is conducted with the constraints that it will face at scale, the program likely does not pass a benefit-cost test, even though it looked incredible in the efficacy test.

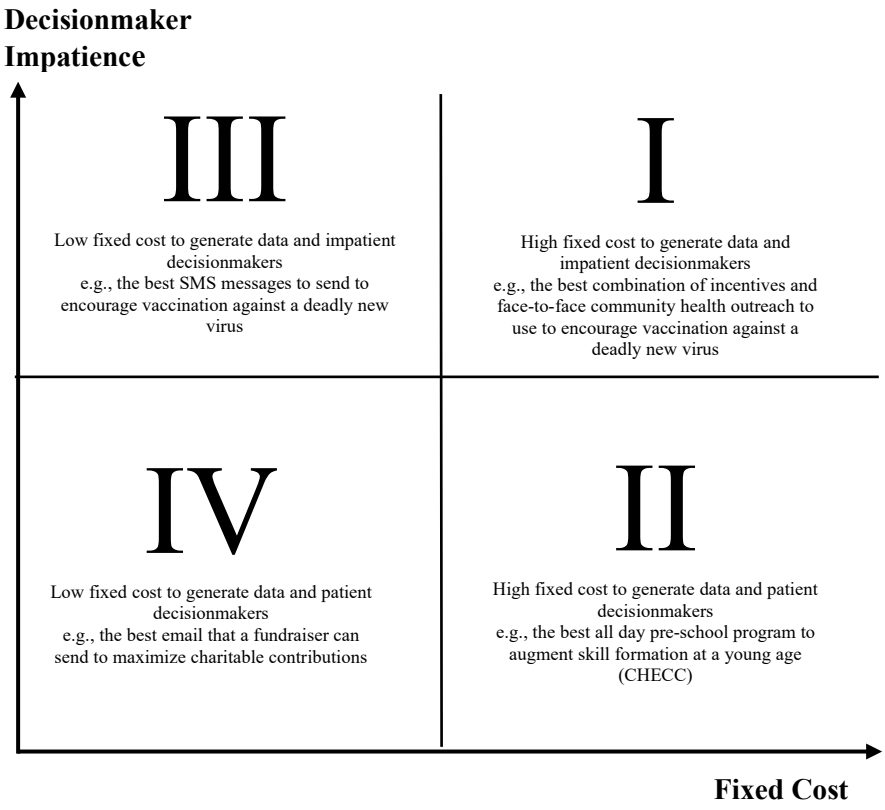

Exhibit 4: Two Dimensions that Affect Optimal Data Generation Sequencing

What types of ideas are most important for the researcher to backward induct? *Ceteris paribus*, the strongest case can be made for ideas in Region I. These are programs that are characterized as high fixed cost programs to generate information and are in topic areas where decisionmakers are impatient. Regions II, III, and IV relax one or both elements. If the idea falls into one of these regions, then I view the prudent action as considering the benefits and costs of flipping the traditional research model. In many such instances, introducing Option C Thinking in the beginning will be optimal. One example is Quadrant IV: with adequate sample sizes, it would be quite simple and cost-effective to explore mechanisms and relevant tests of scale *within* the original charitable giving discovery process.

A useful analogy naturally arises. In engineering, a common approach is that you reverse-engineer the testing from the desired use cases, which usually involve the product working reliably under many different conditions. For instance, when engineering a plane, it would not make sense to follow the typical A/B testing method of starting with a test in a wind tunnel (paired with analogous computer simulations), and then attempting a short journey—say from Chicago to Indianapolis—and then if that works you assume the plane can fly around the world. Instead, you would engineer the plane for all conditions and backwards map the wind tunnel experiments onto how you would want the plane to perform under all possible conditions. Then you would go out and test-fly it, eventually pushing the engineered envelope of performance (which by design has been set beyond the parameters of expected use). After necessary tweaking, you release to the real world. While this is not the current state of art in the social sciences, I argue that it should feature prominently in our toolkit.

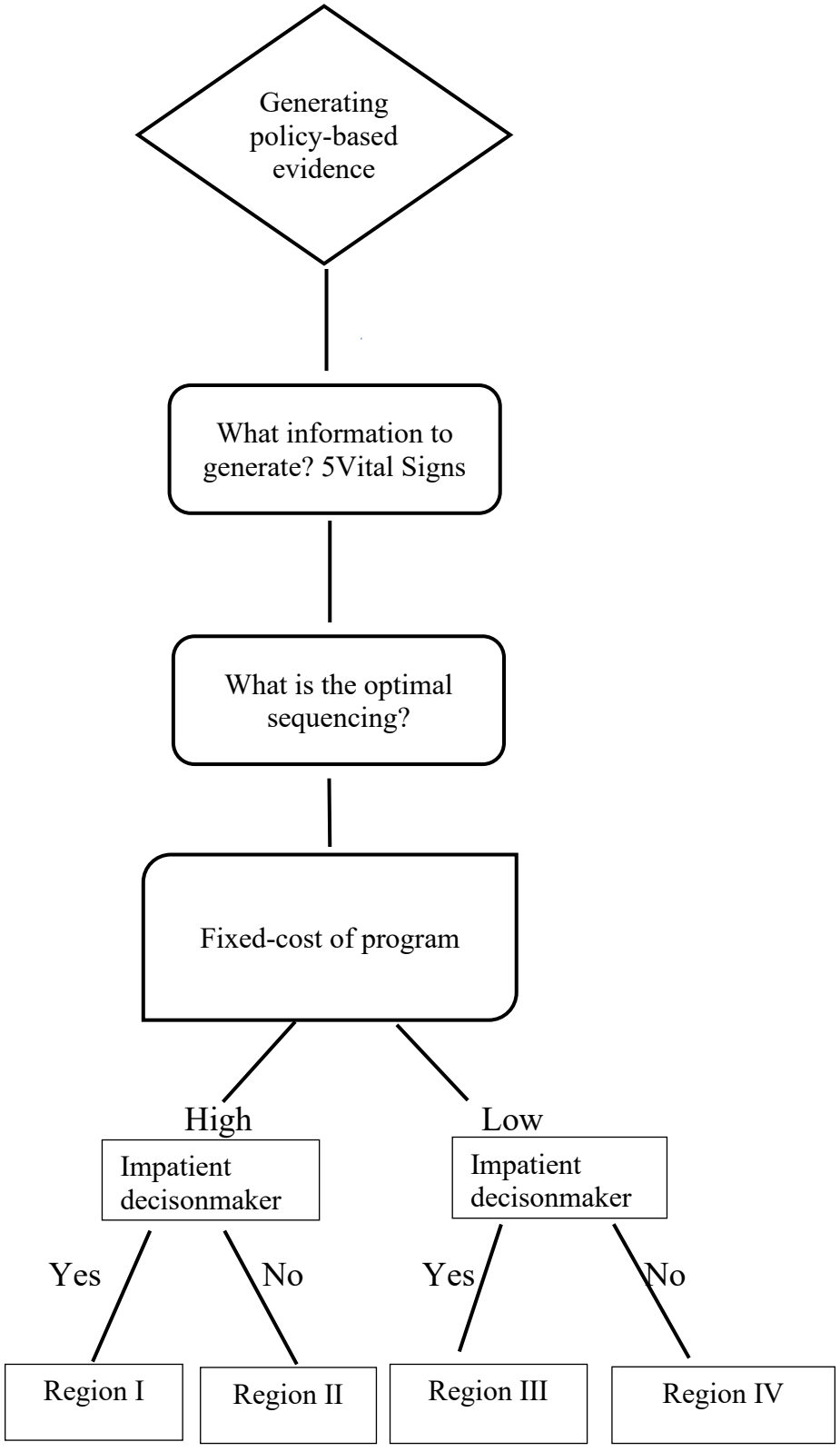

Exhibit 5: Overarching Idea
My general idea of creating policy-based evidence can be viewed through a simple tree diagram, which shows the two dimensions of import and optimal sequencing.

References

1. McCall, W. A. *How to measure in education*. (Macmillan, 1922).
2. Gosnell, H. F. *Getting out the vote*. (University of Chicago Press Chicago, 1927).
3. Fisher, R. A. The Design of Experiments. in *The Design of Experiments* (Oliver and Boyd, 1935).
4. Lewin, K. Field Theory and Experiment in Social Psychology: Concepts and Methods. *American Journal of Sociology* **44**, 868–896 (1939).
5. Smith, V.L. An Experimental Study of Competitive Market Behavior, *Journal of Political Economics* **70**, 111–37 (1962).
6. Harrison, G. W. & List, J. A. Field Experiments. *Journal of Economic Literature* **42**, 1009–1055 (2004).
7. List, J.A., Homo Experimentalis Evolves. *Science* **321**, 207-09 (2008).
8. List, J. A. The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field. *The Quarterly Journal of Economics* **119**, 49–89 (2004).
9. Al-Ubaydli, O. & List, J. A. How natural field experiments have enhanced our understanding of unemployment. *Nature Human Behaviour* **3**, 33–39 (2019).
10. Banerjee, A. V., Duflo, E., Glennerster, R. & Kothari, D. Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *BMJ* **340**, c2220 (2010).
11. Ostrom, E. *Governing the commons: The evolution of institutions for collective action*. (Cambridge university press, 1990).

12. List, J. A. The Market for Charitable Giving. *Journal of Economic Perspectives* **25**, 157–80 (2011).
13. DellaVigna, S., List, J. A. & Malmendier, U. Testing for Altruism and Social Pressure in Charitable Giving. *The Quarterly Journal of Economics* **127**, 1–56 (2012).
14. DellaVigna, S., List, J. A., Malmendier, U. & Rao, G. Estimating Social Preferences and Gift Exchange at Work. *American Economic Review* **112**, 1038–74 (2022).
15. Halperin, B., Ho, B., List, J. A. & Muir, I. Toward An Understanding of the Economics of Apologies: Evidence from a Large-Scale Natural Field Experiment. *The Economic Journal* **132**, 273–298 (2022).
16. Levitt, S.D. & List, J.A. Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review*, **53**, 1-18 (2009).
17. Mobarak, A. M. Assessing Social Aid: the Scale-Up Process Needs Evidence, too. *Nature* **609**, 892–894 (2022).
18. List, J. A. *The Voltage Effect: How to make Good Ideas Great and Great Ideas Scale*. (Currency, 2022).
19. Al-Ubaydli, O., List, J. A. & Suskind, D. L. What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results. *American Economic Review* **107**, 282–86 (2017).
20. Al-Ubaydli, O., List, J. A., LoRe, D. & Suskind, D. Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature. *Journal of Economic Perspectives* **31**, 125–144 (2017).

21. Al-Ubaydli, O., List, J. A. & Suskind, D. 2017 Klein Lecture: The Science of Using Science: Toward an Understanding of the Threats to Scalability. *International Economic Review* **61**, 1387–1409 (2020). *International Economic Review* **61**, 1387–1409 (2020).
22. Al-Ubaydli, O., Lee, M. S., List, J. A., Mackevicius, C. L. & Suskind, D. How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy* **5**, 2–49 (2021).
23. Straight Talk on Evidence, https://www.straighttalkonevidence.org/2018/03/21/how-to-solve-u-s-social-problems-when-most-rigorous-program-evaluations-find-disappointing-effects-part-one-in-a-series/ (2018).
24. Brandon, A., Clapp, C. M., List, J. A., Metcalfe, R. D. & Price, M. The Human Perils of Scaling Smart Technologies: Evidence from Field Experiments. *National Bureau of Economic Research Working Paper Series* No. 30482, (2022).
25. Raikes, H. *et al.* Involvement in Early Head Start home visiting services: Demographic predictors and relations to child and parent outcomes. *Early Childhood Research Quarterly* **21**, 2–24 (2006).
26. Shapley, H. Of Stars and Men; the Human Response to an Expanding Universe, Washington Square Press (*(1964)*.
27. Newton (1687, p. 398, 1966)
28. Brunswik, E. Perception and the representative design of psychological experiments. (1956; 2nd ed.). Berkeley: University of California Press.
29. Campbell, D. T., & J. C. Stanley. Experimental and quasi-experimental designs for research. Chicago: Rand McNally & Company, (1963).

30. Al-Ubaydli, O., & List, J. A. On the Generalizability of Experimental Results in Economics. In Frechette, G. & Schotter, A., Methods of Modern Experimental Economics, Oxford University Press, (2013).
31. List, J. A. Non Est Disputandum de Generalizability? A Glimpse into the External Validity Trial, National Bureau of Economic Research WP #27535 (2020).
32. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspect Psychol Sci* **7**, 615–631 (2012).
33. Jennions, M. D. & Møller, A. P. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* **14**, 438–445 (2003).
34. Camerer, C. F., A. Dreber, E. Forsell, T. Ho, J. Huber, M. Johannesson, M. Kirchler, et al. Evaluating Replicability of Laboratory Experiments in Economics. *Science* 351 (6280): 1433–36. https://doi.org/10.1126/science.aaf0918 (2016).
35. Camerer, C. F., A. Dreber, F. Holzmeister, T. Ho, J. Huber, M. Johannesson, M. Kirchler, et al. Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2 (9): 637–44. https://doi.org/10.1038/s41562-018-0399-z (2018).
36. List, J. A., C. D. Bailey, P. J. Euzent, & T. L. Martin. Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior. Economic Inquiry 39(1):162–170 (2001).
37. Dreber, A. *et al.* Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci USA* 112, 15343 (2015).

38. Benjamin, D. J. *et al.* Redefine statistical significance. *Nature Human Behaviour* **2**, 6–10 (2018).

39. Butera, L. & List, J. A. An Economic Approach to Alleviate the Crises of Confidence in Science: With an Application to the Public Goods Game. (2017) doi:10.3386/w23335.

40. Buck, S. Policy-Based Evidence Doesn't Always Get it Backward, Arnold Ventures, https://www.arnoldventures.org/stories/when-policy-based-evidence-is-exactly-what-is-needed (2019).

41. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLOS Medicine* **2**, e124 (2005).

42. List, J. A. Experimental Economics: Theory and Practice. (2024) Chicago: University of Chicago Press.

43. Maniadis, Z., Tufano, F. & List, J. A. One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *The American Economic Review* **104**, 277–290 (2014).

44. Reed, W. R. A Primer on the 'Reproducibility Crisis' and Ways to Fix It. *Australian Economic Review* **51**, 286–300 (2018).

45. Butera, L., Grossman, P. J., Houser, D., List, J. A. & Villeval, M.-C. A New Mechanism to Alleviate the Crises of Confidence in Science-With An Application to the Public Goods Game. (2020) doi:10.3386/w26801.

46. Maniadis, Z., Tufano, F. & List, J. A. To Replicate or Not to Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study. *The Economic Journal* **127**, F209–F235 (2017).

47. Cleave, B. L., Nikiforakis, N. & Slonim, R. Is there selection bias in laboratory experiments? The case of social and risk preferences. *Exp Econ* **16**, 372–382 (2013).
48. Doty, R. L. & Silverthorne, C. Influence of menstrual cycle on volunteering behaviour. *Nature* **254**, 139–140 (1975).
49. Rosenthal, R. & Rosnow, R. L. *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's classic books*. (Oxford University Press, 2009).
50. Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* **17**, 776–783 (1962).
51. Henrich, J. *et al.* In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review* **91**, 73–78 (2001).
52. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav Brain Sci* 61–83 (2010).
53. Henrich, J., Heine, S. J. & Norenzayan, A. Most people are not WEIRD. *Nature* **466**, (2010).
54. Fehr, E. & List, J. A. The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among Ceos. *Journal of the European Economic Association* **2**, 743–771 (2004).
55. Levitt, S. D. & List, J. A. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives* **21**, 153–174 (2007).

56. Hotz, J. V., Imbens, G. W. & Mortimer, J. H. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* **125**, 241–270 (2005).

57. Kern, H. L., Stuart, E. A., Hill, J. & Green, D. P. Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness* **9**, 103–127 (2016).

58. Yeager, D. S. *et al.* A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019).

59. Yeager, D. S., Krosnick, J. A., Visser, P. S., Holbrook, A. L. & Tahk, A. M. Moderation of classic social psychological effects by demographics in the U.S. adult population: New opportunities for theoretical advancement. *Journal of Personality and Social Psychology* **117**, e84–e99 (2019).

60. Yeager D. S. *et al.* Teacher Mindsets Help Explain Where a Growth-Mindset Intervention Does and Doesn't Work. *Psychol Sci* **33**, 18–32 (2022).

61. Y Tipton, E. *et al.* Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness* **7**, 114–135 (2014).

62. Rudolph KE, Schmidt NM, Glymour MM, Crowder R, Galin J, Ahern J, Osypuk TL. Composition or Context: Using Transportability to Understand Drivers of Site Differences in a Large-scale Housing Experiment. *Epidemiology*. Mar;29(2):199-206 (2018).

63. Miguel, E. & Kremer, M. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* **72**, 159–217 (2004).

64. List, J. A., Momeni, F. & Zenou, Y. Are Measures of Early Education Programs Too Pessimistic? Evidence from a Large-Scale Field Experiment. NBER Working Paper (2019).
65. Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. (Project Gutenberg, 1776).
66. Rabb, N. *et al.* Evidence from a statewide vaccination RCT shows the limits of nudges. *Nature* **604**, E1–E7 (2022).
67. Heller, S. B. *et al.* Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. *The Quarterly Journal of Economics* **132**, 1–54 (2017).
68. Bhatt, M. P., Guryan, J., Ludwig, J. & Shah, A. K. Scope Challenges to Social Impact. *National Bureau of Economic Research Working Paper* #**28406**, (2021).
69. Bettinger, E. P., Long, B. T., Oreopoulos, P. & Sanbonmatsu, L. The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment. *The Quarterly Journal of Economics* **127**, 1205–1242 (2012).
70. Bird, K. A. *et al.* Nudging at scale: Experimental evidence from FAFSA completion campaigns. *Journal of Economic Behavior & Organization* **183**, 105–128 (2021).
71. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour* **5**, 980–989 (2021).
72. List, J. A. On the Interpretation of Giving in Dictator Games. *Journal of Political Economy* **115**, 482–493 (2007).

73. List, J. A. The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy* **114**, 1–37 (2006).
74. Walton, G. M. & Yeager, D. S. Seed and Soil: Psychological Affordances in Contexts Help to Explain Where Wise Interventions Succeed or Fail. *Curr Dir Psychol Sci* **29**, 219–226 (2020).
75. Szaszi, B. *et al.* No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences* **119**, e2200732119 (2022).
76. Holland, P. W. Statistics and Causal Inference. *null* **81**, 945–960 (1986).
77. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* **210**, 2–21 (2018).
78. List, J. A., Pernaudet, J. & Suskind, D. L. Shifting parental beliefs about child development to foster parental investments and improve school readiness outcomes. *Nature Communications* **12**, 5765 (2021).
79. List, J. A. & Shogren, J. F. Calibration of the difference between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior & Organization* **37**, 193–205 (1998).
80. Walton, G. M. *et al.* Where and with whom does a brief social-belonging intervention promote progress in college? *Science* **380**, 499–505 (2023).
81. Blackwell, L. S., Trzesniewski, K. H. & Dweck, C. S. Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention. *Child Development* **78**, 246–263 (2007).

82. Banerjee, A. V., Cole, S., Duflo, E. & Linden, L. Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics* **122**, 1235–1264 (2007).

83. Banerjee, A. *et al.* From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives* **31**, 73–102 (2017).

84. Allcott, H. Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics* **130**, 1117–1165 (2015).

85. Graham, J. R., Harvey, C. R. & Rajgopal, S. The economic implications of corporate financial reporting. *Journal of Accounting and Economics* **40**, 3–73 (2005).

86. Davies, R., Haldane, A. G., Nielsen, M. & Pezzini, S. Measuring the costs of short-termism. *Journal of Financial Stability* **12**, 16–25 (2014).

87. Laverty, K. J. Economic "Short-Termism": The Debate, The Unresolved Issues, and The Implications for Management Practice and Research. *AMR* **21**, 825–860 (1996).

88. Marginson, D. & McAulay, L. Exploring the debate on short-termism: a theoretical and empirical analysis. *Strategic Management Journal* **29**, 273–292 (2008).

89. Caplin, A. & Leahy, J. The Social Discount Rate. *Journal of Political Economy* **112**, 1257–1268 (2004).

90. Stern, N. Stern review: the economics of climate change. (2006).

91. Dasgupta, P. Discounting climate change. *Journal of Risk and Uncertainty* **37**, 141–169 (2008).

92. Weitzman, M. L. On Modeling and Interpreting the Economics of Catastrophic Climate Change. *The Review of Economics and Statistics* **91**, 1–19 (2009).

93. Banerjee, A., Barnhardt, S. & Duflo, E. Can iron-fortified salt control anemia? Evidence from two experiments in rural Bihar. *Journal of Development Economics* **133**, 127–146 (2018).

94. Fryer, R. G., Levitt, S. D., List, J. A. & Samek, A. Towards an understanding of what works in preschool education. *Work in progress* (2017).

95. Fryer, J., Roland G., Levitt, S. D., List, J. A. & Samek, A. Introducing CogX: A New Preschool Education Program Combining Parent and Child Interventions. Working Paper at https://doi.org/10.3386/w27913 (2020).

96. Charness, G., List, J. A., Rustichini, A., Samek, A. & Van De Ven, J. Theory of mind among disadvantaged children: Evidence from a field experiment. *Journal of Economic Behavior & Organization* **166**, 174–194 (2019).

97. Andreoni, J. *et al.* Toward an understanding of the development of time preferences: Evidence from field experiments. *Journal of Public Economics* **177**, 104039 (2019).

98. Andreoni, J., Di Girolamo, A., List, J. A., Mackevicius, C. & Samek, A. Risk preferences of children and adolescents in relation to gender, cognitive skills, soft skills, and executive functions. *Journal of Economic Behavior & Organization* **179**, 729–742 (2020).

99. Cappelen, A., List, J., Samek, A. & Tungodden, B. The Effect of Early-Childhood Education on Social Preferences. *Journal of Political Economy* **128**, 2739–2758 (2020).

100. Islam, A., List, J.A., Vlassopoulos. M., Zenou, Y. Early Childhood Education, Parent Social Networks, and Child Development, mimeo.

101. List, J. A. Field experiments: A Bridge Between Lab and Naturally-Occurring Data. *The BE Journal of Economic Analysis & Policy* **6**, (2007).

102. Hall, J. V., Horton, J. J. & Knoepfle, D. T. Pricing in Designed Markets: The Case of Ride-Sharing. *National Bureau of Economic Research Working Paper* (2021).

103. Chandar, B., Gneezy, U., List, J. A. & Muir, I. The Drivers of Social Preferences: Evidence from a Nationwide Tipping Field Experiment. *National Bureau of Economic Research Working Paper* #26380, (2019).

104. Acemoglu, D., Laibson, D. I. & List, J. A. *Economics*. (Pearson, 2017).