NBER WORKING PAPER SERIES

THE SCIENCE OF USING SCIENCE: TOWARDS AN UNDERSTANDING OF THE THREATS TO SCALING EXPERIMENTS

Omar Al-Ubaydli
John A. List
Dana Suskind

Working Paper 25848
http://www.nber.org/papers/w25848

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

This paper represents the research supporting List's Klein Prize presentation. Seminar participants at Osaka University provided excellent comments that improved the study. We wish to also thank Masaki Aoyagi, Nava Ashraf, Marianne Bertrand, Emir Kamenica, and Dean Karlan for helpful comments. Min Sok Lee and Claire Mackevicious provided excellent research support and comments. Affiliations: Al-Ubaydli: Bahrain Center for Strategic, International and Energy Studies; Department of Economics and the Mercatus Center, George Mason University; College of Industrial Management, King Fahad University of Petroleum and Minerals; List: Department of Economics, University of Chicago; NBER; Suskind: Department of Surgery, University of Chicago Medicine.

The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments
Omar Al-Ubaydli, John A. List, and Dana Suskind
NBER Working Paper No. 25848
May 2019
JEL No. C9,C90,C91,C92,C93,D03

ABSTRACT

Policymakers are increasingly turning to insights gained from the experimental method as a means of informing public policies. Whether—and to what extent—insights from a research study scale to the level of the broader public is, in many situations, based on blind faith. This scale-up problem can lead to a vast waste of resources, a missed opportunity to improve people's lives, and a diminution in the public's trust in the scientific method's ability to contribute to policymaking. This study provides a theoretical lens to deepen our understanding of the science of how to use science. Through a simple model, we highlight three elements of the scale-up problem: (1) when does evidence become actionable (appropriate statistical inference); (2) properties of the population; and (3) properties of the situation. We argue that until these three areas are fully understood and recognized by researchers and policymakers, the threats to scalability will render any scaling exercise as particularly vulnerable. In this way, our work represents a challenge to empiricists to estimate the nature and extent of how important the various threats to scalability are in practice, and to implement those in their original research.

Omar Al-Ubaydli
Department of Economics and Mercatus Center
George Mason University
omar@omar.ec

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and NBER
jlist@uchicago.edu

Dana Suskind
Department of Surgery
University of Chicago Medicine
5841 South Maryland Avenue
MC 1035
Chicago, IL 60637
dsuskind@surgery.bsd.uchicago.edu

1. INTRODUCTION

In the past several decades, experimental methods have evolved from an academic curiosum to a bona fide contributor to scientific knowledge in the social sciences. For their part, economists have generated data from nearly every corner of the world to lend insights into economic theories, such as how markets can be improved, and how public and private organizations can enhance their decision-making (Levitt and List, 2009). Indeed, in most governmental circles, evidence-based programs were once an aspirational goal, then became the gold standard, and now they are the expectation. Even in the most polarized political landscapes, evidence-based policymaking has received widespread bipartisan support.[2] Such a development is entirely reasonable—what policymaker would argue that their policies *should not* be based on scientific evidence?

This naturally leads to an important normative question: if policymakers demand scientific knowledge, then as academics how should we optimally supply our insights? In economics, the tradition of scholarship informing policy decisions arguably goes back to the father of modern economics, Adam Smith, whose most celebrated treatise tackled the issue of how to make people wealthier. Today, improving living standards is considered a core goal for governments, as economists use both theory and empirical work to inform policy. Indeed, data generation via the experimental method has grown to the point that policymakers now expect such wisdom to guide their program choice. Nevertheless, as experimentalists, we have focused almost exclusively on *how best to generate data to explore intervention effects and disentangle mechanisms*. This represented a logical first step, as experimentalists sought to provide deeper empirical insights and theoretical tests as part of the credibility revolution of the 1990s.

Yet, what has been lacking is a scientific understanding of how to make optimal use of the scientific insights generated. In particular, *how should we use the experimental insights for policy purposes*? We denote this as the "scale-up" problem, which revolves around several important questions: do the research results scale to larger markets and settings? When we scale the intervention to broader and larger populations, should we expect the same level of efficacy that we observed in the small-scale setting? If not, then what are the important threats to scalability? What can the researcher do from the beginning of their scholarly pursuit to ensure eventual scalability?

Providing answers to such questions is necessary because understanding when, and how, our experimental insights scale to the broader population is critical to ensuring a robust relationship between scientific research and policymaking. Without such an understanding, empirical research can quickly be undermined in the eyes of the policymaker, broader public, and the scientific community itself. Indeed, in modern economies the chain connecting initial research discovery to the ultimate policy enacted has as its most susceptible link an understanding of the science of how to use science for policy purposes.

[2] For example, the foundations for evidence based policymaking act recently passed with broad bipartisan support in the U.S.: https://www.congress.gov/bill/115th-congress/house-bill/4174/text.

For implementation scientists, some of the issues that we discuss are timeworn. While the implementation science literature is deep, it typically revolves around examining the "voltage effect"—the conjecture that treatment effect sizes observed in research studies diminish substantially when the program is rolled out at larger scale (Kilbourne et al., 2007; Weiss et al., 2014; Supplee and Meyer, 2015; Supplee and Metz, 2015; Gottfredson et al., 2015; Cheng et al., 2017; Al-Ubaydli et al., 2017b). The literature has used this cautionary tale to stress that the voltage effect can severely undermine the optimism advertised in the original research. We suspect that this is one reason why policymakers are slow to adopt and implement the vast amount of science available.

The implementation science literature primarily focuses on fidelity as a key to the voltage problem. For example, consider Head Start home visiting services, an early childhood intervention that found significant improvements in multiple child and parent outcomes in the original research study (Paulsell et. al, 2010). However, variation in quality of home visits was found at larger scale, with home visits for 'at risk' families involving more distractions and less time on child-focused activities, diminishing program effectiveness and increasing attrition (Raikes et al., 2006; Roggman et al., 2008). In this case, the voltage effect likely occurred because the scaled program did not include the fundamental core components that made the initial intervention promising.

While fidelity of the original research study at scale is certainly important, the richness of the economic environment surrounding most of our interventions calls for a more holistic approach to studying the scale-up problem. In our previous research (Al-Ubaydli et al., 2017a), we classified the threats into three bins. First, statistical inference—when is evidence actionable? Second, representativeness of the experimental population—what incentives are in place for researchers to choose a representative subject pool? Third, representativeness of the experimental situation—what situational features are important threats to scalability?[3]

To lend insights into the statistical inference problem, we start with Maniadis et al. (2014), who present a simple model of the inferential problem faced by scholars interpreting initial findings in an area of research with multiple researchers (the interested reader should also see the insightful work of Ioannidis, 2005). Considering representativeness of the experimental population, the extent to which the sample that participates in a study is representative of the broader population is a question that is regularly posed by economists seeking to generalize their findings, whether their data are experimental or observational (Campbell and Stanley, 1963; Al-Ubaydli and List, 2015; Deaton and Cartwright, 2018). While both statistical inference and representativeness of the population are important, the focus of much of the voltage effect literature and our previous work (Al-Ubaydli et al., 2017b) has been on representativeness of the experimental situation. This is because the experimental situation is quite rich and includes many important considerations.

[3] Related excellent work in economics includes Banerjee et al. (2017), Mobarak et al. (2017), Muralidharan and Niehaus, (2017), and Davis et. al. (2017).

Several overarching insights for both policymakers and academics are highlighted by our theoretical framework. A first discussion point is that our model changes the discussion from one that exclusively focuses on the benefit side (voltage effect literature) to a broader metric that includes benefits and costs (BC hereafter). In practice, since in the US every proposed rulemaking that is economically significant has to undergo a formal BC analysis, this change makes sense because many policymakers are attracted to policies that they expect will provide the greatest benefit to the population within time, money, and resource constraints. As such our preferred metric is BC, and if that changes at scale, we classify this as a manifestation of the scale-up effect, because benefits and/or costs change as scale changes.

A second discussion point for policymakers is detailing *why* this might happen, and present guidelines for a proactive approach policymakers and researchers can take to tackle this issue. In terms of the statistical inference bin, our model highlights that there can be a BC drop due to two inferential channels: false positives, and selection of the sampled population by the researcher. For the first channel, in the short run, the expected false positive problem becomes more severe as researcher competition intensifies, a result at odds with intuition. This is because while the results for any given researcher has error that is unconditionally zero on average, the same is not true of the program delivering the largest treatment effects. This leads to the bias strictly increasing in the number of scientists competing in the short run (this is similar to the intuition behind the winner's curse in the auction literature).

The second bin—selection of the research population—occurs in our model when the researcher strategically chooses populations that yield the largest treatment effects. This is a strategic effect due to, for example, publication bias (top journals prefer large treatment effects to small ones, ceteris paribus) or because the researcher is attempting to maximize sample size subject to a fixed budget constraint (a cost-savings, or experimental power effect). This latter relationship holds because experimental participants who expect relatively larger treatment effects may be more willing to select into the experiment, and therefore require less compensation. Our model therefore predicts that even in the absence of strategic effects due to publication bias, a BC drop emerges because scientists exploit heterogeneity as a means of saving money (also see Hunt Alcott's work for a related example).

Our first resolution to these effects is simple advice for policymakers: we need more precise statistical summaries and more frequent replication to help address inference problems. One approach to follow is stipulating a post-study probability of at least 0.95 before enacting policies. This will naturally lead to demand for a greater number of replications and a subsequent change in reward structure. In equilibrium, more dollars for replications from funding agencies would be a natural outcome—and one that we would regard as welcome.

Viewed through the lens of our model, a positive externality of this increased demand for replications is that researchers will place more weight on replicability vis-à-vis cost savings, leading to a smaller strategically-induced bias, and a smaller BC drop in equilibrium. This helps

to reduce a threat to scalability because researchers can take preemptive steps to avoid inadvertently suffering from choosing a non-representative sample.

In terms of our third bin, representativeness of the situation, several insights fall out of the theory. First, negative (positive) network effects and diseconomies (economies) of scale both cause a BC drop (increase). The network effect occurs through the benefit side whereas the economies of scale effect work through the supply side; on cost side considerations of scaling, please see the excellent work of Davis et al. (2017). Second, consonant with the literature, our model showcases that understanding fidelity holds great promise in deepening our knowledge of the threats to scalability (see, e.g., August et al., 2006; Raikes et al., 2006; Roggman et al., 2008; Hippel and Wagner, 2018). Unpacking our production technology reveals several predictions involving fidelity:

- The core components, or 'non-negotiables,' of the intervention should be detailed before scaling to ensure program drift is minimized when implementation takes place.
- Fidelity is increased if facilitators understand the "whys," or the mechanism behind the intervention effect.
- Technology should be used whenever possible, ceteris paribus.
- It is optimal to have the original scientist play an important role in the actual roll out of the program at scale.[4]

What is clear from the literature is that the empirical import of these, and many other important features of the environment, on the scale-up effect are ill-understood. This leads to a generic call to scholars: much like our experimental designs block on features of the population such as age, gender, and race, we should also block on situational features (i.e., scale, inputs, correct dosage, correct program, correct delivery, incentives, substitutes) to not only learn about the intervention, but to also learn about the effects of the environment on our results. In this manner, blocking on situations helps to determine scientifically the core program components when the program is scaled.

A general lesson from our theoretical exercise is that the scholar should backward induct when setting up the original research plan to ensure accurate and swift transference of programs to scale. In this way, even in the case of the insoluble components of the scalability problem, such as upward-sloping supply curves for administrator quality, understanding the source allows scholars to acknowledge it upfront.

A corollary for the policymaker is that when programs are actually scaled, we should take the correct approach to measuring efficacy of the actual implemented program at scale. We prefer using an experiment at scale to measure program effects, but if that is untenable the policymaker should ensure that another appropriate measurement approach (DID, regression discontinuity, or

[4] The innovative work of Ashraf et al. (2017; 2018) using a "cogeneration of knowledge" model that they have implemented in Zambia to explore recruiting of nurses and teaching young women negotiating skills are illustrations of gains to this approach.

the like) can be used. In this manner, an empirical hierarchy for measuring the deliverables at scale should be a policy priority. Overall, for policymakers, this next step demands that they understand the interplay between the research environment and implementation needs necessary at scale and can pinpoint when the threats manifest themselves in the research study. We believe that this type of give and take between scholars and policymakers can yield less "government by guesswork" (Baron, 2018), and more cogeneration of knowledge (Ashraf et al., 2017; 2018).

2. THE SCALING PROBLEM

Before delving into our model, we begin by providing a brief set of examples that help to motivate the three overarching bins in our theory. One interesting example of scaling too quickly is summarized in the work of Hitchcock et al. (2011). Over a series of scientific replications of a Collaborative Strategic Reading (CSR) intervention in 5 different districts in Oklahoma and Texas, Hitchcock et al. (2011) find that overall the program has no discernible effect on reading and comprehension. In essence, the CSR program that showed initial promising results was not effective in other states, providing stark indication of how broad replication before scaling up can prevent wide implementation of ineffective programs.

In this same spirit, Banerjee et al. (2017) describe a study that found no effect of fortified salt on anemia rates, despite earlier programs that found fortified salt reduced anemia rates. They posit that this occurred because original studies specifically sought out adolescent women, and targeting adolescent women in earlier studies led to a measured treatment effect that did not manifest at a larger scale with a broader population. More broadly, Heckman et al. (1998) discuss selection into field experiments, and find that the characteristics of subjects who participate can be distinctly different from those of subjects who do not participate. All of this implies that the measured treatment effect of a small-scale program that compares treatment and control participants from a different population than the set of individuals who participate at scale will not accurately represent the true effect of the program when scaled.

Concerning properties of the situation, August et al. (2006) find that when the situation changed from their initial field study to a broader study, families had reduced engagement in a conduct problems prevention program. This decreased dosage at scale can importantly contribute to the lack of effect in a larger implementation. Likewise, after promising initial results from the Tennessee STAR randomized state-wide class size reduction, Tennessee implemented Project Challenge to reduce class size in k-3 classrooms in the state's poorest school districts (von Hippel and Wagner, 2018). Following an influx of money designated to reducing class sizes, those poorest districts did not actually spend the money to decrease average class sizes. Unsurprisingly, Project Challenge did not result in higher test scores. Indeed, von Hippel and Wagner (2018) find that the average class size decreased from 26 to 25 and overall test scores did not improve. Project Challenge is an example of the entirely wrong program being implemented at scale.

On the cost side, California's statewide implementation of smaller class sizes demonstrated diseconomies of scale in implementation costs (Achilles et al., 1993). Jepsen and Rivkin (2009) examine results of the implementation that forced the state of California to hire from a larger teacher labor market than ever before. To achieve the smaller class sizes, California could have incurred greater costs to maintain similar-quality teachers or continue to pay a similar amount but for lower quality teachers. They find that "the increase in the share of teachers with neither prior experience nor full certification dampened the benefits of smaller classes, particularly in schools with high shares of economically disadvantaged, minority students." When the state of California expanded teacher hiring, they hired less experienced teachers, and the large-scale outcomes of the statewide class size reduction were significantly smaller than the original Tennessee STAR findings.

Alternatively, when exploring scaling up School-Wide Positive Behavioral Interventions and Supports (SWPBIS), Horner et al (2013) explicitly acknowledge that "as states gained local training, coaching, and evaluation capacity, the cost of SWPBIS implementation became less expensive per school and more feasible for scaling up on a geographically distributed level." As the program expanded, costs decreased, or displayed economies of scale. Clearly, on the cost side features of the situation need to be understood to accurately predict whether there will be cost side advantages or disadvantages at scale.

The scaling issue has not gone unnoticed by policymakers, as President Clinton observed that "Nearly every problem has been solved by someone, somewhere. The frustration is that "we can't seem to replicate [those solutions] anywhere else." Echoing this sentiment on the cost side, Larry Summers quipped: "When we use evidence from small interventions to advocate significantly greater public expenditure, we must recognize that we will run into some combination of diminishing returns and higher prices as we scale up programs. It is difficult to quantify this decrease in benefits and increase in costs, but almost certainly, large-scale programs will have lower rates of return than those measured for small-scale programs (see Davis et al., 2017, for these and other quotes).

2.1. THE MODEL

Our model revolves around three main players. The government desires to implement programs that work at scale considering both benefits and costs (BC).[5] Scientists desire to report both replicable findings and important treatment effects. The populace maximizes utility, but we focus on experimental participation for simplicity. With this backdrop, there is a new, proposed intervention, which we refer to as the "program." Scientists are studying the program, while the government is following the research findings with an eye on implementing the program. The

[5] The model can easily be extended to consider the problem of scaling within firms, for which the problem would then revolve around the scale up problem within firms, and non-profit and for profit firms (see, e.g., Lange et al., 2007; Hong et al., 2018) would explore scaling.

program leads to a direct per capita treatment effect T, and it has a per capita cost C. The per capita net treatment effect, which measures the program's impact net of costs, is:

$$\tau = T - C$$

What follows is a simple model designed to explore the scale up effect, that is, changes in the magnitude of τ when moving from the research setting to population-wide implementation.

Let $i \in \{1,2,\dots,N_I\}$ denote the member of the population. Let $S_T \subseteq \{1,2,\dots,N_I\}$ be a set denoting who receives treatment, and let $n_T = |S_T|$, the number of people treated. Let e denote the effort exerted to ensure that the treatment is administered correctly, and let q denote the resultant administration quality.

We define the direct treatment effect of the program on person i as:

$$T_i = f^T\big(w(n_T), b(n_T), q(e, n_T)\big)\bar{T} + \alpha_X X_i$$

$$f^T\big(w(n_T), b(n_T), q(e, n_T)\big) > 0, \bar{T} \sim \big(\mu_T, \sigma_{\bar{T}}^2\big), \alpha_X \geq 0, X_i \sim (0, \sigma_X^2)$$

Therefore, the direct treatment effect is heterogeneous. $\bar{T}$ is a common component of the direct treatment effect, the effect of which is mediated by a strictly positive non-stochastic function f^T, whose arguments we explain below; and $X_i \sim (0, \sigma_X^2)$ is an IID person-specific component, the effect of which is mediated by a weakly positive non-stochastic parameter α_X. We assume that there is at least one person in the population for whom the idiosyncratic component of the treatment effect is zero: $\exists\, i \in \{1,2,\dots,N_I\}: X_i = 0$. We normalize this person to being the first participant, $i = 1$.

We further assume that $\bar{T}$ and X_i are mutually independent for all i. The function f^T has three arguments: within-treatment spillovers (w), between-treatment spillovers (b), and administration quality (q). They affect the common component of the direct treatment effect as follows:

$$f_w^T > 0, f_b^T < 0, f_q^T > 0$$

Each argument is a function of the number of people treated, n_T.

Within-treatment spillovers refer to positive $(w' > 0)$ or negative $(w' < 0)$ spillover effects among the treated group. For example, mobile phones have positive network externalities, meaning that the value of possessing a phone increases with the number of users. Thus, when the intervention is assigning a mobile phone, and the outcome variable is some measure of its utility to the user, increasing the number treated leads to a larger common direct treatment effect. This would suggest that the research study under-estimates the treatment effect at scale.

Between-treatment spillovers refer to positive $(b' > 0)$ or negative $(b' < 0)$ spillover effects from the treated group to the control group, noting that they affect f^T negatively. For example, List et

al. (2018) finds that when evaluating their pre-K intervention, children in the control group who live in the same neighborhood with many treated children have better outcomes than control children who live in neighborhoods with fewer treated children. Thus, all else equal, when taken to scale, the pre-K program should be expected to have larger treatment effects than those observed in the research study.

Administration quality refers to the extent of adherence to the prescribed treatment plan, which includes dosage fidelity. Similar to within- and between-treatment effects, it is a function of the number of people treated, n_T; however, it is also a function of the effort exerted to deliver administrative quality, e. Crucially, as the number of treated rises, the concomitant rise in implementation complexity generates inevitable, organic, random errors. This leads to an attenuation of the common component of the direct treatment effect, in a similar manner to the effect of measurement error in a conventional regression.

Key Assumption 1: As sample size increases, the common component of the direct treatment effect weakly decreases due to a decline in administration quality, $q_{n_T}(e, n_T) \leq 0, f_q^T(w, b, q) > 0$.

We further elaborate on these three effects below.

2.2. Program Costs

Program costs are divided into participation costs and implementation costs.

$$C = P + M$$

The former refers to the cost of inducing people to participate in, and comply with, the program; while the latter is a portmanteau for all remaining costs, including material and administrative costs.

Starting with the participation costs, the cost of compelling person i to enroll and comply is:

$$P_i = f^P(n_T) + \alpha_P p(X_i)$$

$$f^P(n_T) \geq 0, \alpha_P \geq 0, E[p(X_i)] = p(X_1) = 0$$

Thus, the heterogeneous participation cost has a fixed, common component $f^P(n_T)$, where f^P is a non-stochastic function of the number of people being treated. This captures the possibility of economies of scale $\left(f_{n_T}^P < 0\right)$ or diseconomies of scale $\left(f_{n_T}^P > 0\right)$ in participation costs.

There is also an idiosyncratic component of participation costs, $p(X_i)$, which is a function of the idiosyncratic direct treatment effect.

Key Assumption 2: Idiosyncratic participation costs fall as the idiosyncratic direct treatment effect rises, $p'(X_i) < 0$.

This assumption reflects the idea that convincing people to participate in the program, and ensuring their compliance, is easier/cheaper the larger the expected treatment effect. That is, the larger the expected benefits that accrue to that individual. For example, when conducting a trial of the effect of a cancer drug, in principle, getting people who expect the treatment will work to enroll will be easier than getting people to enroll who expect the drug will have minimal effects.

The medical literature features significant support for this assumption. Meta studies of recruitment confirm that those who stand to benefit most from a medical treatment are more likely to participate in trials. For example this disparity is particularly acute in HIV, where confidentiality concerns make recruitment very difficult: those who have reached a stage where they must do something to deal with an advanced stage of the disease, and who are therefore potentially the biggest beneficiaries exhibit much greater readiness to participate (Lovato et al., 1997). In the review due to Cooper et al. (2015), recruitment for medical treatments for type 2 diabetes was significantly easier than for prevention interventions, due to the size, tangibility, and immediacy of the effects of the former. While factors such as altruism and the desire to save money are important determinants of an individual's readiness to participate in a medical trial, surveys also indicate that the perceived benefits are critical, often because prospective participants assume that the medical treatment in a medical trial is of higher quality than conventional, non-experimental treatment (Walsh and Sheridan, 2016).

Subsumed within this assumption is the issue of attrition: just as people with higher treatment effects are more likely to participate at the start, they are also more likely to maintain their participation until the experiment's conclusion. We do not model attrition explicitly, simply because adding it does not yield significant insights beyond those we already offer below.

Without loss of generality, we normalize the expectation of the idiosyncratic term to zero. The non-stochastic parameter α_P captures the importance of the idiosyncratic component.

Turning to the non-participation costs, the implementation cost for i is:

$$M_i = f^M(e, n_T)$$

The implementation cost depends upon the effort exerted in the pursuit of administrative quality, $f_e^M > 0$. The function f_M also captures the possibility of economies of scale $\left(f_{n_T}^M < 0\right)$ or diseconomies of scale $\left(f_{n_T}^M > 0\right)$ in implementation costs.

2.3. The Net Treatment Effect

In light of the above, we can express the net treatment effect for person i, which is the direct treatment effect netting out implementation costs, as follows:

$$\tau_i = f^T\big(w(n_T), b(n_T), q(e, n_T)\big)\bar{T} + \alpha_X X_i - f^P(n_T) - \alpha_P p(X_i) - f^M(e, n_T)$$

For the purposes of this model, this difference structure is our chosen form of BC. Although more generally, we do not take a stand on the form of the BC. We have observed in practice policymakers using the simple BC difference and the BC ratio (the ratio of a program's benefits to its costs). A higher BC difference or ratio indicates a more cost-effective program (Davis et al., 2017).

2.4. Estimation

A scientist conducts an experiment on a participant i, and obtains an estimate of the direct treatment effect:

$$\hat{T}_i = T_i + \varepsilon_i$$

$$\varepsilon_i \sim (\mu_\varepsilon, \sigma_\varepsilon^2)$$

Where ε_i is estimation error (we discuss the statistical sources of estimation error below).

Key Assumption 3: A scientist conducting an experiment on participant i observes the idiosyncratic component of the direct treatment effect, X_i.

This assumption captures the idea that scientists conducting a study have access to more detailed information regarding the unique attributes of the participants, compared to other scientists and other parties who were not involved in the experiment.

Given the non-stochasticity of the functions and parameters, this can be used to estimate the net treatment effect:

$$\hat{\tau}_i = \hat{T}_i - f^P(1) - \alpha_P p(X_i) - f^M(e, 1)$$

$$= f^T\big(w(1), b(1), q(e_1, 1)\big)\bar{T} + \alpha_X X_i + \varepsilon_i - f^P(1) - \alpha_P p(X_i) - f^M(e_1, 1)$$

2.5. Defining the Scaling Problem

If the program is implemented population wide, then conditional on the homogenous component of the direct treatment effect, $\bar{T}$, the net treatment effect for i will be:

$$\tau_i = f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right)\bar{T} + \alpha_X X_i - f^P(N_I) - \alpha_P p(X_i) - f^M\left(e_{N_I}, N_I\right)$$

And on average, this will equal:

$$E(\tau_i) = f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right)\bar{T} - f^P(N_I) - f^M\left(e_{N_I}, N_I\right)$$

If this average is compared to the scientist's estimate from participant i, denoted $\hat{\tau}_i$, then the difference will be:

$$\Delta = E(\tau_i) - \hat{\tau}_i$$

The scaling problem refers to the possibility that Δ is non-zero. If we decompose it into its constituent parts, we have:

$$\begin{aligned}\Delta = &\left[f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right)\bar{T} - f^P(N_I) - f^M\left(e_{N_I}, N_I\right)\right] \\ &- \left[f^T\left(w(1), b(1), q(e_1, 1)\right)\bar{T} + \alpha_X X_i + \varepsilon_i - f^P(1) - \alpha_P p(X_i) - f^M(e_1, 1)\right] \\ = &\underbrace{\left\{\left[f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right) - f^T\left(w(1), b(1), q(e_1, 1)\right)\right]\bar{T}\right\}}_{\delta_1} - \underbrace{\{\alpha_X X_i\}}_{\delta_2} - \underbrace{\{\varepsilon_i\}}_{\delta_3} \\ &- \underbrace{\{f^P(N_I) - f^P(1)\}}_{\delta_4} + \underbrace{\{\alpha_P p(X_i)\}}_{\delta_5} - \underbrace{\left\{f^M\left(e_{N_I}, N_I\right) - f^M(e_1, 1)\right\}}_{\delta_6}\end{aligned}$$

Therefore, there are six possible sources of the scale-up problem:

1. Spillover and administration quality impacts direct treatment effects.
2. The participant being unrepresentative of the population in terms of direct treatment effect.
3. The statistical estimation error.
4. Economies/diseconomies of scale in participation costs.
5. The participant being unrepresentative of the population in terms of participation cost.
6. Economies/diseconomies of scale in implementation costs.

Note that they may cancel each other out, as the sign of each component is indeterminate *ex ante*. This highlights that the pessimism within the scaling literature might be correct, and these terms sum to make Δ negative. Or, it is ill-conceived and too pessimistic because it tends to focus on one slice of the scaling problem and the sum actually makes Δ positive. Within the equation for Δ, Components 1, 4, and 6 are non-stochastic sources of the scaling problem, while 2, 3, and 5 are stochastic sources. We analyze each in turn to provide a deeper intuition of the causes and consequences of each.

We use the term "voltage effect" to refer to the scaling effect arising exclusively in the benefit (treatment effect) side of the equation:

$$\Delta_V = \delta_1 + \delta_2 + \delta_3$$

As aforementioned, the implementation science literature often discusses voltage drop phenomena, whereby observed treatment effects shrink when programs are scaled from the research setting to the population at large. This is sometimes referred to as the scaling effect. To avoid confusion between the voltage effect, and the gross scaling effect, which also considers the cost side of the equation according to our definition, we use a new term, "scale-up" effect, which refers to changes in the *net* treatment effect resulting from changes in scale. We use it for the remainder of this paper, hereby suspending our use of the term scaling effect/problem.

2.6. Scale-Up Effect Taxonomy

As mentioned in the introduction, in our previous work (Al-Ubaydli et al., 2017a), we identified three sources of the scale-up effect: statistical inference; representativeness of the experimental population; and representativeness of the experimental situation. The six sources identified in section 2.5 are not perfectly nested in the tripartite classification, because some of the six sources cut across the three categories. However, both classifications are exhaustive.

For the remainder of this paper, we develop a new classification based on an analysis of the model. In particular, we distinguish between two primary sources of the scale-up effect: non-stochastic sources, which reflect structural and deterministic properties of production functions; and stochastic sources, which cover statistical selection effects and inferential errors.

3. Non-Stochastic Sources of the Scale-Up Effect

3.1. Spillovers and Administration Quality in the Direct Treatment Effect

The first component in Δ_V is composed of three subcomponents, reflecting spillover and administrative quality effects.

$$\delta_1 = \left[f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right) - f^T\left(w(1), b(1), q(e_1, 1)\right)\right]\bar{T}$$

The first subcomponent is the within-treatment spillover effect:

$$w(N_I) \neq w(1), f_w^T > 0$$

Some interventions, such as mobile telephone usage, or literacy, have strong positive spillovers: treating people creates a positive treatment externality on the remainder of the population. One could imagine that early tests of the effects of the use of Facebook could produce quite small treatment effects but as a greater number of people were enrolled in treatment the direct treatment effect increased substantially.

Others suffer from the reverse, especially those that involve ordinal-based payoffs. For example, if the intervention under investigation involves assisting a student in obtaining higher school grades, or an airline decreasing its check-in time, part of the treatment effect may be a ranking effect that does not replicate as a larger number of people is treated: only 5% of students can be in the top 5% of students, and only one airline can have the fastest check-in time, which it can use in an advertising campaign.

The second subcomponent is the between-treatment spillover effect.

$$b(N_I) \neq b(1), f_b^T < 0$$

In some interventions, treating people creates a spillover effect on the untreated. This can be positive—consider the case of a business ethics course where enrollment is assigned to a random subset of a company's employees. Those who do not enroll are still affected positively by the presence of the enrollees, who act as models for them: the greater the number of enrollees (treated group), the smaller the implied treatment effect when comparing treated to untreated in the research study. One recent example of the potential import of this effect is found in List et al. (2018), who report that in their measurement of the effects of a pre-K intervention, control children can gain more than 0.5 standard deviations in cognitive test scores because of treated neighbors.

Alternatively, it could be negative: consider an intervention that improves the school performance of students in a given class. The control group in the same class may, upon seeing an initial improvement in the performance of the treated group, feel demoralized, inducing a further deterioration in their performance, and accentuating the treatment effect.

In both within- and between-treatment spillover effects, we perceive no general rules of thumb regarding which is more likely. We merely note that it can cause a non-zero scale-up effect and empirical measurement of such impacts is important for future research.

In addition to within- and between-treatment spillover effects, there is also the possibility of spillovers from the treated group to people who are not even participating in the experiment, i.e., people beyond the control group. For example, if a small-scale natural field experiment in an Indian village involves giving the participants large amounts of money that exceed daily wages by several orders of magnitude, then the village's macroeconomy may fundamentally change because of the experiment. If these changes then feedback on the treatment and control groups, then the result can be further scale-up effects with an indeterminate sign. For example, inflation from the monetary expansion might diminish the real increase in income experienced by those treated. We do not explicitly model this class of spillover effect; instead, we merely alert readers to its existence (the interested reader should see Banerjee et al. (2017), Muralidharan and Niehaus, (2017), and the citations therein).

The third subcomponent is the administrative quality effect:

$$q\left(e_{N_I}, N_I\right) \neq q(e_1, 1), f_q^T > 0$$

Administering a treatment because more difficult as scale increases, meaning that when increasing the scale, project administrators must exert higher levels of per capita administrative effort to maintain quality. When this does not occur, the treatment effect will shrink:

$$e_{N_I} = e_1 \Rightarrow q\left(e_{N_I}, N_I\right) \leq q(e_1, 1)$$

This reflects the organic rise in complexity of implementation that results from increasing scale, even when all material and human inputs are increased in proportion. They are a form of managerial diseconomies of scale that is reflected in the quality of the output resulting from the inputs.

One frequently-encountered manifestation is political problems, especially when a novel intervention is being implemented. The prevailing regime brings with it significant entrenched interests, which may oppose a novel intervention on the basis of financial interests, or simply because of institutional inertia. Circumventing the barriers erected by opponents in a small-scale experiment is trivial. Yet, at a larger scale, this may require a significant financial outlay, corresponding to diseconomies of scale. Or, in the absence of those outlays (constant per capita administrative effort exerted), the treatment effect will be denuded by counterattacking bureaucrats and other vested interests.

Notably, the effort that researchers and overseers exert when trying to maintain fidelity sometimes reflects their taking the time to explain to newer administrators the reasoning behind the intervention. There is a large literature showing that people are more likely to adhere to instructions when they understand their purpose, and when those issuing the instructions take the time to ensure that people buy in. A good illustration is patient-adherence to medication—when physicians want to maximize the likelihood that their patients take drugs as prescribed, one of the best practices that is grounded in rigorous experimentation is to explain the way in which the drug works to the patient via face-to-face meetings, and to explain the importance of following the instructions (Zullig et al., 2013).

3.2. Scale-Up Effects in Participation Costs

The fourth component in Δ reflects the possibility of increasing/decreasing returns to scale in the cost of securing participants compared to the experiment conducted by the scientist.

$$\delta_4 = f^P(N_I) - f^P(1)$$

A key cost advantage that governments have in this domain is that they can mandate programs, making participation a requirement. Salient examples include attending primary and secondary education, or wearing a seatbelt while driving.

In addition to the weight of legal backing, operating large programs can radically reduce awareness/marketing costs per unit, either for technological reasons (using a television commercial or a government press release has a low per unit cost), or because of positive spillovers, as the more that people talk about a program, the more others become aware of it. Further, monitoring and compliance costs per unit can decline as scale rises, again primarily as the result of technological factors.

However, participation/compliance costs in general may also suffer from diseconomies of scale for the usual laundry list of reasons. For example, at small scales, word-of-mouth can be a cost-free way of generating awareness or marketing a program, but it is insufficient at higher scales. Moreover, scientists operating in academic environments can often secure participants and can monitor compliance at costs that are unusually low, which we expand upon in the next section on scale effects in implementation costs.

More generally, we do not take a definitive stance on whether economies or diseconomies of scale dominate in participation costs, merely noting that both possibilities exist.

3.3. Scale-Up Effects in Implementation Costs

The sixth component in Δ reflects the possibility of increasing/decreasing returns to scale in the cost of implementing the program compared to the experiment conducted by the scientist.

$$\delta_6 = f_M(N_I) - f_M(1)$$

Decreasing returns to scale can sometimes reflect the unusually low costs that scientists can secure due to the unique settings of the academy. Alternatively, many critical inputs may appear "free" according to a scientist's accounts when in fact that they are covered by other sources outside the experimental budget.

An important example is the labor cost of the scientist and research assistants, which will likely not appear in the accounts. Graduate students may often operate as pro bono project managers, since their compensation is a combination of their student stipend and the "reward" of authorship on the resulting paper. Moreover, these graduate students are uniquely qualified in that they are highly intelligent, highly obedient, and fundamentally believe in the mission.

When scaling up, securing implementation staff as competent as the graduate students will likely require significant financial outlays (increasing marginal cost), which is a significant source of diseconomies of scale. The same is potentially true of other key inputs that scientists might be able to secure at a cut-price rate due to the infrastructure that they can access for free as scientists, such as access to rooms/offices in the university, and the ability to meet with important managers due to personal relationships.

Alternatively, conventional economies of scale may apply too, especially those relating to procuring inputs in bulk, or more efficient production processes due to scale. For example, when conducting a mail drive for charitable donations, at a small scale, manual labor will be used, and materials will be purchased at retail rates. At a larger scale, automatic envelope-stuffers can be purchased, and wholesale prices can be accessed for materials.

4. STOCHASTIC SOURCES OF THE SCALE-UP EFFECT

We identified three stochastic sources of the scale-up problem in Δ: heterogeneity in the direct treatment effect, statistical estimation error, and heterogeneity in the participation costs. To understand these mechanisms, we build a simple model of the behavior of scientists, academic journals, and the government.

4.1. PLAYERS

We begin by introducing the preferences/goals of each the three main players, before analyzing predicted behavior.

4.1.1. SCIENTISTS

Let $j \in \{1,2,\dots,N_J\}$ denote the scientist. As above, scientists recruit participants to run experiments. They then estimate the net treatment effect and report it to the world via academic journals. If called upon, they then assist the government in the implementation of the program.

Let $i(j)$ denote the participant that scientist j recruits for an experiment. As above, an experiment on participant i yields an estimate of the direct treatment effect:

$$\hat{T}_{i(j)} = T_{i(j)} + \varepsilon_{i(j)}$$

$$\varepsilon_{i(j)} \sim (\mu_\varepsilon, \sigma_\varepsilon^2)$$

Where $\varepsilon_{i(j)}$ is estimation error. As mentioned above in key assumption 3, when scientist j selects participant i, they know their idiosyncratic direct treatment effect, $X_{i(j)}$. Therefore, the cost to the scientist of running an experiment is:

$$C = f^P(1) + \alpha_P p\left(X_{i(j)}\right) + f^M(\bar{e}, 1)$$

Where we have fixed administration quality effort at $e = \bar{e}$. The reported net treatment effect will be:

$$\hat{\tau}_{i(j)} = \hat{T}_{i(j)} - f^P(1) - \alpha_P p\left(X_{i(j)}\right) - f^M(\bar{e}, 1)$$

$$= f^T\big(w(1), b(1), q(\bar{e}, 1)\big)\overline{T} + \alpha_X X_{i(j)} + \varepsilon_i - f^P(1) - \alpha_P p\big(X_{i(j)}\big) - f^M(\bar{e}, 1)$$

In many cases, scientists explore ways in which to cut costs where possible due to constraints on research budgets. Money saved on a given experiment can be used for running other experiments, so the opportunity cost of research funds can be quite high.

Key Assumption 4: While the idiosyncratic direct treatment effect $X_{i(j)}$ is observable to a scientist running an experiment, it is unobservable to those who see and consume the scientist's reported net treatment effect.

Assumption 4 complements Assumption 3 by adding asymmetric information to the problem.

After doing the research, we assume that reporting results yields three distinct benefits to a scientist. The first is a knowledge-production benefit: a reward for the scientist's contribution to human knowledge.

Key Assumption 5: Ceteris paribus, the scientific community values replicable findings.

$$\pi_K = \overline{K} - \alpha_K\big(\hat{\tau}_{i(j)} - \hat{\tau}_1\big)^2$$

$$\overline{K} > 0, \alpha_K \geq 0$$

$\overline{K}$ represents the scientist's reward, while the latter term is a penalty for non-replicability of findings. Future scientists investigating the original scientist's findings will re-run the experiment with $i = 1$, the person for whom $X_i = 0$ (equivalently, imagine picking X_i randomly and performing many replications), and compare their estimated net treatment effect with the figure originally reported by the scientist. The parameter α_K captures the strength of the penalty for imperfect replicability, a penalty that can be avoided by running the original experiment on $i = 1$.

The second benefit accruing to the scientist from reporting estimated net treatment effects is the prestige from reporting dramatic and eye-catching results.

Key Assumption 6: Ceteris paribus, the scientific community values experiments that report net estimated treatment effects that are large in absolute value.

$$\pi_L = \alpha_L l\big(\hat{\tau}_{i(j)}\big)$$

$$\alpha_L \geq 0, \hat{\tau}_{i(j)} > 0 \Rightarrow l' > 0, \hat{\tau}_{i(j)} < 0 \Rightarrow l' < 0$$

The function l captures the reward for reporting large net treatment effects, while the parameter α_L measures the importance of such rewards. The key assumption is based on the well-documented bias that both professional academics and laypeople suffer, whereby they regard large net treatment effects as more noteworthy. We expand upon this point below when we discuss scientific journals.

The third and final benefit that a scientist reaps when reporting estimated net treatment effects relates to the government's response.

Key Assumption 7: Upon observing a scientist's reported estimated net treatment effect, if the government decides to implement the program at large scale (the level of the population), then the scientist earns material and psychological benefits.

$$\pi_G = \alpha_G g\left(\hat{\tau}_{i(j)}\right)$$

$$\alpha_G \geq 0, g \in [0,1]$$

Where the function g represents the probability that the program is adopted at scale by the government, as a function of the reported estimated net treatment effect. We discuss the nature of this function when we discuss the government. Parameter α_G reflects the importance of this reward from the scientist's perspective.

Thus, a scientist's objective function is:

$$U_S = \pi_K + \pi_L + \pi_G - C$$

$$= \overline{K} - \alpha_K\left(\hat{\tau}_{i(j)} - \hat{\tau}_1\right)^2 + \alpha_L l\left(\hat{\tau}_{i(j)}\right) + \alpha_G g\left(\hat{\tau}_{i(j)}\right) - \left[f^P(1) + \alpha_P p\left(X_{i(j)}\right) + f^M(\bar{e}, 1)\right]$$

Therefore, with this maximization problem in mind, the scientist selects the participant i to realize the following potentially conflicting goals:

1. Maximizing replicability when scientists compare to the situation where $X_i = 0$.
2. Maximizing the estimated net treatment effect to make results eye-catching.
3. Maximizing the likelihood that the government implements the program at scale.
4. Minimizing the cost of experiment.

Note that despite goal (4), for the simplicity of exposition, we are exogenizing the scientist's effort decision regarding administration quality. A more sophisticated model would endogenize it, and, in the event that it is partially unobservable, this may lead to additional scale-up effects. We hope that future research takes on this goal.

4.1.2. SCIENTIFIC JOURNALS

After conducting experiments, scientists submit their estimated net treatment effects to scientific journals for publication. Consumers of scientific journals demand studies that report large net treatment effects. They reward journals via the purchase of subscriptions and by citing the papers within a journal. We treat these two goals as perfectly aligned.

In the interests of parsimony, we do not model the process by which journals compete with each other over papers submitted, and over subscriptions and citations. Instead, we treat journals as a

unitary decision-maker that receives a fixed number of submissions, and must decide how to allocate a fixed prestige/exposure pie among the submissions. We assume that all costs are overhead, i.e., that all possible distributions of prestige/exposure entail the same cost, meaning that the journals' problem is simply choosing the distribution that maximizes subscriptions/citations.

We do not explicitly post or solve this problem in this paper. Rather, to focus on the scale-up effect, we proceed directly to the implied reward function for scientists. Expected rewards accruing to scientist j are:

$$r_j\left(\hat{\tau}_{i(j)}; \hat{\tau}_{i(-j)}\right) \geq 0$$

$$\sum_{j=1}^{N_J} r_j\left(\hat{\tau}_{i(j)}; \hat{\tau}_{i(-j)}\right) = \bar{R}$$

$$\frac{\partial r_j}{\partial \hat{\tau}_{i(j)}} \geq 0, \frac{\partial r_j}{\partial \hat{\tau}_{i(k \neq j)}} \leq 0$$

Where $\hat{\tau}_{i(-j)}$ is the vector of results reported by all scientists except j. $\bar{R}$ denotes the fixed size of the pie. As described in key assumption 6, each scientist's expected reward is increasing in the estimated net treatment effect that they report, while it is decreasing in the estimated net treatment effect reported by others. This is the result of the well-documented bias that journal editors express in favor of studies that report large estimated net treatment effects.

The mechanism embodied by the function r_j can take many forms. For example, it could reflect the largest reported net treatment effects being published in the top journals, or receiving the most citations. Accordingly, the function r_j is equal to π_L from the scientist's objective function. To formally reconcile the two, we express l to be a function of the entire vector of reported net treatment effects.

$$\pi_L = \alpha_L l\left(\hat{\tau}_{i(j)}; \hat{\tau}_{i(-j)}\right)$$

Note that we are assuming journal editors naively interpret the reported findings of scientists, and disregard the underlying source of variation in reported net treatment effects. We do not believe this to be literally true. After all, journal editors are invariably some of the most accomplished scientists themselves. However, we regard this to be a reasonable approximation of what actually occurs in practice for some journals. Ultimately, this may be because the editors' patrons—journal readers and citers—are the ones who obsess over large net treatment effects with insufficient attention to their underlying cause.

4.1.3. THE GOVERNMENT

We treat the government as a surrogate of the general population, meaning that its preferences are identical to those of the representative individual, the one for whom $X_i = 0$, in the event that the program is adopted.

$$U_G = E(\hat{\tau}_i) = f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right)\bar{T} - f^P(N_I) - f^M\left(e_{N_I}, N_I\right)$$

And, it is zero in the event that the program is rejected. The government does not know the true value of the net treatment effect when it makes a decision about adopting the program, and so it must rely on the estimates published in scientific journals.

Key Assumption 8: The government naively reads results reported in the scientist literature; it does not account for the potential non-representativeness of the participants in published studies, estimation bias, economies of scale, spillovers, or administration quality effects.

Thus, it treats $\hat{\tau}_i$ as its best estimate of $E(\hat{\tau}_i)$. While this is an exaggerated characterization of the government's actual naivety when interpreting scientific results, similar to our discussion of scientific journals above, it is likely to be a more accurate representation than assuming government omniscience towards potential sources of inferential bias.

In the event that the program is adopted by the government, it gives prestige/consulting benefits to the scientist responsible for the findings. Similar to the benefits doled out by scientific journals, these are zero-sum, and thus we adjust the scientist's government reward to make it a function of all reported net treatment effects.

$$\pi_G = \alpha_G g\left(\hat{\tau}_{i(j)}; \hat{\tau}_{i(-j)}\right)$$

4.1.4. SUMMARY

Given the behavior of scientific journals and the government, scientist i's problem is to maximize the following with respect to $X_{i(j)}$:

$$U_S = \pi_K + \pi_L + \pi_G - C$$

$$= \bar{K} - \alpha_K\left(\alpha_X X_{i(j)} - \alpha_P p\left(X_{i(j)}\right)\right)^2 + \alpha_L l\left(\hat{\tau}_{i(j)}\right) + \alpha_G g\left(\hat{\tau}_{i(j)}\right) - \left[f^P(1) + \alpha_P p\left(X_{i(j)}\right) + f^M(\bar{e}, 1)\right]$$

Recall that the scale-up problem is equivalent to the following expression being non-zero:

$$\Delta = \left[f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right)\bar{T} - f^P(N_I) - f^M\left(e_{N_I}, N_I\right)\right] - \left[f^T\left(w(1), b(1), q(e_1, 1)\right)\bar{T} + \alpha_X X_i + \varepsilon_i - f^P(1) - \alpha_P p(X_i) - f^M(e_1, 1)\right]$$

$$= \underbrace{\left\{\left[f^T\left(w(N_I), b(N_I), q\left(e_{N_I}, N_I\right)\right) - f^T\left(w(1), b(1), q(e_1, 1)\right)\right]\bar{T}\right\}}_{\delta_1} - \underbrace{\{\alpha_X X_i\}}_{\delta_2} - \underbrace{\{\varepsilon_i\}}_{\delta_3}$$
$$- \underbrace{\{f^P(N_I) - f^P(1)\}}_{\delta_4} + \underbrace{\{\alpha_P p(X_i)\}}_{\delta_5} - \underbrace{\left\{f^M\left(e_{N_I}, N_I\right) - f^M(e_1, 1)\right\}}_{\delta_6}$$

Our focus is on how scientist behavior affects the three terms $(\delta_2, \delta_3, \delta_5)$. We focus initially on (δ_2, δ_5). In particular:

$$\delta_2 - \delta_5 = \alpha_X X_{i(j)} - \alpha_P p\left(X_{i(j)}\right)$$

$$\frac{\partial \Delta}{\partial(\delta_2 - \delta_5)} < 0$$

Therefore, as $\delta_2 - \delta_5 = \alpha_X X_{i(j)} - \alpha_P p\left(X_{i(j)}\right)$ increases, the net treatment effect shrinks.

4.2. PARTICIPANT UNREPRESENTATIVENESS

A non-representative participant pool can be caused by several factors. For example, it could be quite direct, as when the FDA guidance recommended the exclusion of what they defined as "women of childbearing potential" from Phase I and (early) Phase II clinical cancer drug trials in 1977, a policy that was eventually rescinded in 1993.[6] This could certainly lead to skewed results when the cancer drugs are taken to scale, especially in cases of "fast-tracking" that might have left population specific analysis among certain groups as speculative. Our model focuses on a very different purpose for the sampled population to affect scale-up: in the scientific marketplace, researcher incentives dictate a subject pool choice that is more likely to find larger treatment effects than a random sample would support.

Participant unrepresentativeness as a source of the scaling problem is defined as deviations of Δ from zero caused by $X_{i(j)}$ being non-zero, i.e., by scientists using participants who are not representative of the general population.

$$X_{i(j)} = 0 \Rightarrow \delta_2 = \delta_5 = 0 \Rightarrow \delta_2 - \delta_5 = 0$$

$$\frac{\partial(\delta_2 - \delta_5)}{\partial X_{i(j)}} = \alpha_X - \alpha_P p'\left(X_{i(j)}\right) \geq 0$$

$$\frac{\partial \Delta}{\partial(\delta_2 - \delta_5)} = -1 \Rightarrow \frac{\partial \Delta}{\partial X_{i(j)}} \leq 0$$

[6] See https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071682.pdf and https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/WomensHealthResearch/UCM131204.pdf).

To understand why scientists might use unrepresentative participants, we analyze the scientist's objective function and its relationship with the endogenous variable, $X_{i(j)}$, and the parameters $(\alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)$.

$$U_S = U_S\left(X_{i(j)}; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G\right)$$

Let $X^*(\alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)$ denote the solution to the scientist's problem. The first- and second-order conditions are:

$$\frac{\partial U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)}} = 0, \frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)}^2} < 0$$

We perform the traditional comparative statics manipulations by differentiating through the first-order condition and using the second-order condition:

$$\frac{\partial X^*}{\partial \alpha} = -\frac{\frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)} \partial \alpha}}{\frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)}^2}}$$

$$\Rightarrow sign\left(\frac{\partial X^*}{\partial \alpha}\right) = sign\left(\frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)} \partial \alpha}\right)$$

Therefore, if we wish to determine the sign of a comparative static, then we need only calculate the sign of the cross-partial of utility. Moreover, from above, we have:

$$\frac{\partial \Delta}{\partial X_{i(j)}} \leq 0 \Rightarrow sign\left(\frac{\partial \Delta}{\partial \alpha}\right) = sign\left(-\frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)} \partial \alpha}\right)$$

We begin by fully expressing the first-order condition.

$$\left[\alpha_L l'\left(\hat{\tau}_{i(j)}\right) + \alpha_G g'\left(\hat{\tau}_{i(j)}\right) - 2\alpha_K \left(\alpha_X X_{i(j)} - \alpha_P p\left(X_{i(j)}\right)\right)\right]\left(\alpha_X - \alpha_P p'\left(X_{i(j)}\right)\right) - \alpha_P p'\left(X_{i(j)}\right) = 0$$

Rearranging this term and using the fact that all parameters are weakly positive, and the assumptions on the derivatives of the functions l (positive), g (positive), and p (negative), yields:

$$\alpha_X X_{i(j)} - \alpha_P p\left(X_{i(j)}\right) = \frac{1}{2\alpha_K}\left[\alpha_L l'\left(\hat{\tau}_{i(j)}\right) + \alpha_G g'\left(\hat{\tau}_{i(j)}\right) - \frac{\alpha_P p'\left(X_{i(j)}\right)}{\alpha_X - \alpha_P p'\left(X_{i(j)}\right)}\right] \geq 0$$

$$\Rightarrow \delta_2 - \delta_5 \geq 0$$

The left-hand side of the above equation represents the marginal cost of deviating from a representative participant: the penalty caused by non-replicability of the results. The right-hand

side represents the marginal benefit: the sum of the larger treatment effect that is rewarded by scientific journals, the larger treatment effect that is rewarded by the government, and the cost savings.

Remark 1.1: If $\alpha_X = \alpha_P = 0$, then the scientist is indifferent to the choice of X.

Under these conditions, the marginal effect of varying X on the scientist's utility is everywhere zero. Therefore, for the problem to be non-trivial, at least one of α_X and α_P must be non-zero.

Remark 1.2: $\alpha_K > 0$ is a necessary condition for X^* to be less than its maximum possible value.

From the original utility function, setting $\alpha_K = 0$ eliminates the penalty for having an non-representative participant, which pushes the scientist toward picking the largest possible value.

We henceforth assume that $\alpha_X, \alpha_K > 0$. We now turn to the comparative statics.

Result 1.1: Increasing the non-replicability parameter, α_K, diminishes the non-representativeness of the participant, X^*, and decreases the magnitude of the scale-up drop.

Proof:

$$\frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)} \partial \alpha_K} = -2\left(\alpha_X X_{i(j)} - \alpha_P p(X^*)\right)\left(\alpha_X - \alpha_P p'(X^*)\right) < 0$$

$$\Rightarrow \frac{\partial X^*}{\partial \alpha_K} < 0, \frac{\partial \Delta}{\partial \alpha_K} > 0 \blacksquare$$

Intuitively, this is equivalent to increasing the marginal cost of non-representativeness, without affecting the marginal benefit, meaning a decrease in optimal non-representativeness.

Result 1.2: Increasing the parameters denoting the scientific (α_L) or government (α_G) reward for reporting a large net treatment effect increases the non-representativeness of the participant, X^*, and increases magnitude of the scale-up drop.

Proof:

$$\frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)} \partial \alpha_L} = \left(\alpha_X - \alpha_P p'(X^*)\right) l'(\hat{\tau}^*) > 0$$

$$\frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)} \partial \alpha_G} = \left(\alpha_X - \alpha_P p'(X^*)\right) g'(\hat{\tau}^*) > 0$$

$$\Rightarrow \frac{\partial X^*}{\partial \alpha_L}, \frac{\partial X^*}{\partial \alpha_G} < 0, \frac{\partial \Delta}{\partial \alpha_L}, \frac{\partial \Delta}{\partial \alpha_G} < 0 \blacksquare$$

Intuitively, this is equivalent to increasing the marginal benefit of non-representativeness, without affecting the marginal cost, meaning an increase in optimal non-representativeness.

Result 1.3: Increasing the idiosyncratic participation cost parameter, α_P, has an indeterminate effect on the participant's non-representativeness.

Proof:

$$\frac{\partial^2 U_S(X^*;\alpha_X,\alpha_K,\alpha_P,\alpha_L,\alpha_G)}{\partial X_{i(j)}\partial\alpha_P}$$
$$= [2\alpha_K - \alpha_L l''(\hat{\tau}^*) - \alpha_G g''(\hat{\tau}^*)]\big(\alpha_X - \alpha_P p'(X^*)\big)p(X^*)$$
$$- \big[\alpha_L l'(\hat{\tau}^*) + \alpha_G g'(\hat{\tau}^*) - 2\alpha_K\big(\alpha_X X^* - \alpha_P p(X^*)\big)\big]p'(X^*) - p'(X^*)$$

Substituting in from the first-order condition yields:

$$= [2\alpha_K - \alpha_L l''(\hat{\tau}^*) - \alpha_G g''(\hat{\tau}^*)]\big(\alpha_X - \alpha_P p'(X^*)\big)p(X^*) - \frac{\alpha_P p'^2(X^*)}{\alpha_X - \alpha_P p'(X^*)} - p'(X^*)$$

$$= [2\alpha_K - \alpha_L l''(\hat{\tau}^*) - \alpha_G g''(\hat{\tau}^*)]\big(\alpha_X - \alpha_P p'(X^*)\big)p(X^*) - \frac{\alpha_X p'(X^*)}{\alpha_X - \alpha_P p'(X^*)}$$

Since $X^* > 0$, it follows that $p(X^*) < 0$, meaning that the first term in the above expression is negative (assuming concavity of the functions l, g), and a negative term (the second one) is being subtracted from it, yielding an indeterminate sign. ■

Intuitively, changing α_P changes both the marginal cost and the marginal benefit of non-representativeness: it makes the net treatment effect's absolute deviation from the representative case larger, because of the larger saving on idiosyncratic participation costs, which raises the non-replicability while increasing the returns to non-replicability.

Result 1.4: Increasing the idiosyncratic direct treatment effect parameter, α_X, decreases the participant's non-representativeness, and decreases the magnitude of the scale-up drop.

Proof:

$$\frac{\partial^2 U_S(X^*;\alpha_X,\alpha_K,\alpha_P,\alpha_L,\alpha_G)}{\partial X_{i(j)}\partial\alpha_X}$$
$$= \big[\alpha_L l'(\hat{\tau}^*) + \alpha_G g'(\hat{\tau}^*) - 2\alpha_K\big(\alpha_X X^* - \alpha_P p(X^*)\big)\big]$$
$$+ \big(\alpha_X - \alpha_P p'(X^*)\big)(\alpha_L l''(\hat{\tau}^*) + \alpha_G g''(\hat{\tau}^*) - 2\alpha_K X^*)$$

The term in square brackets is negative due to the first-order conditions, while concavity of l and g implies that the latter term is negative.

$$\Rightarrow \frac{\partial^2 U_S(X^*; \alpha_X, \alpha_K, \alpha_P, \alpha_L, \alpha_G)}{\partial X_{i(j)} \partial \alpha_X} < 0$$

$$\Rightarrow \frac{\partial X^*}{\partial \alpha_X} < 0, \frac{\partial \Delta}{\partial \alpha_X} > 0 \blacksquare$$

Similar to α_P, increasing α_X increases both the marginal benefit and the marginal cost of non-representativeness. Yet, unlike changes in α_P, changes in α_X amplify the marginal benefit less than they amplify the marginal cost, because one part of the marginal benefit (cost savings) is unaffected by changes in α_X. Therefore, at the margin, the net effect on non-representativeness is negative.

This result is somewhat paradoxical. In one case, if there is no heterogeneity, then there is no scale-up drop from having an unrepresentative pool of participants. Alternatively, as the degree of heterogeneity increases, the cost of having an unrepresentative pool of participants increases, pushing scientists toward selecting more representative samples. But this is an oversimplification, since there are in fact two sources of heterogeneity: α_X and α_P. And, in the case of the latter, it is possible that increasing heterogeneity leads to a higher scale-up drop, for example when $\alpha_X = 0$.

4.3. Inferential Errors

Returning to the scale-up equation in section 4.1.4, the error in the estimation of the net treatment effect, ε, causes scale-up drop.

$$\frac{\partial \Delta}{\partial \varepsilon} = \frac{\partial \Delta}{\partial \delta_3} < 0$$

What factors might lead to a systematically biased estimation error? The model highlights two inferential channels that imply a voltage drop, $\Delta^* < 0$: researcher white noise term and the sampled population is drawn strategically, $X^*_{i(j)} > 0$. We discussed the latter above, so here we focus on the former.

A first insight is that while the white noise is unconditionally zero on average, $E(\varepsilon_j) = 0$, the same is not true of the "winning" program. Moreover the bias is strictly increasing in the number of scientists competing, J. To understand the intuition, consider the winner's curse, an adverse-selection problem that arises because the winning bidder in a common value auction holds the most overly-optimistic information concerning the value of the auctioned item. As such, bidders must bid more conservatively as the number of competing bidders increases because winning implies a greater winner's curse. The same phenomenon is happening here—as the number of scientists working on related programs increases, the "winning program" will be overly optimistic, leading to an inferential error.

To lend insights into potential solutions, we use the approach described in Maniadis et al. (2014), which investigates false positives (which is closely related to the problem of publication bias (Young et al., 2008). Maniadis et al. (2014) key theoretical result focuses on the concept of a post-study probability (PSP): the probability that a declaration of a research finding, made upon reaching statistical significance, is true. The PSP is defined as follows:

$$PSP = \frac{(1-\theta)\omega}{(1-\theta)\omega + \eta(1-\omega)}$$

Where η is the level of statistical significance, $(1-\theta)$ is the level of power, and ω is the prior. As the exhibits in Maniadis et al. (2014) reveal, even after an initial research proclamation, the PSP can be quite low, implying that naïve policymakers will be making quite dramatic errors if they base important decisions upon such inferences—false positives are important, especially when the empirical results are deemed "surprising" or "large."

Second, the PSP can be raised substantially if the initial positive findings pass as little as two or three independent replications. This is an important insight, because in our experience many decision makers in government and the private sector wish to rush new insights into practice. Proper incentives for independent replication therefore help mitigate two problems:

1. White noise draw leading to adoption.
2. Strategic $X^*_{i(j)}$ draw leading to adoption.

This leads to our proposal concerning inference: *before advancing policies, the PSP should be at least 0.95*.

While of course there is an ad hoc nature of this proposal, in equilibrium, this choice has implications that would permeate various parts of the modeling. For example, it naturally leads to a greater number of replications and a subsequent change in reward structure. In equilibrium, more dollars for replications from funding agencies would be a natural outcome. A positive externality of this increased demand in replications is that researchers will place more weight on replicability vis-à-vis cost savings, leading to a smaller strategically-induced bias, and a smaller BC drop in equilibrium. This helps to reduce a threat to scalability because researchers can take preemptive steps to avoid inadvertently suffering from choosing a non-representative sample.

5. EPILOGUE

Major societal advances will not occur until we revamp the entire system of knowledge discovery for policymaking: from soup to nuts. This involves three major steps. First, we must fund basic research so scientists have the means to carry out credible science. This involves discussions around the philanthropy of science (see, e.g., List, 2011). Second, we must provide the knowledge creation market with the optimal incentives for researchers to design, implement, and report

scientific results. Third, we must develop a system whereby policymakers have the appropriate incentives to adopt effective policies, and once adopted they must develop strategies to implement those policies with rigorous evaluation methods to ensure continual improvement (see, e.g., Komro et al., 2019; Chambers et al., 2013).

We view a firm understanding of all three of these areas as necessary for any modern government to use an evidence based approach. While we provide insights into all three links, our study focuses attention on the middle link: the knowledge creation market. In advancing an economic model of scaling, we highlight the various incentives actors face in this market. By juxtaposing the actors and their various incentives, we provide straightforward insights into the causes of the scale up effect and where and when it is likely to occur. We show that the benefit cost relationship changes at scale simply due to the nature of the incentives in the system. Our framework also features areas where behavioral relationships are known, where more empirical evidence is necessary, and how we can adjust incentives for researchers to provide information in their original research concerning the likelihood of their intervention scaling effectively.

As academics, we often ask why more scholarly research is not implemented into public policy. One argument for the lack of scientifically-driven policies is that the current approach is broken. When going from science to policy we typically follow the traditional formula of documenting effects on small groups over short time-spans and testing their statistical significance. We then ask policymakers to adopt the programs that have large treatment effects. This is because when scaling, we oftentimes generalize our results to both a population of situations and a population of people when we typically only speak to the issue of the latter. Yet, such an empirical approach can be quickly undermined in the eyes of the policymaker, broader public, and the scientific community if the promises of the original research are not delivered.

Our research advocates flipping the traditional model, calling on scholars to place themselves in the shoes of the people whom they are trying to influence. Our call is for policy research that starts by imagining what a successful intervention would look like fully implemented in the field, applied to the entire subject population, sustained over a long period of time, and working as it is expected because its mechanisms are understood.

To accomplish this goal, our original experimental designs must address each of these needs. For example, providing a list of "non-negotiables" is important in that these are features of the program that must be implemented with fidelity. To complete this exercise, our experiments should block on situations when doing experiments just like we block on individual characteristics (i.e., scale, inputs, human's delivering, correct dosage, program, delivery, incentives, substitutes).

One illustration of this idea in action revolves around human capital. If the research study uses 20 classroom teachers but at scale we will need 20,000, then simply hiring the 20 best teachers for the research study is ill-conceived if one has scaling in mind. Rather, much like Fryer et al.'s (2012) approach in their Chicago Heights studies, a broader pool should be considered and then a

random sample chosen from that pool. This is carefully done in Davis et al. (2017) to explore scientifically the effects of this factor on scaling.

Another example of how to use the original research design to provide empirical content to the features of our scaling model is to use multi-site designs optimally (for excellent recent discussions see Raudenbush and Bloom (2015) and Weiss et al., (2017)). In carrying out such an agenda, the analyst can not only measure the average treatment effect, but explore how the treatment effect varies across sites. By using appropriate variation in site specific characteristics, the design of multi-site trials can provide empirical content into why effects might not scale and give empirical hints where more research is necessary before scaling.

While our theory highlights many other reasons why the BC ratio may differ across research studies and programs at scale, empirical work must be completed to determine which pieces of our model have empirical relevance. Measuring the nature and extent of the effects of the non-stochastic and stochastic factors we discuss will usher in a new and innovative way to generate and use experimental data. We hope that this promise will be fulfilled as we strive to enhance the efficacy and usage of evidence-based policies.

References

Achilles, C.M., Nye, B.A, Zaharias, J.B., and Fulton, B.D. (1993), 'The Lasting Benefits Study (LBS) in grades 4 and 5 (1990-1991): A legacy from Tennessee's four-year (K-3) class-size study (1985-1989), Project STAR', Paper presented at the North Carolina Association for Research in Education. Greensboro, North Carolina, January 14, 1993.

Al-Ubaydli, O. and List, J. A. (2015). Do Natural Field Experiments Afford Researchers More or Less Control than Laboratory Experiments? The American Economic Review, 105(5):462–466.

Al-Ubaydli, O., List, J. A., LoRe, D., and Suskind, D. L. (2017a). Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature. Journal of Economic Perspectives, 31(4):125–144.

Al-Ubaydli, O., List, J. A., and Suskind, D. L. (2017b). What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results. American Economic Review, 107(5):282–286.

Ashraf, Nava, Oriana Bandiera and Scott Lee. (2017). "Losing Prosociality In The Quest For Talent? Sorting, Selection, And Productivity In The Delivery Of Public Services." Working paper.

Ashraf, Nava, Natalie Bau, Corinne Low and Kathleen McGinn. (2018). "Negotiating a Better Future: How Interpersonal Skills Facilitate Inter-Generational Investment." Working paper.

August, G., Bloomquist, M., Lee, S., Realmuto, G. and Hektner, G. (2006), 'Can Evidence-Based Prevention Programs be Sustained in Community Practice Settings? The Early Risers' Advanced-Stage Effectiveness Trial', Prevention Science, 7(2): 151-165.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., and Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives, 31(4):73–102.

Baron, J. (2018). A Brief History of Evidence-Based Policy. The ANNALS of the American Academy of Political and Social Science, 678(1):40–50.

Campbell, D. T. and Stanley, J. (1963). Experimental and Quasi-Experimental Designs for Research. Cengage Learning, Boston, 1 edition edition.

Chambers, D.A., Glasgow, R.E., & Strange, K. C. (2013). The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. Implementation Science, 8, Published Online. http://www.implementationscience.com/content/8/1/117. doi: 10.1186/1748-5908-8-117.

Cheng, S., McDonald, E. J., Cheung, M. C., Arciero, V. S., Qureshi, M., Jiang, D., Ezeife, D., Sabharwal, M., Chambers, A., Han, D., Leighl, N., Sabarre, K.-A., and Chan, K. K. (2017). Do the American Society of Clinical Oncology Value Framework and the European Society of Medical Oncology Magnitude of Clinical Benefit Scale Measure the Same Construct of Clinical Benefit? Journal of Clinical Oncology, 35(24):2764–2771.

Cooper, C. L., Hind, D., Duncan, R., Walters, S., Lartey, A., Lee, E., and Bradburn, M. (2015). A rapid review indicated higher recruitment rates in treatment trials than in prevention trials. Journal of Clinical Epidemiology, 68(3):347–354.

Davis, J. M., Guryan, J., Hallberg, K., and Ludwig, J. (2017). The Economics of Scale-Up. Working Paper 23925, National Bureau of Economic Research.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210:2–21.

Foundations for Evidence-Based Policymaking Act of 2017, P. L. N. H. . (2019). Foundations for Evidence-Based Policymaking Act of 2017.

Fryer, Roland G, J., Levitt, S. D., List, J., and Sadoff, S. (2012). Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. Working Paper 18237, National Bureau of Economic Research.

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., and Zafft, K. M. (2015). Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation. Prevention Science, 16(7):893–926.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998), 'Characterizing Selection Bias Using Experimental Data', Econometrica, 66(5): 1017-1098.

Hippel, P. von and Wagner, C. (2018), 'Does a Successful Randomized Experiment Lead to Successful Policy? Project Challenge and What Happened in Tennessee After Project STAR (March 31, 2018). Available at SSRN: https://ssrn.com/abstract=3153503.

Hitchcock, J., Dimino, J., Kurki, A., Wilkins, C. and Gersten, R. (2011). 'The Impact of Collaborative Strategic Reading on the Reading Comprehension of Grade 5 Students in Linguistically Diverse Schools', NCEE 2011-4001, U.S. Department of Education.

Hong, F., Hossain, T., List J.A., and Migiwa Tanaka (2018). "Testing The Theory Of Multitasking: Evidence From a Natural Field Experiment In Chinese Factories," International Economic Review, vol. 59(2), pages 511-536.

Horner R. H., Kincaid, D., Sugai, G., Lewis, T., Eber, L., Barrett, S., Dickey, C. R., Richter, M., Sullivan, E., Boezio, C., Algozzine, B., Reynolds, H. and Johnson, N. (2014), 'Scaling Up School-

Wide Positive Behavioral Interventions and Supports: Experiences of Seven States With Documented Success', Journal of Positive Behavior Interventions, 16(4): 197-208.

Ioannidis, J. (2005). "Why Most Published Research Findings Are False." PLoS Medicine 2 (8): 1418–22.

Jepsen, C. and Rivkin, S. (2009), 'Class Size Reduction and Student Achievement: The Potential Tradeoff between Teach Quality and Class Size', Journal of Human Resources, 44(1): 223-250.

Kilbourne, A. M., Neumann, M. S., Pincus, H. A., Bauer, M. S., and Stall, R. (2007). Implementing evidence-based interventions in health care: application of the replicating effective programs framework. Implementation science: IS, 2:42.

Komro, KA, Flay, B.R., Biglan, A., Wagenaar, A.C., (2019), "Research design issues for evaluating complex multicomponent interventions in neighborhoods and communities," Trans Behav Med., in press.

Lange, A., List, J.A., and M.K. Price, (2007) "Using Lotteries to Finance Public Goods: Theory and Experimental Evidence," International Economic Review, 48(3), pp. 901-927.

Levitt, S. D. and List, J.A. (2009). Field experiments in economics: The past, the present, and the future. European Economic Review, 53(1):1–18.

List, J.A. (2011) "The Market for Charitable Giving," Journal of Economic Perspectives, 25(2): pp. 157-180.

List, J.A., Momeni, F., and Zenou, Y. (2018). "Are Estimates of Early Education Programs Too Pessimistic? Evidence from a Large-Scale Field Experiment that Causally Measures Neighbor Effects." Working paper.

Lovato, L. C., Hill, K., Hertert, S., Hunninghake, D. B., and Probstfield, J. L. (1997). Recruitment for controlled clinical trials: literature summary and annotated bibliography. Controlled Clinical Trials, 18(4):328–352.

Maniadis, Z., Tufano, F., and List, J. A. (2014). One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. American Economic Review, 104(1):277–290.

Mobarak, A. M., K. Levy, M. Reimao, (2017), "The path to scale: Replication, general equilibrium effects, and new settings," *Voxdev.org*, November 21.

Muralidharan, K. and Niehaus, P. (2017). Experimentation at Scale. Journal of Economic Perspectives, 31(4):103–124.

Paulsell, D., Porter, T., Kirby, G., Boller, K., Martin, E. S., Burwick, A., Ross, C., and Begnoche, C. (2010). Supporting Quality in Home-Based Child Care Initiative: Design and Evaluation Options. Technical Report 3887af819cdc4b2e9f0e830c0fd3f97a, Mathematica Policy Research.

Raikes, H., Pan, B. A., Luze, G., Tamis-LeMonda, C. S., Brooks-Gunn, J., Constantine, J., Tarullo, L. B., Raikes, H. A., and Rodriguez, E. T. (2006). Mother-child bookreading in low-income families: correlates and outcomes during the first three years of life. Child Development, 77(4):924– 953.

Raudenbush, S. W., and Bloom, H. S. (2015). Learning About and From a Distribution of Program Impacts Using Multisite Trials. American Journal of Evaluation. doi: 10.1177/ 1098214015600515.

Roggman, L. A., Cook, G. A., Peterson, C. A., and Raikes, H. H. (2008). Who Drops Out of Early Head Start Home Visiting Programs? Early Education and Development, 19(4):574–599.

Supplee, L. and Metz, A. (2015). Opportunities and Challenges in Evidence-based Social Policy. Technical Report V27, 4, Society for Research in Child Development.

Supplee, L. H. and Meyer, A. L. (2015). The Intersection Between Prevention Science and Evidence- Based Policy: How the SPR Evidence Standards Support Human Services Prevention Programs. Prevention Science, 16(7):938–942.

United States. (1977). General Considerations for the Clinical Evaluation of Drugs. DHEW publication ; no. (FDA) 77-3040. U.S. Dept. of Health, Education, and Welfare, Public Health Service, Food and Drug Administration ; for sale by the Supt. of Docs., U.S. Govt. Print. Off., Rockville, Md.: Washington.

U.S. Food and Drug Administration (1993). Guideline for the Study and Evaluation of Gender Differ- ences in the Clinical Evaluation of Drugs; Notice. Federal Register, 58(139):39406–39416.

Walsh, E. and Sheridan, A. (2016). Factors affecting patient participation in clinical trials in Ireland: A narrative review. Contemporary Clinical Trials Communications, 3:23–31.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A Conceptual Framework for Studying the Sources of Variation in Program Effects. Journal of Policy Analysis and Management, 33(3).

Weiss, M.J. Bloom, H.S., Verbitsky-Savitz, N., Gupta, H., Vigil, A.E., and Daniel N. Cullinan (2017) How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials, Journal of Research on Educational Effectiveness, 10:4, 843-876,

Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. (2008). Why Current Publication Practices May Distort Science. PLOS Medicine, 5(10):e201.

Zullig, L. L., Peterson, E. D., and Bosworth, H. B. (2013). Ingredients of successful interventions to improve medication adherence. JAMA, 310(24):2611–2612.