What Can We Learn From Experiments?
Understanding the Threats to the Scalability of Experimental Results

By OMAR AL-UBAYDLI, JOHN A. LIST AND DANA L. SUSKIND *

* Al-Ubaydli: Bahrain Center for Strategic, International and Energy Studies, PO Box 496, Manama, Bahrain (e-mail: omar@omar.ec), and Department of Economics and Mercatus Center, George Mason University. List: University of Chicago, 1126 E. 59th St., Chicago IL, 60637 (e-mail: jlist@uchicago.edu), and NBER. Suskind: The University of Chicago Medicine, 5841 S. Maryland Avenue, MC 1035, Chicago IL, 60637(email:dsuskind@surgery.bsd.uchicago.edu)

Policymakers often consider interventions at the scale of the population, or some other significant scale, and seek to ground their decisions in scientific research. 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. Improving living standards is now considered a core goal for governments, as reflected in the 2016 US presidential election campaign, where Hillary Clinton proposed significant reductions in the cost of preschool, while Donald Trump espoused tax cuts, both as indirect methods of creating a better life for US citizens.

Among the scholarly sources of information about the potential effects of such interventions are experimental studies conducted at a significantly smaller scale, such as programs designed to tackle social problems, such as health, education, employment, and family support issues.

A common occurrence is for such research programs to never be scaled, or when they are scaled the program (treatment) effects diminish substantially in size when applied at the larger scale (this is commonly denoted "voltage drop" in the literature), even though such predictable changes are not accounted for in benefit-cost analysis. We refer to this as the "scalability" problem. Generally, the issue revolves around the query: I just found a 0.2 standard deviation effect in my experiment, should I expect to observe such an effect when scaled to a city, state, or country?

A simple example due to program drift illustrates one set of reasons for the scalability problem. Consider Early Head Start home visiting services, one of the largest federally funded early childhood interventions in the world. The program demonstrated significantly improved school readiness for children aged up to three years old, improved family economic self-sufficiency, and

parenting practices through high-quality efficacy trials (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. Lower proportion of time on child-focused activities and lower parental engagement was associated with diminished effectiveness for both child and parent outcomes and higher dropout rates (Raikes et al., 2006, Roggman et al., 2008).

In choosing such an exploration, we are changing the conversation from the "whys" and "hows" of economic experiments to the *science of using science*. This movement is necessary because learning to understand when, and how, our experimental results scale to the broader population is critical to ensuring a robust relationship between scientific research and policymaking. Without such an understanding, empirical research can be quickly undermined in the eyes of the policymaker, broader public, and the scientific community.

As a first step toward formulating a theory of the science of using science, this paper discusses the ingredients necessary to understand several important threats to scalability. In a companion set of studies, we theoretically model the scaling problem (Al-Ubaydli et al., 2017a), and use that theory to understand the scaling problem in medicine (Al-Ubaydli et al., 2017b).

To characterize scalability and highlight certain relevant threats, we divide the problem into three components: a statistical inference procedure applied to the data gathered, representativeness of the population, and representativeness of the situation. In this study, we walk through each in turn, and provide a glimpse of how economics and experiments can lend insights into their import.

I. Statistical Inference and Scalability

We open with a discussion of the role of erroneous statistical inference in compromising scalability because it is in some sense the most straightforward to model generally, and the easiest to remedy, owing to the fact that it is underlain by factors that are almost completely under the control of the scientific community.

Maniadis et al. (2014) present a simple model of the inferential problem faced by scholars interpreting initial findings in an area of research where multiple researchers are working. Their key theoretical result focuses on the concept of a post-study probability (PSP), which is the probability that a declaration of a research finding, made upon reaching statistical significance, is true. This

can be interpreted as the likelihood that a naïve scholar is *ex post* correct in taking an initial, significant finding at face value. The word "naïve" distinguishes the scholar from scholars deploying rational expectations in their inference.

The authors find that the larger the number of researchers investigating a relationship, the smaller the PSP, implying that competition between independently operating research teams will cause naïve scholars to commit greater inferential errors when interpreting an initial, statistically significant finding.

Drawing further theoretical deductions from the model requires knowledge of parameters that are generally unknown, such as the proportion of associations being investigated that are actually true. However, the authors demonstrate, according to a wide range of plausible parameter values, two key insights.

First, even after an initial research proclamation, the PSP can be quite low. Implying that naïve scholars will be making quite dramatic errors if they base important decisions upon their inferences—false positives are important, especially when the empirical results are deemed "surprising".

Second, the PSP can be raised substantially if the initial positive findings are subjected to—and pass—as little as two independent replications. This is an important insight, because in our experience many decision-makers—governmental policymakers or CEOs of non-profit and for-profit firms—wish to rush new insights into practice.

Continuing with the analogy to rational expectations, naïve scholars' biases can be abusive, just as governments can exploit agents deploying adaptive expectations to force unemployment below equilibrium. Unscrupulous researchers might cherry pick certain results or data, or interpret ambiguous findings in favor of significant results, for example by not sharing the results of initial trials (Babcock and Lowenstein 1997).

Publication bias, often characterized by journal editors favoring studies that report significant results, exacerbates these problems by providing researchers with an additional incentive to conduct suspect inference and other intellectual contortions (Young et al., 2008). Yet, even absent malfeasance, the fact that significant and surprising results are favored by academic journals provides a natural setting to promote Type 1 error in policymaking, in much the same way the winner's curse operates in common value auctions (drawing from a distribution of results, the significant ones are published, and those are the results used by policymakers).

Naturally, the sort of naïve inference modeled by Maniadis et al. (2014) constitutes

a significant threat to scalability, and one can find examples across a wide variety of disciplines, where false positives lead to vast amounts of wasted public resources. One example is mammograms, where in about 10 to 15 percent of the cases a false positive results. Within academia, stereotype threat appears to have risen to an unwarranted prominent status (see Fryer et al., 2008). Note that this problem is related to, but not the same as the classic multiple hypothesis testing problem.

Fortunately, unlike some of the other threats to scalability, there are remedies to these problems. In the case of the abstract inferential problem considered by Maniadis et al. (2014), there is the solution of replication described above. There are also a wide variety of best practices that should be adopted by journal editors to combat publication bias, such as guaranteeing journal space for replication studies and for studies that yield statistically insignificant results, as well as requiring studies to be declared and registered in advance of their execution as a way of combating the selective presentation of results (Young et al., 2008).

II. Representativeness of the Population: The Attributes and Behavior of Participants and Scalability

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 looking to scale findings, whether the original study is based on naturally-occurring data, field experimental data, or laboratory experimental data. In fact, there exists a lively debate over the relative merits of the aforementioned data types in forming the basis of more general inference (see, e.g., Levitt and List 2007, Al-Ubaydli and List 2015, Deaton and Cartwright 2016).

A less considered issue is the possibility that experimental studies of all forms suffer from inherent biases toward finding estimated causal effects that become weaker under scaling.

One common source of scaling bias is adverse heterogeneity, whereby the participants' attributes make them systematically predisposed to exhibiting a stronger relationship than in the population at large. For example, if parents have priors whether Head Start will help their child, and those priors are correct, then the treatment effect from children who take part in Head

Start will be an upwardly biased estimate of the program at larger scale.

This sort of adverse heterogeneity bias has multiple potential sources. In the case of studies that involve informed consent, if the proposed intervention is a desirable one, such as a financial subsidy, or enrollment in an after-school program, then those who stand to benefit the most will have the biggest incentive to participate, while those who are unaffected, or who might suffer, will systematically opt out. A perusal of the sampled populations in medical trials provides an indication that this sort of effect extends well beyond social programs.

Beyond this, due to the prevalence of publication bias, researchers themselves have an incentive to seek participants who will yield the largest treatment effects. Acting on such an incentive might not even be conscious, and scholars may forgetfully or otherwise omit to mention any implicit grooming when picking participants. In other cases, scholars proclaim that they are using a protocol or sampled population to give the theory or a program "its best chance to succeed" (Smith 1962).

Returning to the literature on field experiments, lab experiments, and generalizability, experimental studies are often characterized by features of the environment that promote unnaturally high levels of compliance, compared to the general population. This could be due to the fact that studies attract compliant participants through selection procedures, that the researcher seeks compliant people by design, or that the physical environment in which the study is conducted induces higher levels of compliance. Laboratory experiments in economics measuring short-run substitution effects include each of these three features, as they typically recruit college students making choices in a college classroom or lab.

In laboratory experiments, compliance may simply be the result of the paucity of strategic options available to participants, such as in a public goods game, where participants are picking one number in each round of the experiment; in the real public goods environment considered by Elinor Ostrom, participants can "walk away" or simply do their own thing.

In natural field experiments, funding-constrained researchers will naturally favor the unique environments where people will most likely comply with the intervention, even if such levels of compliance are unnaturally high. In such instances, compliance is natural in the sense that subjects are acting in their normal course of business and do not know that they are part of an experiment. Consider

Karlan and List's (2007) natural field experiment where they sent charitable solicitation letters to thousands of individuals. They could not measure who actually opened the letters, but they could measure an intent to treat effect comparing the treatment donors to the control group donors. In this way, they measured what the practitioner was ultimately interested in measuring.

One manifestation of non-compliance is non-random attrition, which can reinforce scaling problems. This problem is particularly acute when long-run, or longitudinal estimates, are measured.

Non-compliance is an acute problem in the medical sciences, where clinical supervision is usually significantly higher during the study than can be expected under a population-level rollout. This suggests that patients will comply with prescribed treatments as prescribed in the study, but will exhibit much lower levels of adherence to instructions when scaled.

This result even extends to professional subjects, such as professional health specialists in a best practice experiment that uses hand washing as a treatment. Although numerous studies have illustrated decreased hospital-borne infections with proper hand hygiene practices, noncompliance with these practices are higher in the real-world setting, and impede hospitals from reaching ideal infection control (Grol et al, 2003).

III. Representativeness of the Situation: The Attributes and Behavior of Administrators and Scalability

Analogous difficulties arise on the administrator and "situational" side of the equation. Most of the experimental studies published in the economics literature are administered by the principal investigators, or their lieutenants, such as graduate students. They have a strong incentive to comply with whatever protocol they are investigating, as they seek to maximize the scientific value of their projected discoveries, as well as ensuring the highest possible level of replicability.

When such insights are scaled up, however, it is no longer practically possible for the principal investigators to maintain the role of chief administrator, often because the matter falls under the jurisdiction of much bigger governmental or non-governmental institutions. Moreover, the researchers may even have little interest in following up on the matter, assuming the discovery is complete.

For example, in a review of health interventions targeting HIV and sexual health issues, the majority of studies contained no empirical examination of acceptability, feasible delivery, local needs, or coverage for

implementation (Bonell et al, 2006). And, even when overarching control is retained, the primary researchers will surely have to rely on the administrative assistance of many new people across many differing locales.

Each of these potential threats point to a substantial diminution of control, and in turn noisier and less faithful adherence to the original protocol, and ultimately, therefore, smaller observed treatment effects. For example, an evidence-based 4Real Health teen pregnancy prevention program paired with small community-based organizations for implementation, but encountered barriers, such as inadequate facilities lacking consistent classroom space, inability to hire health educators, and insufficient administrative staff (Demby et al, 2014). These factors, combined with competition with other after school programs, resulted in inadequate recruitment to properly conduct the program and inability to complete all eight sessions.

To some extent, this aspect of the scaling problem reflects the increasing cost of moving up the supply curve. At the small scale associated with the original study, the researchers are able to secure high quality inputs for a relatively low cost—such as bright, keen graduate students willing to administer the experiment in exchange for a good recommendation letter, and using the office printer to print materials without drawing down the research budget.

As the scale increases, professional administrators must be hired, and tenders have to be put out for the material inputs. The economics of the situation naturally lends itself to inferior inputs, or if similar input quality is obtained, a richer price tag accompanies such services. This will especially undermine treatment effects measured in benefit-cost terms, where the cost of provision enters negatively, such as any social program that is compared to other programs that are attempting to attract the attention of policymakers.

Problems stemming from inadvertently chaotic implementation of the original protocol are compounded by those relating to conflicts of interest, especially when rolling out revolutionary ideas, as these often challenge the power and established practices of incumbent organizations.

Implementation frameworks, joining researchers and community leaders, can aid in scaling by rolling out programs with the community needs, resources, and targeted outcomes in mind. Supplee and Metz's (2004) review emphasized the need for greater collaboration among all stakeholders from the beginning of program design, not just implementation, in order to best address

community needs and allow for continuous feedback to drive quality improvement.

Implementation programs, such as PROSPER and Communities that Care, have proven effective in scaling evidence-based preventive programs for youth substance abuse (Hawkins et al., 2012; Spoth and Greenberg, 2007). Programs with greater community coalition functions, communication to key stakeholders, and sustainability planning were more likely to be sustained for 2 or more years beyond their initial funding (Cooper et al., 2015).

Interestingly, the literature has shown that if the original research study sheds light on the 'whys' behind the causal effect observed, fidelity to the original program is more likely (see, e.g., McCoy and Diana 2015). An emphasis on the "whys" also allows for proper identification of all factors within the study contributing to an effect, and these factors may then be used to identify larger populations for scaling. Therefore, casual thinking can be a guide for not only original investigation, but also implementation to further understand what factors and conditions to select so programs can successfully be scaled (Kainz, 2017).

IV. Discussion

Speaking to policymakers has been a major goal of economists for centuries. The experimental method provides a particularly attractive means to continue the discussion, as experiments—lab and field—importantly complement traditional empirical approaches. Whereas experimental economists today focus on how best to obtain parameter estimates—to test theory, speak to firms and policymakers, and in general to improve social welfare—the next important step we must take is understanding how best to use and implement research. We denote this next step as the science of using science.

This next step demands that experimentalists understand the interplay between the research environment and implementation needs necessary at scale. In this way, the scholar must backward induct when setting up the original research plan to ensure swift transference of programs to scale.

Our overview of the primary threats to fluid scaling of programs and their concomitant results should assist scholars in several ways.

First, 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 in the conclusions of their studies, diminishing

the likelihood of spectacular research findings falling flat when policymakers seek deployment.

Second, for a certain class of sources, researchers can take preemptive steps to avoid inadvertently suffering from them. For example, trying to select a sample that will be as compliant with instructions as the population that they are supposedly representing.

Third, some of them can be solved, such as more precise statistical inference, and more prudent journal editing. Our hope is that the rapid advance, and understanding, of the science of using science will permit a step in the right direction to the profession's impact on society.

REFERENCES

Al-Ubaydli, O. and List, J.A. 2015. Do Natural Field Experiments Afford Researchers More or Less Control than Laboratory Experiments? American Economic Review P&P, 105(5):462-66.

Al-Ubaydli O, List, J.A., and Suskind, D. 2017a. The Science of Using Science. In preparation for the International Economic Review.

Al-Ubaydli O., List, J.A., and Suskind, D.L. 2017b XXX In preparation for the Journal of Economic Perspectives.

Babcock, L. and Loewenstein, G., 1997. Explaining bargaining impasse: The role of self-serving biases. The Journal of Economic Perspectives, 11(1), pp.109-126.

Bonell, C., Oakley, A., Hargreaves, J., Strange, V. and Rees, R., 2006. Research methodology: Assessment of generalisability in trials of health interventions: suggested framework and systematic review. BMJ: British Medical Journal, pp.346-349.

Chambers, D.A., Glasgow, R.E. and Stange, K.C., 2013. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. Implementation Science, 8(1), p.1.

Cooper, B.R., Bumbarger, B.K. and Moore, J.E., 2015. Sustaining evidence-based prevention programs: Correlates in a large-scale dissemination initiative. Prevention Science, 16(1), pp.145-157.

Deaton A. and Cartwright, N. 2016. Understanding and Misunderstanding Randomized Control Trials. NBER Working Paper 22595.

Demby, H., Gregory, A., Broussard, M., Dickherber, J., Atkins, S. and Jenner, L.W., 2014. Implementation lessons: The importance of assessing organizational "fit"

and external factors when implementing evidence-based teen pregnancy prevention programs. Journal of Adolescent Health, 54(3), pp. S37-S44.

Feinberg, M.E., Jones, D., Greenberg, M.T., Osgood, D.W. and Bontempo, D., 2010. Effects of the Communities That Care model in Pennsylvania on change in adolescent risk and problem behaviors. Prevention Science, 11(2), pp.163-171.

Fryer, R.G., Levitt, S.D. Levitt, List J.A., 2008. Exploring the Impact of Financial Incentives on Stereotype Threat: Evidence from a Pilot Study, American Economic Review P&P, 98(2), pp. 370-375.

Grol, R. and Grimshaw, J., 2003. From best evidence to best practice: effective implementation of change in patients' care. The lancet, 362(9391), pp.1225-1230.

Karlan, D. and List, J.A. 2007. Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment, American Economic Review, 97(5), pp. 1774-1793.

Levitt, S.D. and List, J.A. 2007. What do Laboratory Experiments Measuring Social Preferences Reveal About the Real World, Journal of Economic Perspectives, 21(2), pp. 153-174.

Maniadis, Z., Tufano, F. and List, J.A., 2014. One swallow doesn't make a summer: New evidence on anchoring effects. The American Economic Review, 104(1), pp.277-290.

Paulsell, D., Avellar, S., Martin, E.S. and Del Grosso, P., 2010. Home visiting evidence of effectiveness review: Executive summary (No. 5254a2ab30e146ce900220dbc4f41900). Mathematica Policy Research.

Raikes, H., Green, B.L., Atwater, J., Kisker, E., Constantine, J. and Chazan-Cohen, R., 2006. Involvement in Early Head Start home visiting services: Demographic predictors and relations to child and parent outcomes. Early Childhood Research Quarterly, 21(1), pp.2-24.

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), pp.574-599.

Saldana, L., 2014. The stages of implementation completion for evidence-based practice: protocol for a mixed methods study. Implementation Science, 9(1), p.1.

Smith, V.L. 1962. An Experimental Study of Competitive Market Behavior. Journal of Political Economy, Vol. 70 , pp. 111–137.

Supplee, L.H. and Metz, A., 2015. Opportunities and challenges in evidence-based social policy. Social Policy Report, 28(4).

Spoth, R., Redmond, C., Shin, C., Greenberg, M., Clair, S. and Feinberg, M., 2007. Substance-use outcomes at 18 months past baseline: The PROSPER community–university partnership trial. American journal of preventive medicine, 32(5), pp.395-402.

Young, N.S., Ioannidis, J.P. and Al-Ubaydli, O., 2008. Why current publication practices may distort science. PLoS Med, 5(10), p.e201.