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.

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 health, education, and employment 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, even though such predictable changes are not accounted for in benefit-cost analysis. We refer to this as the "scalability" problem.

As an example, consider Early Head Start home visiting services, an early childhood intervention that found significant improvements in multiple child and parent outcomes (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).

Understand the scalability of experimental results is critical maintaining the faith of policymakers and the general public in the value of empirical research.

This paper discusses several important threats to scalability. In a companion set of studies, we model the scaling problem (Al-Ubaydli et al., 2017a), and use that theory to explore what lessons economists can learn from medicine (Al-Ubaydli et al., 2017b).

To characterize scalability and highlight certain relevant threats, we divide the problem into three components, which we examine in turn: a statistical procedure applied to the data gathered, representativeness of the population, and representativeness of the situation.

I. Statistical Inference and Scalability

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): 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 those deploying rational expectations.

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 pass as little as two 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.

Continuing with the analogy to rational expectations, naïve scholars' biases can be exploited, 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 editors favoring studies that report significant

results, worsens these problems by providing researchers with an added incentive to conduct suspect inference (Young et al. 2008).

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 (Fryer et al. 2008). Note that this problem is related to, but distinct from 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. Population Representativeness 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 (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 shrink 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. 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 Head Start, 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 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 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 atypically high, yet still natural for the setting.

One manifestation of non-compliance is non-random attrition, which can reinforce scaling problems. This problem is particularly acute when measuring longitudinal effects.

Non-compliance is an acute problem in medicine, where clinical supervision is usually higher during the study than can be expected under a larger rollout. Patients will typically comply with 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 studies of hand-washing and health specialists (Grol et al. 2003).

III. Representativeness of the Situation and Scalability

Analogous difficulties arise on the 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 institutions. Moreover, even when overarching control is retained, the primary researchers will surely have to rely on the administrative assistance of new people across differing locales. Many medical trials do not anticipate such scaling problems (Bonell et al. 2006).

Each of these potential threats point to a substantial diminution of control, less faithful adherence to the original protocol, and, therefore, smaller observed treatment effects. For example, the 4Real Health teen pregnancy prevention program paired with community-based organizations for implementation, but encountered barriers, such as inadequate facilities, inconsistent classroom space, and insufficient hiring of health educators and administrative staff (Demby et al. 2014).

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.

As the scale increases, professional administrators must be hired. This will especially undermine treatment effects measured in benefit-cost terms, where the cost of provision enters negatively.

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.

The above elevates the value of engaging stakeholders, as programs with greater community coalition functions, communication with key stakeholders, and sustainability planning are more likely to be sustained for two 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 mechanism underlying a causal effect observed, fidelity to the original program is more likely (McCoy and Diana 2015), and the researcher’s ability to preemptively identify potential scaling issues is enhanced

IV. Discussion

Speaking to policymakers has been a major goal of economists for centuries. This requires 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 upon larger-scale 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 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 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. International Economic Review (f/coming).

Al-Ubaydli O., List, J.A., and Suskind, D.L. 2017b. Scaling for Economists: Lessons from the Medical Literature. Journal of Economic Perspectives (f/coming)

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.

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.

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. 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.

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

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