INTRODUCTION

The Voltage Effect in Behavioral Economics

John A. List

University of Chicago, ANU and NBER

All happy families are alike; each unhappy family is unhappy in its own way.

-Leo Tolstoy, Anna Karenina.

For decades, a school district in the upper Midwest of the US has been struggling with kindergarten readiness.[1] Administrators have tried a long list of solutions with little success, leaving the District Superintendent, Greta, at her wit's end. A new member of the school board, and a devoted follower of behavioral economics (BE), Mason, recently read an article about an early education behavioral intervention with impressive results: peer-reviewed by academic experts, the study showed large treatment effects on several school readiness indicators. At the end-of-the-year school board meeting, while others discussed the district's woes, Mason brought up the idea of implementing a similar BE program, a potential silver bullet to address the district's pervasive issues with kindergarten readiness. The benefit-cost ratio was astronomical, he assured Greta and the rest of school board. Armed with the science and associated statistical jargon that few could understand, the school board chose to trust Mason and adopt the BE program to scale in their district.

That fall, the school district began to introduce the program, rolling it out in an experimental fashion so that officials could credibly isolate its impacts and prove its benefits to the community. At every fish fry and rotary club meeting they attended, Greta and Mason mentioned the BE program, being sure to mention that economic thinkers as far removed as Adam Smith, Herbert Simon, Gary Becker, Daniel Kahneman, and Dick Thaler produced BE ideas. Just wait until these students apply to college—our first Harvard matriculants are coming soon, Mason boasted at the Lion's Club pancake breakfast. After one year, the results arrived. Mason and Greta pored over the costs, benefits, and outcomes, as measured by standardized cognitive and behavioral tests.

The results: unequivocally mediocre. "The BE program does not even pass a benefit cost test, much less yield the silver bullet that was promised by the original study results. I guess the science got it wrong this time; those BE results didn't scale," Mason concluded.

But *did* the science get it wrong?

I believe that the science likely had it right but that the original results were overinterpreted. The program that Greta and Mason tried to replicate could never carry the water that they had hoped. Indeed, most of us think that scalable ideas have some 'silver bullet' feature, i.e., some quality that bestows a 'can't miss' appeal. That kind of thinking is fundamentally wrong. There is no single quality that distinguishes ideas that have the potential to succeed at scale with those that do not do so. In this manner, moving from an initial research study to one that will have an attractive benefit cost profile at scale is much more complex than most imagine.

And, in most cases, scaling produces a *voltage drop*—the original BE insights lose considerable voltage when scaled. The problem, *ex ante*, is determining whether (and why) that voltage drop will occur. When scaling ideas, one can look to Tolstoy for a bit of wisdom, because in my travels I have learned that *all successfully scaled ideas are alike; all unsuccessfully scaled ideas fail in their own way.* What this lesson inherently means is that scaling, in the end, is a weakest link problem: the endeavor is only as strong as the weakest link in the chain. However, via theory and empirical work, various colleagues and I (see, e.g., Al-Ubaydli & List, 2013; Al-Ubaydli

[1] This Greta and Mason opening passage leans heavily on Al-Ubaydli et al. (2020b). As noted there, except for the names and a few other changes, this is a true story.

et al., 2017a,b; 2020a,b; 2021; Supplee et al., 2021) find that there are five specific traits that scalable ideas must possess—what I call the 'BIG5'. These are the five 'key signatures' of ideas that scale. A deficiency in any one can render an idea unscalable, even for the most ingenious among us.

Before immersing ourselves in the details of the BIG5, it is useful to step back. To start, it is important to note that BE and field experiments have contributed immensely to the "credibility revolution" of the last three decades in the social sciences (see Harrison & List, 2004). In this way, field experiments have become a useful tool for providing causal estimates that are difficult to obtain using other approaches. Yet, since the early 1990s, field experiments have focused primarily on testing BE theories, uncovering BE mechanisms, and estimating program effects. 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.

Nevertheless, what has been lacking is a scientific understanding of how to make optimal use of the scientific insights generated for policy purposes. I denote this as the 'scale-up' problem, which revolves around several important questions, such as: do the BE insights we find in the petri dish scale to larger markets and settings? When we scale the BE 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 BE 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, the 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.

As mentioned previously, several colleagues and I have put together a series of studies that both theoretically and empirically explore these questions. 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 general 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 expected, because its mechanisms are understood. To accomplish this goal, our original experimental designs must provide insights along five key dimensions, to ensure that we are actually scaling ideas and polices that have a chance to make a deep impact, or at least keep the promise of their initial results.

These needs can be broken down into what I call the BIG5. First is the inference problem: how much evidence should be gathered before scaling? I advocate that a post-study probability of at least 0.95 is achieved before enacting public policies (see Maniadis et al., 2014).[2] In practice, this amounts to three or four well-powered independent replications of the original finding. In the case of Greta and Mason, perhaps the original BE results they read about were simply a false positive. A first truth about false positives is that they can be considered 'lies' or 'false alarms'. These are cases whereby, due to statistical error, there was never any voltage in the first place. At the most basic level, a false positive occurs when you interpret some piece of evidence or data as proof that something is true when in fact it is not so. For example, when I visited a high-tech plant in China that produced headsets, if a headset working properly was marked as defective due to human error, that was a false positive. Unfortunately, false positives are ubiquitous across contexts; in a forthcoming book titled *The Voltage Effect* (List, 2022), I summarize findings that suggest a wealth of policies and ideas that fail to scale are simply the result of false positives.

The second element of the BIG5 is representa-

[2] Formally, post-study probability is defined as the share of true associations which are declared true divided by the share of all associations which are declared true.

tiveness of the population. Often, this is the result of failing to know your audience—or assuming that the small subset of people for whom the idea worked originally are representative of the general population that needs to be served, so that when you expand your idea it falls short for a broader set of people. Following the vignette above, in the original study, the researcher might have gathered a sample of students that was much different than the students Greta and Mason had in their district. Greta's school district might have had students with much different characteristics, including observables such as demographics and educational background that did not match the original study. In addition, the original researcher might have reached a population of students that minimized participation costs, or perhaps a population that had characteristics that might yield a larger treatment effect (a 'let's give the idea its best shot of working' recruiting strategy). A medical example of this type of selection effect can be found in meta-studies of recruitment, which confirm that those who stand to benefit most from a medical treatment are more likely to participate in trials (see Al-Ubaydli et al., 2020a). Such selection effects might yield a good journal publication and future grant funding, but it portends a voltage drop at scale as the program is rolled out to everyone.

In a nutshell, researcher choice/bias, selection bias/sorting of the study's population into the program, non-random attrition, and (dis)economies of scale in participation costs all affect the representativeness of the population studied, which in turn might affect the promise of scaling (see Bell & Stuart, 2016).

Third is the representativeness of the situation. A subtle fact is that the research and policy communities oftentimes generalize results to both a population of situations and a population of people, even though we often only speak of the latter. This is particularly troubling considering that the data, thus far, suggest that representativeness of the situation is much more important than representativeness of the population when it comes to generalizing or scaling (see, e.g., List, 2007). For instance, when Greta's school district scaled up the BE program, they did it within their infrastructure, which might have been entirely different from that of the original study, in that certain logistical constraints were present that affected the roll out. If the original results are dependent on the specific context, or they are not done in a policy relevant environment, we can expect the benefit-cost profile to change at scale.

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 the quality of home visits was found on a larger scale, with home visits for 'at risk' families involving more distractions and less time on child-focused activities, thereby diminishing program effectiveness and increasing attrition (Al-Ubaydli et al., 2020a). 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.

The implementation literature sometimes calls this 'context-dependence'. Likewise, in conjunction with curriculum specialists, the original researcher created a curriculum for a pre-kindergarten program, trained the teachers, and provided hands-on support throughout the process. When the school district scaled up the program, they might not have used the exact same curriculum and care as the original implementation, due to local constraints. This is often described as 'program drift' in the literature. This third reason behind voltage effects is generally caused by not understanding that the initial success depended on unscalable ingredients—unique circumstances that cannot be replicated at scale.

A fourth key aspect pertains to spillovers (network effects) and general equilibrium effects of scaling. Concerning the midwestern school district, spillovers could be negative from the treated group to the control group. While the intervention improves the school performance of students in a given class, the control group may, upon seeing an initial improvement in the performance of the treated group, feel demoralized, inducing a deterioration in their performance, accentuating the measured treatment effect (psychologists denote this effect as "resent-

ful demoralization"). Of course, the effect could run in the opposite direction.

Related to spillovers are what economists call *general equilibrium effects*, a term describing shifts in an overall market or system that likely do not manifest on a small scale. To illustrate this notion, let's say that I conducted an experiment wherein I randomly chose 100 college sophomores, forced half of them to change their major to Economics, then examined how much they were earning in their first job compared to the 50 students who did not change majors. I would likely find the Economics majors doing quite well. Now, instead, let's say that I had 50% of all college sophomores around the world change their major to Economics and the other 50% constituted the control group. What would happen a few years later when they all entered the workforce? Assuming no sudden spike in employer demand for Economics majors, a large influx of new economists on the market (increased supply) would cause their wages to plummet: a huge voltage drop.

Here is the rub: our BE experiments typically give us answers along the lines of a small-scale experiment; they don't speak to large movements, such as everyone, or even 50% of college sophomores changing majors. Yet, in a very real sense, this is exactly what we want to know before we scale, especially in the policy world: what are the total effects of my idea in a world where *everyone* changes and *anything and everything* else can change? Ideas do not exist in petri dishes. And an innovation can have negative consequences that are at odds with its purpose but *only become visible at scale.*

Representativeness of the population and the situation as potential threats to scalability underline how fundamental it is to understand 'sites' (i.e., the environment where the original research was implemented) to address the scale-up problem. The literature treats 'sites' loosely whereby some disciplines focus on the population of sites while others emphasize the situational characteristics. I define 'sites' as having multi-dimensional characteristics, which our theory guides into population and situational categories. It is, thus, critical for researchers to describe comprehensively the environment in which the research is carried out, going beyond a cursory description. In this spirit, I advocate that original researchers should stratify (block) on situations when doing experiments, just like we commonly stratify on individual characteristics in modern experimentation (for example, we typically are sure to include both women and men in treatment and control groups, and we do so by stratification; we should do the same for potential non-negotiables in our programs, such as the actual human's delivering, correct dosage, program, delivery, incentives, substitutes, etc.).

Finally, we consider marginal cost considerations. This fifth element of the BIG5 represents the 'supply-side economics' of scaling—does your idea have economies or diseconomies of scale? Greta needed high-quality teachers to run the BE program she was attempting to scale. While the original study only needed 10 teachers, Greta needed 100 for her school district. There was just one problem: the best teachers are also very expensive to retain and hire in the first place. In this case, teachers are very difficult to scale while retaining a reasonable budget. As you 'buy'" more of them at scale, the price invariably goes up, unlike the wholesale price of lettuce going down for Costco when it buys thousands of heads for its locations every week. Indeed, the opposite happens: teachers become more expensive. This is because to attract more high-quality people into the teaching profession, you must raise the teacher salary in order to compete with employers that might pay them more, such as a Wall Street bank or a Silicon Valley tech company.

This key element calls on the analyst to not only measure benefits and how they might scale, but also carefully consider the cost side. This is typically not discussed in the literature, but an idea that has economies of scale is much more likely to scale effectively than one with severe diseconomies of scale. The cost side of the equation just cannot be ignored, and benefit-cost profiles should be computed not only in the petri dish but also at scale.

So where does this leave us? After you clear these BIG5 hurdles, you will know that you have an idea that scales. More generally, while our running example pertained to a public policy, I do not view these insights as limited to helping policymakers. By highlighting the key potential economic sources threatening the scalability of programs and bring-

ing them to the attention of researchers, I hope that those preparing to conduct new studies might consider modifying their own designs such that their reported treatment effect estimates more accurately inform what is likely to occur, should the program be scaled. In this way, as mentioned above, the new demand on scholars is that we backward induct when setting up our original research plan, to ensure accurate and swift transference of programs to scale with minimal uncertainty.

Yet, after the BIG5 are cleared, we are not done. When the program is actually scaled, the correct empirical approach should be taken to measure efficacy, and continuous measurement should be a priority. The first best approach to estimating the effects of the program at scale is to do a large-scale RCT. One can then compare these estimates with the results from the original studies, to explore efficacy at scale. If this approach is untenable, then it is critical to adopt an empirical approach that allows stakeholders to measure its efficacy without unrealistic assumptions. While an exhaustive summary of such approaches is beyond the scope of our work, I point the interested reader to List (2007), who discusses various empirical approaches to policy evaluation as an empirical spectrum, which includes examples of econometric models that make necessary assumptions to identify treatment effects from naturally occurring data. Some of these approaches, such as interrupted time series designs or regression discontinuity analysis, can get pretty close to addressing the internal validity that RCTs solve.

In closing, scaling of ideas is not a silver bullet problem. This is because *all successfully scaled ideas are alike; all unsuccessfully scaled ideas fail in their own way.* I have documented five key reasons why most (if not all) policies and ideas fail to scale. Find an idea that failed to scale, and it will revolve around one or several deficiencies associated with the BIG5. Find ideas that do scale, and they will each be devoid of the BIG5. My work showcases that moving from evidence-based policy to policy-based evidence forces the researcher to backward induct from what a successful idea or policy looks like at scale and test those features in the petri dish.

Nearly every problem has been solved by someone, somewhere. The frustration is that we can't seem to replicate [those solutions] anywhere else.

–President Bill Clinton.

The Author

John A. List is Kenneth C. Griffin Distinguished Service Professor in Economics at the University of Chicago. His research focuses on combining field experiments with economic theory to deepen our understanding of the economic science. In the early 1990s, List pioneered field experiments as a methodology for testing behavioral theories and learning about behavioral principles that are shared across different domains. He was elected a Member of the American Academy of Arts and Sciences in 2011, and a Fellow of the Econometric Society in 2015. List received the 2010 Kenneth Galbraith Award, the 2008 Arrow Prize for Senior Economists for his research in behavioral economics in the field, and was the 2012 Yrjo Jahnsson Lecture Prize recipient. He is a current editor of the Journal of Political Economy.

References

Al-Ubaydli, O., & List, J. A. (2013). On the generalizability of experimental results in economics: In G. Frechette & A. Schotter (Eds.), *Methods of modern experimental economics* (pp. 420-462). Oxford University Press.

Al-Ubaydli, O., List, J. A., LoRe, D., & Suskind, D. (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., & 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.

Al-Ubaydli, O, List, J. A., & Suskind, D. (2020a). 2017 Klein lecture: The science of using science: Towards an understanding of the threats to scaled experiments. *International Economic Review, 61*(4), 1387-1409.

Al-Ubaydli, O., Lee, M., List, J., Mackevicius, C., & Suskind, D. (2020b). How can experiments play a greater role in public policy? Twelve proposals

from an economic model of scaling. *Behavioural Public Policy, 5*(1), 1-48.

Al-Ubaydli, O., Lee, M., List, J., & Suskind, D. (2021). The science of using science: A new framework for understanding the threats to scaling evidence-based policies. In J. List, D. Suskind, & L. Supplee (Eds.), *The scale-up effect* (pp. 104-125). Routledge.

Bell, S. H., & Stuart, E. A. (2016). On the "where" of social experiments: The nature and extent of the generalizability problem. *New Directions for Evaluation, 2016*(152), 47-59.

Harrison, G., & List, J. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009-1055.

List, J. (2007). Field experiments: A bridge between lab and naturally occurring data. *The B.E. Journal of Economic Analysis & Policy, 5*(2), 1-47.

List, J. (2022). *The voltage effect: How to make good ideas great and great ideas scale.* Penguin Randomhouse.

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

Paulsell, D., Porter, T., Kirby, G., Boller, K., Martin, E. S., Burwick, A., Ross, C., & Begnoche, C. (2010). *Supporting quality in home-based child care initiative: Design and evaluation options.* Technical Report 3887af819cdc4b2e9f0e830c0fd3f97a, Mathematica Policy Research.

Supplee, L. H., Ammerman, R. T., Duggan, A., List, J., & Suskind, D. (2021) *The role of open science practices in scaling evidence-based prevention programs.* Working paper.