OMAR AL-UBAYDLI AND JOHN A. LIST

Will It Scale?

During the bloodiest battles of the First World War, a young French sergeant named André Maginot was injured in the fighting, earning him a medal for valor. He later rose to the position of French minister of war and became famous for building a series of defensive fortifications along the Franco-German border. The design of the Maginot Line, as it came to be known, was not the result of abstract musings. Instead, it was deeply informed by Maginot's own experience—particularly his assessments about optimal military tactics in the fierce, close-quarters combat of the Great War.

But then the Second World War came, and the Maginot Line failed when Nazi Germany's military machine circumvented the fortifications by invading France through Belgium. The skirmishes Maginot witnessed along the Western Front from 1914 to 1918 constituted a small-scale proof-of-concept, convincing him that extending fortifications along the entire Franco-German border would deter invasion. Instead, the strategy ended up being a deadly manifestation of what social scientists refer to as the *scaling problem*: the efficacy estimated from small or pilot programs shrinks or evaporates when programs are expanded. The static defenses of the Maginot Line were effective for narrowly delimited battlefronts, where the enemy had the option of either going forward or retreating, but they failed at a larger scale, where the invading army could just choose another path.

To this day, orthodox scientific methods remain aligned with the ideas about scaling that animated Maginot's failed fortifications. This is exemplified by the biomedical trials used to evaluate new pharmaceuticals. Ideas are first tested in a restricted environment, such as a petri dish; or in Maginot's case, various battles in the mud-soaked fields of the Western Front. If it works in the petri dish, that is taken as a signal to scale it up systematically. We refer to this approach—testing no intervention versus testing an intervention under a limited (usually the best possible) scenario—as *A/B testing*.

The A/B testing approach invites promising early results that are unlikely to be realized in a larger setting. We argue that within the social sciences, a fundamentally different approach is needed; we call it *option C thinking*. Put simply, a twenty-first-century team of civil servants and social scientists should lead with experiments that anticipate likely causes of failure at scale, even if doing so requires more time, effort, and resources initially.

Scale-up letdowns

Statistical flukes, well-intentioned errors, cognitive biases, other oversights, and even willful deceit readily boost estimates of effectiveness at the proof-of-concept stage. This produces seductive, unreliable evidence that can lead decisionmakers astray, particularly if they are seeking out research for new solutions to old problems.

Scaling ineffective programs can waste money, time, and cause people harm by blocking more promising alternatives. Consider the Drug Abuse Resistance Education (D.A.R.E) program, which built on social inoculation theory and aimed to inoculate kids against the temptation of drugs. Early on, a study in Honolulu that found D.A.R.E to be effective estimated there was a 2% chance the researchers' data could yield a false positive, leading the government to scale up the program. Subsequent studies found that the program did not work. The researchers either made a statistical error or the result fell within that 2%. Another example is a small-scale experiment in Seattle, in which one of us (List) and colleagues found that Uber users who got a $5-off coupon took more rides than those who did not receive the coupons, and the increased earnings from those rides offset the effect of the discount. But when the initiative was scaled up to a larger group of Seattle riders, the shortage of drivers resulted in higher fares and wait times, which led to an overall decrease in demand.

In both cases, the actual conditions under which an intervention was implemented differed from the idealized form in which it was tested. Considering that difference is the crux of option C thinking, and this requires a thought process we've called "backward inducting from reality." A relevant example comes from a preschool List and colleagues started in Chicago in order to identify programs that could decrease the achievement gap—the Chicago Heights Early Childhood Center. A typical A/B test would recruit stellar teachers and compare learning in our program versus a traditional one. But we realized that, at a scale of thousands of schools, not every teacher could be exceptional. So we designed our study to examine whether our curriculum would work with teachers of varying abilities. We recruited teachers who would typically come and work in a school district like Chicago Heights. This choice provided the A/B efficacy test because we had several stellar teachers, but by populating option C as well, we ensured that the situation was representative—at least on the dimension of teacher quality.

For another example of option C thinking, consider how American company Opower, alongside Honeywell, implemented a new smart thermostat with great energy-saving promise. It would modify the temperature when occupants weren't home and reduce costs by turning off during peak hours. However, when taken to scale, the benefits failed to come to fruition. A team (including List) came in to determine what went wrong. It turned out most customers undid all the power saving settings. With option C thinking, the engineers developing the technology would have tested it using not only the optimal settings, but also trying the ones actually chosen by the end user. If this had been done, the engineers could have considered ways to encourage people to use the energy conserving settings before taking their product to scale.

Causes of "voltage drop"

Scaling failures are often dismissed with a handwave or a shrug, as if what happened was unforeseeable. However, we have found such failures fall into five general causes. In a 2022 book, *The Voltage Effect*, List linked the idea of scaling failures to the metaphor of voltage drops. Just as voltage along an electrical cable decreases with distance from the source, the effectiveness of a tested intervention will fall with its distance from its "ideal" real-world implementation.

The first reason for failure to scale is the *false positive*, or a statistical fluke in the original research. Especially for small-scale experiments, the probabilistic checks used to conclude that a program works are not foolproof; the possibility that a positive result was a lucky but ultimately misleading inference can never be ruled out.

The second and third causes relate to *epistemological representativeness*. Sometimes, the population used in a positive trial looks very different from the general population that the program will be rolled out to. For example, when testing energy efficiency programs, early adopters are often those most excited about conserving energy and therefore most responsive to the intervention. However, the population at large is far more likely to contain disinterested and obstinate energy consumers. Health and education programs have failed for similar reasons, including not accounting for varying needs by age, wealth, and other demographics. The other epistemological mismatch occurs when the situation or context during the trial is unrepresentative, as in the case of the Maginot Line failing to account for the possibility of circuventing the fortifications. Another example includes COVID-19 vaccination messaging that failed to incentivize people who were reluctant to receive the new vaccines.

A twenty-first-century team of civil servants and social scientists should lead with experiments that anticipate likely causes of failure at scale, even if doing so requires more time, effort, and resources initially.

The fourth cause of failure is the effect of *negative spillovers*, which describes what happens when a program has a positive effect on those enrolled, but a negative one on the unenrolled that is imperceptible in small samples but shows up when the program is expanded. For example, in one test, increasing Uber drivers' salaries by raising fares led to more hours worked. However, at scale, the benefits were muted because some of the extra hours included greater efforts to steal passengers from fellow Uber drivers, rather than more time transporting paying passengers.

A fifth cause of the scaling problem can be attributed to *supply-side factors*, reflecting the fact that the unit costs of small-scale experiments can be misleadingly low. For example, Saudi Arabia has considered plans to introduce Chinese language lessons to all children at school. Hypothetically, a small trial would be relatively inexpensive because it would employ a few local teachers already qualified in the language. However, at scale, costs may rise sharply because the government would need to greatly expand the supply of instructors, which would likely mean paying professionals to relocate to Saudi Arabia.

Underplaying scaling problems

Unfortunately, current research infrastructure and incentives contribute to the scaling problem. Overly eager—or intentionally corrupt—scholars and program officers may artificially inflate the results of small-scale trials and so increase the likelihood a program expands—and then fails.

One way scientists may inflate results is by rerunning experiments (or analyses) until they get lucky, and presenting that lucky trial as their sole attempt. Alternatively, they can handpick the participants in their trial to maximize the effect of the program, exploiting their knowledge of who is most likely to benefit. They might overwork their PhD students and artificially decrease program delivery costs. The possibilities are endless, but the consequences can be devastating for the budgets and credibility of government entities that roll out ineffective programs. Civil servants need to be wary of scientific doping.

Fortunately, countermeasures exist to combat these practices, whether they are caused by an overly eager, incompetent, or deceitful scientists. For example, the scholarly community has started to demand replication of certain scientific findings by independent teams of scientists before results are afforded credibility. Increasingly, researchers are required to post all results and data, which can reveal cherry-picking or otherwise biased analyses. When drawing on research to design public programs, decisionmakers should take their cue from the scientific community and find approaches to ensure findings are reliable.

Countering scale failures by testing bigger

In our opinion, option C thinking enhances reliability in the social sciences. It effectively means asking—before research begins—why an idea would fail at scale. In our experience, such reasons are not difficult to anticipate if one has the discipline to face the question.

Leveraging option C thinking may begin with a new treatment arm, such as testing a smart thermostat with the end user. Ultimately, option C thinking is a mindset that augments the traditional A/B approach by proactively producing the type of policy-based evidence that the science of scaling demands. In a nutshell, it means starting with a recognition of the big picture and anticipating what information is needed to scale up. The goal is to gather the evidence that provides greater scaling confidence from the initial design.

As economists like to say, there is no such thing as a free lunch: adopting an option C approach brings significant risks and downsides that decisionmakers must bear in mind. A key virtue of A/B testing is that an individual trial is cheap, even though the cost of mistakenly scaling an intervention could be catastrophic. Implementing option C thinking builds that consideration into the equation before the (overly encouraging) results come in. It also accommodates instances where executing a trial will have large fixed costs, or where benefits might only be apparent at a large scale.

The higher cost of option C testing brings another risk that is easy to overlook. A decisionmaker entertaining proposals for research or interventions will naturally be biased toward submissions from famous professors working at elite scientific institutions. The greater the proposed cost, the more acute this bias will be. Smaller entities, with their more limited means, struggle to submit credible bids, and policymakers look to hide behind the safety of a big name lest the experiment turns out to be a total failure. Thus, there is the possibility that option C thinking reinforces existing inequities in science, and pushes the process of knowledge production closer to a winner-takes-all format.

There are no straightforward solutions to these valid concerns. However, just sticking to A/B testing is far from ideal. Option C thinking should be part of the toolkit belonging to any scholar hoping to influence policy, and any decisionmaker involved in program implementation. When government officials are considering new programs, they must resist the natural urge to fixate on small-scale testing as a first stop for ideas, and instead be open to the benefits of starting with a large-scale experiment that captures more of the big picture. We'd also like to see the social sciences adapt, so that researchers reflexively design studies to consider the most likely causes of failure in advance, priming public programs for success rather than failure.

While option C testing requires going big, there are some small steps that government officials can take to support it. The simplest is to explicitly require evaluators to consider the five causes of voltage drop, which will bolster research that anticipates stumbling blocks. This might include working within the scientific infrastructure on how grants are evaluated, contracts awarded, and publications assessed. Another step is to select a handful of priority areas that can particularly benefit from option C thinking and provide sufficient resources, ideally while also introducing mechanisms to ensure a diverse group of researchers is recruited.

We will never know what might have happened on May 10, 1940, when Hitler commanded his forces to outflank the Maginot Line, had Maginot previously been exposed to the modern epistemological analysis of the scaling problem. However, by adopting option C thinking, decisionmakers give themselves the best chance of conceiving of and rolling out effective programs at a time when resources are scarce and the public's trust in government is worryingly low.

Omar Al-Ubaydli *is the director of research at the Bahrain Center for Strategic, International, and Energy Studies.* ***John A. List*** *is a professor of economics at the University of Chicago known for his work in establishing field experiments as a tool in empirical economic analysis.*