Behavioural Public Policy, Page 1 of 9
© The Author(s) 2020. Published by Cambridge University Press doi:10.1017/bpp.2020.20

Success at scale: six suggestions from implementation and policy sciences

TECK-HUA HO*
National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077
CHING LEONG
National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077
CATHERINE YEUNG[1]
Chinese University of Hong Kong, Shatin, Hong Kong

Abstract: Very often, significantly smaller benefits are observed in final policy outcomes than are indicated by initial research discoveries. Al-Ubaydli *et al.* have identified a poor understanding of the 'science of scaling' as the underlying cause of this discrepancy. They propose a framework to increase our understanding of the science of scaling. We build on this framework by making six specific suggestions capturing three key ideas. First, researchers need to move away from their preoccupation with general theoretical models and focus on subject-specific theories of intervention, leading to individualized treatments. Second, there should be greater collaboration between researchers and policymakers, as well as more transparency in reporting findings, to ensure that the research environment is more representative of the policy environment. Third, researchers should recognize that policymakers do not always maximize social welfare; policymakers may have their own short-term incentives. Therefore, researchers must consider policymakers' short-term incentives in designing interventions in order to increase the chances of a research intervention becoming a policy.

Submitted 31 October 2019; accepted 6 April 2020

Introduction

What could be more sensible than to try at a small scale what we would like to do at a large one? As US Supreme Court associate justice Louis Brandeis said (*New State Ice Company v. Liebmann,* 285 U.S. 262, 1932), "[A] single courageous State may, if its citizens choose, serve as a laboratory; and try novel

* Correspondence to: National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077 E-mail: teck@nus.edu.sg.

[1] Authors are listed in alphabetical order. All authors contributed equally.

social and economic experiments without risk to the rest of the country." The assumption here is clear: one state can serve as a small-scale replica of the country as a whole. However, this assumption is seldom valid.

Analogously, we observe this scalability problem when we translate research insights gained from field experiments into actual public policy. Often, policymakers find that there is a significantly smaller benefit than indicated by the original research outcomes – what researchers call a 'voltage drop' (Kilbourne *et al.*, 2007). Al-Ubaydli *et al.* argue that the cause of this voltage drop is a poor understanding of the 'science of scaling'.

We commend the authors for bringing this well-known challenge in economics into sharp focus. They have developed a framework of 12 proposals that address threats associated with the scalability problem. The framework provides important insights into two key areas: first, identifying the motivations and the incentives for different actors (the public, researchers and policymakers); and second, looking at the issue of representativeness, both of the target population and of the eventual policy environment.

We build on the authors' framework by positing that human motivation must be the centerpiece of the policy translation process. We must understand and address the human motivations of these three groups – the public, researchers and policymakers – in order to overcome the scalability problem.

The public is highly heterogeneous in motivation and is boundedly rational. As a consequence, any policy involving the assumption that the public is fully rational (as assumed by standard economic theory) can be problematic. One example is that of recycled drinking water (RDW) (Lazarova *et al.*, 2001). RDW is accepted by participants in small-scale pilot projects, but is rejected by large populations because of the "yuck" factor – a majority of the public still believes that purified water is not clean, as happened with the San Diego water repurification project in the late 1990s (Po *et al.*, 2003).

Researchers are not properly incentivized to work with policymakers when researchers design their proof-of-concept studies. As a result, the context of the proof-of-concept is often quite different from the environment in which a policy will operate. The current publication process also penalizes researchers who document the specifics of the environment for their proof-of-concept studies because reviewers see the specifics as limitations on the generalizability of the reported findings. As a result, the contexts of the reported findings are often quite different from the contexts of the actual policies.

Policymakers operate in a muddy world of accountability, legitimacy and public perception. Public policy researchers capture the political nature of this world through the phenomenon of credit and blame (Weaver, 1986; Hood, 2011). Political leaders and government agencies are beholden to the masses they serve. They may receive praise for their actions – credit, which allows political advancement – or they may be blamed for their choices. As a consequence, policymakers and public officials often engage in self-interested behavior such as 'credit claiming' and 'blame avoidance' (Leong & Howlett, 2017). In short, they may not maximize social welfare.

In this paper, we focus on the motivations of the public, researchers and policymakers, and we build on the framework in Al-Ubaydli *et al.* by making six specific suggestions: (1) develop subject-specific theories of intervention; (2) quantify the non-representativeness of the population under study and adjust the reported treatment effect; (3) create a formal empirical methodology to customize interventions for individuals or segments of a policy population; (4) engage policymakers in proof-of-concept studies in order to understand constraints when designing interventions; (5) ensure transparency in describing research protocols, design choices and sampling methodologies; and finally (6) seek interventions that align social welfare with short-term incentives that are attractive to policymakers.

Representativeness of the population

Studies involving human subjects must seek approval from an institutional review board (IRB). A cornerstone of IRB approval is voluntary participation and explicit subject consent. As a result, subjects in small-scale proof-of-concept studies are necessarily unrepresentative of a target policy population. This issue will never go away.

To make the situation worse, researchers may use a convenience sample (e.g., student subjects) to conduct their studies. If the primary purpose of the study is to test theory and the theory is silent about the influence of subject characteristics, the unrepresentativeness of the sample is by itself not a major issue for publication. Yet, when the same intervention is applied to a policy population with different characteristics, the outcome can be very different.

In a large field experiment involving overweight people, Ho *et al.* (2019) show that a simple financial incentive for weight loss only works on men. Therefore, the effect will be significantly lower if the same financial incentive is applied at scale to the overweight population. Similarly, sending SMS reminders to all diabetic patients may not significantly improve medication adherence or slow disease complications because the intervention only works for the small segment of patients whose medication non-adherence was due to forgetfulness or absentmindedness (Vervloet *et al.*, 2012). Put simply, if the public is heterogeneous and if proof-of-concept studies are necessarily unrepresentative, there are always limits to building a general intervention theory for populations at scale. Instead, what we need are individual-specific intervention theories.

We are not simply asking for transparent reporting of subject characteristics. In fact, Consolidated Standards of Reporting Trials (CONSORT) guidelines already require that subject characteristics be reported so that readers can judge how relevant the results of a trial might be to a target population. While reporting subject characteristics is necessary, it does not guarantee that policymakers can judge the relevance of findings if connections between subject characteristics and research outcomes have not been established. Consider an example from pharmaceutical science. Historically, prescription drugs have mostly been tested on men (women with child-bearing potential were often excluded because of concerns about potential adverse effects on pregnancy). Yet policymakers assumed that any effects also applied to women, until researchers pointed out systematic gender differences in drug response (Liu & Dipietro Mager, 2016).

This leads to our first suggestion.

> *Suggestion #1:* Researchers need to develop subject-specific theories of an intervention that postulate not just the expected treatment effect, but also the specific populations that will demonstrate such a treatment effect.

Given that most existing theories are not subject-specific, it is crucial for researchers to estimate the degree of non-representativeness in their studies. Business and social science researchers have empirically demonstrated that findings can be biased if one does not account for non-responders (people who do not respond and therefore are not in the sample). However, it is possible to correct these biases using a follow-up survey. For example, Peytchev *et al.* (2009) revisited a national survey conducted for the Centers for Disease Control and Prevention aimed at estimating the prevalence of sexual violence. The authors drew a random sample of non-responders to the original survey and contacted them for a follow-up survey. To motivate the non-responders to participate, they offered a larger incentive and halved the number of questions in the survey. These changes brought in 'new' respondents who significantly differed on key survey variables from respondents to the original survey. Adding these new responses to the original dataset allowed the authors to provide a better estimate of the prevalence of sexual violence.

> *Suggestion #2:* Proof-of-concept studies should quantify the non-representativeness of the population in a follow-up study and make adjustments to the reported treatment effect.

In field experiments, many policy interventions work, but only for a subset of the population. Loss aversion is perhaps the most reported behavioral tendency in laboratory experiments, but its existence is more often found inside laboratories than in the field. Camerer (1998) provides a comprehensive survey of the applications of prospect theory in the wild and suggests that loss aversion is valuable in explaining naturally occurring phenomena, but only when people are narrowly bracketing the relevant decisions. Similarly, List (2003) shows that only inexperienced stock traders exhibit loss aversion in markets.

If a behavioral intervention works for only a subset of people, its scaled implementation (even with the highest level of fidelity) can at best affect a subset of the population. If a cost is associated with the treatment for each subject, this phenomenon often results in an unprofitable policy implementation. Put differently, a one-size-fits-all approach to policy implementation simply does not work.

In many disciplines, the science of scaling lies in individualization. In artificial intelligence, computer scientists propose reinforcement learning as a way to allow policymakers to customize interventions in real time so that different people receive different treatments in order to increase the average effect size of a catalog of treatments designed for a population. Using artificial intelligence, it is possible for policymakers to make granular interventions that affect specific individuals.

For example, educational psychologists have developed and tested numerous theories to describe human learning and to prescribe educational practices. Each theory tackles a specific learning challenge and only applies to a subset of students. Using machine learning, computer scientists can now design algorithms that customize learning materials based on these theories for each student. Some computer-assisted learning programs have shown promising results in improving academic achievement at scale (Kulik & Fletcher, 2016; Banerjee *et al.*, 2017; J-PAL Evidence Review, 2019).

It is important to point out that each individualized or segment-based intervention should be informed by theory and must go through a rigorous research and testing process before it can be put into practice.

> *Suggestion #3:* A formal empirical methodology should be developed to customize interventions for an individual or segment within a policy population in order to increase the average effect of a group of treatments applied to that population.

Representativeness of the situation

The importance of working with policymakers in proof-of-concept field experiments cannot be over-emphasized. Researchers must understand the constraints of policy realities and design interventions that are practical. This includes being attentive to the associated costs of their prescribed interventions that change existing work practices. Banerjee *et al.* (2017) demonstrate this point by showing how scaling up a simple intervention can fail.

The authors reported on taking the 'Teaching at the Right Level' (TaRL) pedagogical approach from proof-of-concept to scalable implementation. TaRL organizes students, for a fixed duration of time, by level of knowledge rather than age. Despite promising results from the proof-of-concept randomized controlled trial, there was resistance from teachers and parents when it was scaled up. They insisted on following the grade-level curriculum, despite the fact that some slower students were not up to the grade-level standard. This example demonstrates the importance of getting insight and buy-in from 'insiders' early on. Hence, including policymakers in the research team will certainly help to lower resistance to change.

> *Suggestion #4:* Researchers should engage policymakers in proof-of-concept studies in order to understand the major constraints faced by policymakers when designing interventions.

If it is not possible to involve policymakers early in the proof-of-concept experiment, it is important to ensure that the research environment and design choices are transparent and formally documented. We are proposing this as a principle of good practice. Once transparency becomes the standard, researchers can rigorously test the limits of replicability, giving policymakers confidence in a study's findings and scalability.

A good example can be found in the site selection experiments in Allcott (2015, pp. 1131–1134). Researchers must be transparent about the rationale behind site and sample selection, especially when selection is not random (e.g., forced upon them by circumstance, or the selection gives a more promising outcome). Policymakers must have full knowledge of this rationale.

> *Suggestion #5:* Researchers should adhere to the principle of transparency in describing their research protocol, design choices and sampling methodology.

Aligning incentives between researchers and policymakers

Bureaucrats and politicians are two groups of actors with a large effect on policy. It is therefore useful to see how credit claiming and blame avoidance affect their actions. Most bureaucrats prefer inaction (i.e., status quo bias) (Weaver, 1986; Howlett, 2014; Leong & Howlett, 2017) because of the fear of making mistakes and attracting blame. Frequently, not making mistakes (rather than gaining praise for policies done right) is sufficient for career progression. In fact, research shows that policymakers are risk averse due to bureaucracy, incentives and the external environment (Rainey, 2009; Demircioglu, 2018). Politicians may be differently motivated – for example, to maximize their chances of re-election, politicians may offer short-term rewards at the expense of society's long-term interests (Jacobs, 2008; Kang & Reich, 2014; Leong & Howlett, 2017; Marx, 2017).

In short, politicians may seek the rewards that come with populist decision-making instead of trying to maximize social welfare. Conversely, policymakers may not implement a good policy because it is not easy to claim credit from it, such as investing in vital infrastructure projects that have low visibility. Marx (2017) discusses several case studies in Africa that demonstrate this point. For example, in 2006 in Mozambique, the government reduced the number of primary schools they would build from 12,000 to 6000 in order to build secondary schools, which are more notable. The World Bank, which co-funded the project, reported: "This was in contrast to the initial objective of building smaller and more dispersed schools and as a result, fewer communities benefited from the construction program" (World Bank, 2006, p. 26). In short, the political gains from building fewer but larger and more visible secondary schools trumped the policy objective of extending primary education in the country.

> *Suggestion #6:* Researchers must not assume that policymakers maximize social welfare. They must actively seek interventions that align social welfare with short-term incentives that are attractive to policymakers.

Conclusion

Building on the framework provided by Al-Ubaydli *et al.*, this paper makes six suggestions to further our understanding of the science of scaling and to improve the chances of success at scale. We focus on understanding the human motivations of three actors (the public, researchers and policymakers) in the chain of evidence-based public policymaking.

Individuals are heterogeneous, so it is important to develop individual-specific theories of intervention. Since people respond to interventions differently, interventions must be customized based on individual characteristics. While we are waiting for more individual-specific theories and interventions to materialize, researchers must quantify the representativeness of the sample of subjects involved in the proof-of-concept study with respect to the target population. This can be done through a follow-up study that captures subject responses that are different from those in the original study.

There is an urgent need to make research environments more representative of the eventual policy environment by involving policymakers early on, when designing the proof-of-concept study. This is crucial because policy environments often face bureaucratic constraints and rigid work practices. If early involvement is not possible, researchers must be transparent in describing their unique research environment so that policymakers can judge how representative the research environment is of the eventual policy environment.

Finally, we emphasize the potential misalignment between policymakers' short-term incentives with societal, long-term welfare. Policymakers are self-interested and behave so as to claim credit and avoid blame; they do not necessarily maximize social welfare. Therefore, researchers must be sensitive to this misalignment and do their best to design interventions that align policymakers' interests and social welfare.

We believe that all six suggestions are critical to improving the chances of success at scale. This will be a long and difficult process, but one that will allow economists to play a greater and more impactful role in policy creation and implementation.

References

Allcott, H. (2015), 'Site Selection Bias in Program Evaluation', *The Quarterly Journal of Economics*, **130**(3): 1117–1165.

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

Camerer, C. (1998), 'Prospect Theory in the Wild: Evidence from the Field. Reprinted', in Max H. Bazerman (ed), 2003. *Negotiation, Decision Making and Conflict Management*, UK: Edgar Elgar Publishing, Ltd.

Demircioglu, M. (2018), 'The Effects of Empowerment Practices on Perceived Barriers to Innovation: Evidence from Public Organizations', *International Journal of Public Administration*, **41**(15): 1302–1313.

Ho, T. H., C. Yeung, N. Lim, R. M. van Dam, R. Sato, K. W. Tham and H. C. Tan (2019), *Cash Incentives for Weight Loss Work Only for Males*. Working Paper.

Hood, C. (2011), *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*, Princeton University Press.

Howlett, M. (2014), 'Why are policy innovations rare and so often negative? Blame avoidance and problem denial in climate change policy-making', *Global Environmental Change* **29**, 395–403. https://doi.org/10.1016/j.gloenvcha.2013.12.009

J-PAL Evidence Review. (2019), *Will Technology Transform Education for the Better?* Cambridge, MA, Abdul Latif Jameel Poverty Action Lab.

Jacobs, A.M., (2008), 'The Politics of When: Redistribution, Investment and Policy Making for the Long Term', *British Journal of Political Science* **38**, 193–220. https://doi.org/10.1017/S0007123408000112

Kang, M. and M.R. Reich, (2014), 'Between credit claiming and blame avoidance: the changing politics of priority-setting for Korea's National Health Insurance System', *Health Policy* **115**, 9–17. https://doi.org/10.1016/j.healthpol.2013.09.015

Kilbourne, A. M., M. S. Neumann, H. A. Pincus, M. S. Bauer and R. Stall (2007), 'Implementing Evidence-Based Interventions in Health Care: Application of the Replicating Effective Programs Framework', *Implementation Science*, **2**(1): 42.

Kulik, J. A. and J. D. Fletcher (2016), 'Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review', *Review of Educational Research*, **86**(1): 42–78.

Lazarova, V., B. Levine, J. Sack, G. Cirelli, P. Jeffrey, H. Muntau, M. Salgot and F. Brissaud (2001), ‘Role of Water Reuse for Enhancing Integrated Water Management in Europe and Mediterranean Countries’, *Water Science and Technology: A Journal of the International Association on Water Pollution Research*, **43**(10): 25–33.

Leong, C. and M. Howlett (2017), ‘On Credit and Blame: Disentangling the Motivations of Public Policy Decision-Making Behaviour’, *Policy Sciences*, **50**(4): 599–618.

List, J. A. (2003), ‘Does Market Experience Eliminate Market Anomalies? *The Quarterly Journal of Economics*, **118**(1): 41–71.

Liu, K. A. and N. A. Dipietro Mager (2016), ‘Women’s Involvement in Clinical Trials: Historical Perspective and Future Implications’, *Pharmacy Practice*, **14**(1): 708.

Marx, B. (2017), Elections as Incentives: Project Completion and Visibility in African Politics. Working Paper. https://www.tse-fr.eu/elections-incentives-project-completion-and-visibility-african-politics.

Peytchev, A., R. K. Baxter and L. R. Carley-Baxter (2009), ‘Not All Survey Effort Is Equal: Reduction of Nonresponse Bias and Nonresponse Error’, *Public Opinion Quarterly*, **73**(4): 785–806.

Po, M., J. D. Kaercher and B. E. Nancarrow (2003), *Literature Review of Factors Influencing Public Perceptions of Water Reuse*. CSIRO Land and Water.

Rainey, H. (2009), *Understanding and Managing Public Organizations*. San Francisco, CA: John Wiley & Sons.

US Supreme Court. (1932), New State Ice Co. v. Liebmann, 285 U.S. 262. Retrieved from https://supreme.justia.com/cases/federal/us/285/262/.

Vervloet, M., A. J. Linn, J. C. van Weert, D. H. De Bakker, M. L. Bouvy and L. Van Dijk (2012), ‘The effectiveness of interventions using electronic reminders to improve adherence to chronic medication: A systematic review of the literature’, *Journal of the American Medical Informatics Association*, **19**(5): 696–704.

Weaver, R. K. (1986), ‘The Politics of Blame Avoidance’, *Journal of Public Policy*, **6**(4): 371–398.

World Bank. (2006), *Implementation Completion and Results Report: Education Sector Strategy Program, Republic of Mozambique*. Report No: ICR000029.