Behavioural Public Policy (2021), 5: 1, 80–89
© The Author(s), 2020. Published by Cambridge University Press doi:10.1017/bpp.2020.24

Successfully scaled solutions need not be homogenous

DILIP SOMAN*

Canada Research Chair in Behavioural Science and Economics at the Rotman School of Management, University of Toronto, Toronto, Canada

TANJIM HOSSAIN

Professor of Marketing in the Department of Management at the University of Toronto Mississauga, Mississauga, Canada

and

Rotman School of Management, University of Toronto, Toronto, Canada

Abstract: Al-Ubaydli *et al.* point out that many research findings experience a reduction in magnitude of treatment effects when scaled, and they make a number of proposals to improve the scalability of pilot project findings. While we agree that scalability is important for policy relevance, we argue that non-scalability does not always render a research finding useless in practice. Three practices ensuring (1) that the intervention is appropriate for the context; (2) that heterogeneity in treatment effects are understood; and (3) that the temptation to try multiple interventions simultaneously is avoided can allow us to customize successful policy prescriptions to specific real-world settings.

Submitted 23 January 2020; accepted 20 March 2020

The publication of *Nudge* (Thaler & Sunstein, 2008) ushered in an era in which governments and businesses started increasingly depending on the results of evidence to make better product, program and process design decisions (OECD, 2017). Over a decade after the publication of *Nudge*, Soman and Yeung (in press) argue that while the relevance, acceptability and methods of behavioral science in organizations are now not questioned, two challenges for the next decade of applied behavioral scientists remain. The first challenge relates to creating organizations that truly understand the nature of behavioral science and embed it deeply in their structures and methods. The second challenge is one of scaling – how do we ensure that the results of small-scale

* Correspondence to Rotman School of Management, University of Toronto, 105 St. George St., Toronto, ON M5S 3E6, Canada. E-mail: dilip.soman@rotman.utoronto.ca

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

experiments and pilots successfully scale into interventions that reliably improve consumer welfare?

In their lead article, Al-Ubaydli and colleagues acknowledge the increasing role of the experimental method to inform large-scale public policy programs and welfare initiatives, and they provide a list of 12 proposals that will advance the ability of policymakers to scale the results of pilots. We applaud their thoughtful analysis and agree that without a better science of using science, the promise of behavioral science in policymaking and end-user welfare will not be fully realized. We also agree that the 'voltage drop' – the drop in effectiveness of a scaled-up intervention relative to the original pilot (Kilbourne *et al.*, 2007) – presents a significant scaling challenge. Voltage drops occur because: (1) researchers often make imprecise and exaggerated inferences based on the original pilot; and (2) the original pilot might not be representative of either the larger population or the situation in which the intervention will be scaled up.

While our opinions are largely similar, we diverge and extend their thinking in two ways. First, unlike the model they draw heavily on (Al-Ubaydli *et al.*, 2019), we do not necessarily believe that government entities and researchers are completely independent actors, but instead make a call for a different kind of evidence (*in situ* testing) created by collaborations between the two entities. Second, we believe that the lack of representativeness of the situation can not only result in a voltage drop, but might also result in backfiring effects and produce the opposite result of what they were intended to do.

We identify three specific scaling challenges that are worthy of further investigation. First, one of the biggest themes from the research on behavioral science and judgment and decision-making is the notion of context dependence (Tversky & Simonson, 1993). We now know that a large number of elements in the context (e.g., the presence of irrelevant options, medium of the message, time of the day, information framing and ambient factors, to name a few) influence choices (for overviews, see Soman, 2015; Thaler, 2015). A rich literature of preference reversals has shown, for example, that consumers' preferences might reverse with seemingly irrelevant changes in the context (Lichtenstein & Slovic, 2006). For policymakers, the implication should be clear – the fact that a particular intervention worked well in one particular context should not guarantee its success in a different context.

Second, like Al-Ubaydli *et al.*, we agree that the lack of representativeness of pilot study participants poses a challenge. One area of analysis that is relatively uncommon in this literature is heterogeneous treatment effects. While it is important to understand the efficacy of a treatment among a representative set of participants, we posit that it is equally important to understand the heterogeneity in treatment effect. As small-scale research studies may focus on the

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

set of people for whom the treatment effects are likely to be the largest, a goal for large-scale replications can be to figure out the effect sizes for different groups of the population. If done successfully, this may lead to successful use of treatments that may not scale well to the whole population, but can be very effective on a subgroup of that population. By identifying the boundaries of the effectiveness of an intervention and understanding the heterogeneous effects, policymakers and researchers can target interventions to the appropriate population. Heterogeneous treatment effects imply that successful policies do not have to be one-size-fits-all.

Third, there is often a temptation on the part of policymakers to adopt a kitchen sink approach in which multiple interventions that have been successfully tested independently are deployed simultaneously in the naive belief of improving efficacy. Unfortunately, these multiple interventions can interact in complex and unpredictable ways, which might result in backfiring.

We next provide three examples to illustrate some of these scaling challenges. We then speculate on how these scaling challenges might play a role if we were to scale up the results from a field experiment that one of us was involved in (Hossain & List, 2012). We will then conclude with a discussion and offer additional propositions.

Example 1: credit card reminders in South Korea

Research shows that consumers tend to spend more when using a credit card rather than paying with cash or by check (Feinberg, 1986; Prelec & Simester, 2001; Soman, 2001). One reason for the credit card premium is the fact that continued use of the credit card weakens memory of past expenses. Consequently, giving people feedback (reminders) on how much they have spent on prior expenses will improve mental tracking and mitigate spending differences (Soman, 2001).

In response to growing concerns about credit card debt, the government in South Korea (in 2010) mandated that credit card providers introduce a text messaging service that would remind consumers about transactions that they had recently incurred. In addition to prevent fraudulent use of credit cards, the policy was designed to "improve rational behavior in credit card spending" (Lee & Yoon, 2011, translated from Korean). In general, the expectation was that an introduction of the credit card spending text alert system would help people control their credit card spending and hence would reduce credit card balances.

An evaluation showed that the policy had the intended outcome for only about 12% of the population – the heavy spenders. For the remaining 88%, the policy actually resulted in an increase in spending (Kim *et al.*, 2020).

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

This backfiring effect might be due to an important difference in the manner in which the reminder was delivered in the pilot studies (where it was available on the same screen as that where the spending decisions were made) versus in the scaled-up intervention (where it appeared on a mobile device). This created a degree of 'digital dependency'(Sparrow *et al.*, 2011) by which consumers believed that they could easily access their past spending if they needed to, thereby reducing the motivation to mentally track it.

Example 2: backfiring effect of a retirement pension contribution intervention in Mexico

Mexico is facing an elderly poverty crisis (Fertig *et al.*, 2018). Salaried employees in Mexico make a 6.5% mandatory contribution to their pensions, while projections show that they need to make an additional 5% through voluntary contributions in order to retire comfortably. Unfortunately, the voluntary contribution rate in Mexico is abysmally low.

In working with the pension authority in Mexico, Shah *et al.* (2020) redesigned the quarterly pension statement that every salaried employee in Mexico receives. Prior research shows that simplifying communication and making it engaging will increase the likelihood that recipients of that communication will consume the information and act on it (Bhargava & Manoli, 2015). Accordingly, the redesigned quarterly statement was significantly more engaging than the original, and it differed in two ways. First, it provided a simplified visual illustration (in the form of a categorical thermometer) showing whether the recipient's current stock of pension savings was inadequate for retirement. Second, the statement included one of multiple behavioral interventions (gain versus loss framing, a wallet cutout to increase implementation intentions, an appeal that made the family's welfare salient and a fresh-start intervention encouraging recipients to start saving after a particular temporal landmark) that had been shown to be successful elsewhere. In a large-scale field trial with members of two pension funds, the intervention was a success with one of the funds (it increased the contribution incidence, the contribution amounts and the contribution frequency), but backfired in the other fund. Why did this happen?

Backfiring happened because of a specific design feature unique to the Mexican pension system. In particular, Mexicans needed to first choose a pension fund and then make contribution decisions. The quarterly pension statements displayed the performance of each fund in a tabular form. By making these statements more engaging, the researchers increased the attention that was being paid to the table. If the fund was high performing, then the engaging statement ended up improving voluntary contributions because it

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

increased the motivation to save. If, on the other hand, the recipient had chosen a low-performing fund, then the improved engagement of the statement actually resulted in demotivation because it highlighted attention to the low performance of the fund.

Example 3: heterogeneity in responses to text message reminders

In ongoing work with pension contributions in Mexico, Shah *et al.* (2019) compared a control condition with a condition in which recipients received the redesigned quarterly statement. In a third condition, they additionally received a text message with a variety of call-to-action messages. Text reminders have been used successfully to convert intention to action (Karlan *et al.*, 2016).

The results of the field trial showed that compared to the control, a condition in which recipients received a text message emphasizing their family's financial security in addition to the redesigned statement increased contribution rates significantly. However, do both men and women care equally about families? What about people who do not have families? Does age matter? Did it matter whether the recipient of the text message was an existing contributor or was a new contributor? Using the machine learning technique of causal tree and causal forest (Athey & Imbens, 2016), the researchers identified heterogeneous treatment effects across the scaled-up population. Their results were unsurprising, and they yet worked to show exactly for which sub-segments in the population the intervention was particularly successful and where it was not. For instance, they found that the family security text messaging increased contribution rates for people who were aged between 28 and 42. It did not have a significant effect for people above the age of 43, and it had a backfiring effect for people under the age of 27.

Knowing exactly where the family security text message (as well as each of the other messages that were tested) was most effective could now allow them to target the sub-segments of the population properly and customize the content of the text messages to different sub-segments in the population accordingly. Successful scaling, in this case, would embrace heterogeneity by using a family of interventions, each targeted to segments where it might be most effective.

Potential scaling challenges

We next illustrate how some of these scaling challenges might play a role if the intervention from a landmark paper (Hossain & List, 2012) were to be scaled. This paper used one of the most robust findings from behavioral science – the

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

framing of monetary outcomes as gains or losses. In a field study in a high-tech manufacturing facility, the authors manipulated the manner in which a productivity bonus was presented to workers. In particular, the same bonus was presented to a random subset of workers as a potential gain and to another subset as a potential loss. The authors found that both the gain and loss framing of incentives increased productivity compared to a control condition in which there were no incentives. Incentives clearly worked! More importantly, framing the incentive as a potential loss led to a small but significant increase in productivity over presenting it as a potential gain.[1]

While they found that loss aversion is manifested in a small treatment effect, it is quite important that this small increase in productivity happened in a real-work setting. If this small gain in productivity can be scaled up to large populations of workers, this would indeed have a significant impact on welfare. The question remains as to whether scaling up happens at exactly the same rate or whether there will be a voltage drop. Is it possible that the intervention might backfire?

It is noteworthy that Hossain and List did not find any significant framing effect for individual workers. The sign of the impact fluctuated between positive and negative across econometric specifications. If this intervention is scaled up with individual workers, it may in fact backfire. Moreover, in their study, the intervention was implemented for only about a month. While the impact was evident throughout the experiment, a negative framing of incentives might be less likely to be successful if the intervention is permanent or much longer in duration.

The effect of loss aversion on productivity depends on the interaction of the framing and the underlying incentive scheme (Hong *et al.*, 2015). Hence, one needs to be careful in choosing the economic part of the intervention while scaling up to ensure that it does not backfire. Moreover, the effectiveness of loss framing may also depend on the size of the economic incentives. In Hossain and List, the size of the bonus was above 20% of any of the workers' base salaries. The treatment effect for smaller-sized incentives may be insignificant or opposite in direction, as Hoffman and Thommes (2020) found for individual workers.

Why do we have scaling challenges?

Our views of why scaling challenges exist are similar to those of Al-Ubaydli *et al.*, and they are grounded in organizational realities. In particular, we

1 While a number of subsequent studies found similar framing effects in different contexts, a notable exception in Castelo *et al.* (2015), who found no framing effect.

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

would like to highlight three organizational reasons for this, as well as one reason to do with the nature of evidence.

We believe that there are three organizational reasons that drive some of the scaling challenges. The first reason is solution-mindedness. Unlike scientists, governments are typically under enormous time pressures to solve problems. Second, as the applications of our field have spread, organizations increasingly rely on non-specialists to design and deploy behavioral interventions. In an effort to help non-specialists, a number of heuristic frameworks (e.g., Easy, Attractive, Social, Timely (EAST); Service *et al.*, 2014) have been developed. Frameworks are elegant in that they allow a non-specialist to try to design interventions based on the learnings of others. However, as Soman (2017) argues, frameworks can often be counterproductive because they change the process of intervention design from one that begins with a careful audit of the context to a heuristic-type checklist-based approach. Third, most organizations that we have worked with operate in silos of capability. In the context of some of the work described here, the behavioral scientists typically are located in a different department from the data scientists and the design teams. Some of the scaling challenges will be best addressed if we have large and diverse teams working on these projects. However, organizational realities often make it difficult to achieve this result. Collectively, and perhaps most importantly, the cost of experimentation (including the costs of collecting and analyzing data, as well as organizational costs such as fear of failure, time commitment, organizational buy-in) in most organizations is high; therefore, they are more likely to embrace off-the-shelf results (Feng *et al.*, in press).

We also note that the nature of evidence in the behavioral sciences is significantly different from evidence in other sciences. Consider fields such as physics or medicine. As an example, objects released from a height will fall. They will do so irrespective of who releases them, the height of release, their color and whether there are other objects being released at the same time or not. This is because the theory of gravity has a large bracket of context surrounding it, and how the theory works in foreseeable but uncommon contexts is also well understood. For example, releasing the same object on the surface of the moon (or in outer space) will not result in the same sort of drop (there will be a 'voltage drop'). Nevertheless, the theory of gravity clearly predicts how the ball will drop, as we properly understand the difference between the two contexts. Unfortunately, unlike these sciences, the field of behavioral science has very narrow brackets of context. Therefore, it is important to emphasize that the documented robust demonstration of an effect in a particular context with a particular type of end user at a particular point in time does not guarantee that it will successfully replicate across those brackets of context.

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

Therefore, unlike the natural or medical sciences, where a rigorous meta-analysis of available evidence is typically enough to help predict what would happen in a scaled-up scenario, behavioral interventions often call for a sequential approach to evidence. A meta-analysis is helpful to serve as a starting point for developing interventions. These interventions would then probably need to be tested in low-cost and quick-win environments such as laboratory studies or online panels. The successful interventions in these domains might then be scaled into pilot studies – and then into larger pilots – before being then enshrined in policy. However, it is of utmost importance that these pilots are done in the context in which the intervention is going to be scaled up (*in situ* evidence). *In situ* testing in South Korea should have flagged the potential backfiring effect of reminders at a stage where it could have been corrected. Finally, scaling up must be accompanied by continuous testing, learning and adapting (Haynes *et al.*, 2013) in order to detect and respond to potential backfiring due to context changes, due to interactions between different interventions and due to heterogeneity. For this to happen, we need to push down the cost of experimentation in organizations.

We close with a few additional proposals:

Proposal A: Rather than viewing researchers purely as producers and governments purely as consumers of research, we should strive to co-create evidence.
Proposal B: Researchers should be incentivized to articulate all of the dimensions of the context (or features of the situation) under which their documented effects hold. They should further be incentivized to test for the effect under different contexts.
Proposal C: When *in situ* testing is not possible, we should document all of the dimensions of the context of the scaled-up situation that might be different from the testing phase and invite the original researcher to speculate on whether these differences will change the results.
Proposal D: We should test for heterogeneous treatment effects of interventions and use a family of appropriately targeted interventions to scale solutions.
Proposal E: We should encourage researchers and governments to create project teams of behavioral scientists, designers and data scientists to collectively design, test, scale and monitor interventions.
Proposal F: We should strive to reduce the cost of experimentation within organizations.

Disclaimer

Both authors are fellows at the Behavioural Economics in Action at Rotman (BEAR) research center and acknowledge support from the center.

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

References

Al-Ubaydli, O., J. A. List and D. Suskind (2019), The science of using science: towards an understanding of the threats to scaling experiments. *NBER Working Paper Series*, No. 25848.

Athey, S. and G. Imbens (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences of the United States of America*, **113**(27): 7353–7360. https://doi.org/10.1073/pnas.1510489113

Bhargava, S. and D. Manoli (2015), 'Psychological frictions and the incomplete take-up of social benefits: evidence from an IRS field experiment', *American Economic Review*, **105**(11): 3489–3529.

Castelo, N., E. Hardy, J. House, N. Mazar, C. Tsai and M. Zhao (2015), 'Moving citizens online: using salience & message framing to motivate behavior change', *Behavioral Science & Policy*, **1**(2): 69–81.

Feinberg, R. A. (1986), 'Credit cards as spending facilitating stimuli: a conditioning interpretation', *Journal of Consumer Research*, **13**(3): 348–356. doi: 10.1086/209074.

Feng, B., M. Kim and D. Soman (in press), 'Embedding behavioural insights in organizations', *Behaviorally informed organizations*, Toronto: University of Toronto Press.

Fertig, A., A. Fishbane and J. Lefkowiz (2018), Using behavioral science to increase retirement savings in Mexico: a look at what we have learned over three years. Available at: http://www.ideas42.org/wp-content/uploads/2018/11/I42-1046_MetLifeLatAm_paper_ENG_Final.pdf.

Haynes, L., O. Service, B. Goldacre and D. Torgerson (2013), Test, learn, adapt: developing public policy with randomized controlled trials. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf.

Hoffman, C. and K. Thommes (2020), Using loss aversion to incentivize energy efficiency in a principal-agent context – evidence from a field experiment, forthcoming, *Economics Letter*.

Hong, F., T. Hossain and J. A. List (2015), 'Framing manipulations in contests: a natural field experiment', *Journal of Economic Behavior and Organization*, **118**, 372–382.

Hossain, T. and J. A. List (2012), 'The behavioralist visits the factory: increasing productivity using simple framing manipulations', *Management Science*, **58**(12): 2151–2167.

Karlan, D., M. McConnell, S. Mullainathan and J. Zinman (2016), 'Getting to the top of mind: how reminders increase saving', *Management Science*, **62**(12): 3393–3411. https://doi.org/10.1287/mnsc.2015.2296

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**(42). doi:10.1186/1748-5908-2-42

Kim, J. K., J. H. Yoon, Y. H. Choi and D. Soman (2020), Do text reminders about credit card spending help reduce spending? A quasi-experimental evaluation. *Working paper*.

Lee, J. S. and Y. Y. Yoon (2011), Introduction of SMS transaction reminder (with cumulative spending). Available through: Financial Supervisory Service https://www.fss.or.kr.

Lichtenstein, S. and P. Slovic. eds., (2006), *The construction of preference*, New York: Cambridge University Press.

OECD, (2017), *Behavioural insights and public policy: lessons from around the world*, Paris: OECD Publishing. doi:10.1787/9789264270480-en.

Prelec, D. and D. Simester (2001), 'Always leave home without it: a further investigation of the credit-card effect on willingness to pay', *Marketing Letters*, **12**(1): 5–12.

Service, O., M. Hallsworth, D. Halpern, F. Algate, R. Gallagher, S. Nguyen, S. Ruda, M. Sanders (2014), EAST: four simple ways to apply behavioural insights. Available at: https://www.bi.team/wp-content/uploads/2015/07/BIT-Publication-EAST_FA_WEB.pdf.

Shah, A., M. Osborne, J. Lefkowitz, A. Fishbane and D. Soman (2019), Can making family salient increase financial savings? Quantifying heterogeneous treatment effects in voluntary

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press

retirement contributions using a field experiment in Mexico. Available at SSRN: https://ssrn.com/abstract=3460722 or http://dx.doi.org/10.2139/ssrn.3460722.

Shah, A., M. Osborne, J. Lefkowitz, A. Fishbane and D. Soman (2020), The simplification paradox: when reducing cognitive complexity can impede retirement savings contributions. Available at SSRN: https://ssrn.com/abstract=3460722 or http://dx.doi.org/10.2139/ssrn.3460722.

Soman, D. (2001), 'Effects of payment mechanism on spending behavior: the role of rehearsal and immediacy of payments', *Journal of Consumer Research*, **27**, 460–474.

Soman, D. (2015), *The last mile: creating social and economic value from behavioural insights*, Toronto: Rotman-UTP Publishing.

Soman, D. (2017), The elegant simplicity (and potential pitfalls) of simple frameworks, in OPRE Report 2017–23, *Nudging Change in Human Services*, Office of Planning, Research and Evaluation, U.S. Administration for Children and Families.

Soman, D. and C. Yeung (in press), Preface, *Behaviorally informed organizations*, Toronto: University of Toronto Press.

Sparrow, B., J. Liu and D. M. Wegner (2011), 'Google effects on memory: cognitive consequences of having information at our fingertips', *Science*, **333**(6043): 776–778.

Thaler, R. H. (2015), *Misbehaving: the making of behavioral economics*, New York: W.W. Norton & Company.

Thaler, R. H. and C. R. Sunstein (2008), *Nudge: improving decisions about health, wealth, and happiness*, Updated ed. New York, NY: Penguin Books.

Tversky, A. and I. Simonson (1993), 'Context-dependent preferences', *Management Science*, **39**(10): 1179–1189.

https://doi.org/10.1017/bpp.2020.24 Published online by Cambridge University Press