Behavioural Public Policy (2021), 5: 1, 117–124
© The Author(s) 2020. Published by Cambridge University Press doi:10.1017/bpp.2020.28

Field experiments and public policy: *festina lente*

GLENN W. HARRISON*
Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, Atlanta, GA, USA
School of Economics, University of Cape Town, Cape Town, South Africa

Abstract: The current state of the art in field experiments does not give me any confidence that we should be assuming that we have anything worth scaling, assuming we really care about the expected welfare of those about to receive the instant intervention. At the very least, we should be honest and explicit about the need for strong priors about the welfare effects of changes in averages of observables to warrant scaling. What we need is a healthy dose of theory and the implied econometrics.

Submitted 8 March 2020; accepted 28 March 2020

The problem of scaling results from one setting to another, whether that be an experimental setting or not, deserves formal attention. In environmental economics, there has been a long literature on 'benefit transfer': the transfer of valuations of environmental amenities in one location to valuations in another location. This can be called *horizontal* scaling. The latest challenge comes from field experiments, for the moment left undefined, being *vertically* scaled, to apply to wider and wider populations. And with that type of scaling to large-scale, we presumably deal with *temporal* scaling: seeing how behavior is affected over time.[1]

I do not see this as the fundamental missing link needed to see field experiments play a more significant role in public policy.[2] Nor do I see the latest

* Correspondence to: Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, Atlanta, GA, USA. E-mail: gharrison@gsu.edu.

[1] See Friedlander and Burtless (1995), Moffitt (1998) and de Haan and Lind (2018).

[2] These issues are not new. The earlier enthusiasm for social policy experiments in the USA in the 1970s spurred thoughtful assessments in Ferber and Hirsch (1978, 1982) and Hausman and Wise (1985). And the issues have garnered considerable attention already in the development field: see Banerjee *et al.* (2017).

fascination with 'replication crises' as particularly deep or constructive. Instead, my concern is with the glossing of the premise behind any kind of scaling at all: that the instant field experiment has delivered something worth scaling. For a variety of reasons, I offer some cautions about that premise, and I encourage attention to those issues before we get too far into scaling issues or the selling of scaling issues as the Rosetta Stone of behavioral public policy.

The expression 'field experiments' has come to mean two very different things in the economics literature. For one group, it just means any experiment conducted in the field that uses randomization; for others, it means any experiment that is conducted in the field or that uses field referents, whether or not randomization has been used. This semantic distinction does have some bite in terms of how people design experiments and what they expect to get out of them, so it is not 'just a semantic distinction'. I will argue for the complementarity of these two types of field experiments, in the interests of being diplomatic here.[3] If one adopts the latter definition from the start, one does not have to worry about complementing a randomized evaluation with structural insights from another type of field experiment. You simply design the field experiment to answer the policy question at issue, and it may or may not include a randomization component.

The problem with many field experiments is that they avoid wanting to make any structural claims about *why* things work or that they work to improve the *welfare* of individuals. The slogan sadly tells it all when someone proclaims loudly that they only care about *what* works. There is, to be sure, a cursory hand-wave at the theoretical literature on possible behavioral factors at work, but when it comes to what the evidence shows, and is intended to show, we get a net effect inside a theoretical black box. What is needed, in addition, are experiments to provide some structural insight into the processes at work. Which of the Big Four behavioral moving parts – in my view, risk attitudes, subjective beliefs, time preferences and social preferences – might account for observed behavior? Only if we obtain some estimates of these structural parameters[4] will we have any hope of describing why something is

[3] I speak more directly on this issue in Harrison (2013, §1). It is also proper to stress that randomized evaluations are excellent methods for doing what they set out to do in the narrow sense explained below. I want more interesting policy questions answered, and I fully expect that excellent answers to the less interesting questions will be needed as part of that broader policy objective.

[4] Some field experimenters do try to elicit measures of preferences, such as Jakiela and Ozier (2019) for risk preferences, but they do so using hypothetical survey instruments that have well-documented biases. But the effort to elicit preferences is laudable and can easily be implemented correctly. At least with respect to the use of incentives, good examples include Fisman *et al.* (2017) and Balakrishnan *et al.* (2020).

working or not, and then going further and undertaking a welfare evaluation.[5]

There are several reasons why most field experiments fall short. One is that they limit themselves to evaluations of *observables*: this price change in delivering that product leads to what revealed change in demand? Another is that they limit themselves to *average* treatment effects (ATEs): what is the average change in demand?

General concerns with any instant experiment

In virtually all cases that I am aware of, attention has been trained on making causal statements about observables. In the end, I do not care about such causal statements. Call me old-fashioned, but I care about whether the policy intervention is doing harm to individuals, as measured by the subjective consumer surplus.[6] In turn, shame on me, in order to calculate the consumer surplus, I need to know some things about the individual, such as preferences and beliefs. And also in turn, further shame on me, I probably need to make a fair number of parametric assumptions to map estimates of preferences and beliefs into consumer surplus. You can visualize the methodological fingers wagging now. Who knows if that theory of preferences and beliefs is the right one? Who knows if those parametric assumptions are the right ones? I put my hands up: you got me.

But then I get ethical about things, to turn the heat back on the methodological nay-sayers. How on earth do you know – or even have a reasonable prior – that your intervention is doing no harm? It sure is not theory, since plumbers do not need any serious theory. Just as a wind tunnel[7] is literally not the same thing as field turbulence, would you ever step in a plane that had not survived the most elementary tests in a wind tunnel? Of course, you just assume that this is done, and we all know about the illusion of risk regulation in some quarters when it comes to aircraft safety. And there are some simple, canonical examples where successful improvement of the intended observable outcome leads to welfare losses.[8] I pose the same challenge to

[5] Advocates of randomized interventions sometimes pose a false dichotomy between 'all-in theological' modeling via structural assumptions or 'agnostic eyeballing' of the average effects: Heckman (2010) takes aim squarely at this false tradeoff.

[6] In many instances of importance, such as the purchase of insurance, it is *expected* consumer surplus that is relevant. That should just be understood, and it has nothing to do with whether or not the insurance actually generates a payment to the individual, as some believe.

[7] https://en.wikipedia.org/wiki/Wind_tunnel.

[8] The best example is insurance, where take-up has been the observable metric, under the assumption that it automatically tells us that there is a subjective welfare improvement. Relaxing the assumption of naive revealed preference, in the spirit of behavioral welfare economics, one finds that

those that would run a field experiment in the first place, and I pose it even more loudly and urgently to those that would seek to scale the thing up.

Gaussianity, and the lure of the ATEs

From experience, I know that economists hear 'social welfare' when anyone utters the word 'welfare', so it is important to stress that the word is used here to refer to welfare of the individual. I want to measure the *distribution* of individual welfare effects of a policy. If one then looks, descriptively and in passing, at the average of this distribution, that should not be mistaken for any specific social welfare function at play. On the other hand, if one *only* looks at the average, then one has lashed any policy advice to the mast of a problematic social welfare function.

A more important normative implication flows from looking at distributions. One might uncover identifiable sub-populations that do benefit from some intervention, allowing behaviorally smart *conditional* interventions. What if men lose from an intervention and women gain, and one is allowed (legally and culturally) to condition on gender? Are we to throw that information away on the altar of parsimony, when we only look at the average effect? No, and the better studies know this, but most do not.

A related point, of course, is about the econometric methods used to evaluate average effects. I admit to being the person in seminars and referee reports that complains about ordinary least squares on any dependent variable that is not defined over ±∞.[9] And that reminds everyone that even Angrist and Pischke (2009, chapter 7) have a place in their hearts for quantile regressions.[10] And is *anyone* concerned that known randomization might generate a familiar

increased take-up is generally associated with welfare losses (Harrison, 2019; Harrison & Ng, 2016, 2018, 2019).

[9] One constructive solution in advanced graduate classes is to have students replicate estimates of influential papers in major journals, ideally on a topic they are writing their thesis on: these days, one finds good documentation of data and code online. Then have them apply the methods they learned about in their first graduate econometrics class. You know the methods: probit for binary dependent variables, beta or fractional regressions for dependent variables in the unit interval and hurdle models (*not* tobit) for dependent variables with histogram-evident spikes at zero or some other point – and average marginal effects, of course. The one that often jumps out with different results is the modest hurdle model, a staple of health econometrics, but not as widely used as it should be. In literal terms, this is not a replication crisis, thank goodness, but a methodological crisis nonetheless.

[10] There have also been advances in *flexible* parametric quantile regression methods, due to Frumento and Bottai (2016, 2017) and implemented by Bottai and Orsini (2019). Similarly, there have been advances in the econometric methods for comparing unconditional distributions, due to Goldman and Kaplan (2018) and implemented by Kaplan (2019). Carter *et al.* (2019) offer a great illustration of how one can use information 'in the tails' to drive better development policy.

sample selection bias, as well as an attrition bias, with respect to heterogeneous risk preferences?[11]

The deeper implications of all of this attention on experimental design in field experiments is the debate over experimental methods versus observational methods.[12] On a good day in a field experiment, randomization ensures 'covariate balance', in the sense that the potentially confounding covariates with respect to the effects of the treatment on the outcome are then the same, or balanced, when one looks at the treated sample and the untreated sample. Potential outcomes for treated and untreated are imputed by looking at averages of 'similar individuals' that do or do not receive the treatment, where similarity might be determined by a scored propensity to be one of the treated. So in this instance we can say that the measured effect is solely due to the treatment.

Three issues arise here. The first is that pure randomization may be an inefficient way to generate data for the inferences required.[13] The second is the amusing combination of formal univariate tests and informal cross-variate eyeballing to determine balance, as if the covariances are all zero. The third is that there is no real debate between experimental methods and observational methods: they are complementary. Powerful statistical tools, largely developed and fully reviewed by Imbens and Rubin (2015), allow non-experimental,

[11] See Harrison *et al.* (2009, 2020).

[12] Also a debate long ago: see Hausman and Wise (1985) and Moffitt (1986).

[13] Alternatives to pure randomization when matching treated and untreated build on the idea that complex survey designs might randomize within clusters, which are paired with other clusters according to certain characteristics prior to randomization to treatment. Thus, a whole cluster receives the treatment or not, and individuals within that cluster are evaluated and, since they are in the same cluster, are not statistically independent. But when the clusters are blocked, prior to random allocation to treatment, there is valuable information contained in the blocking criteria (e.g., urban or rural). This blocking information can be used, and it is explicitly not used in propensity score or nearest-neighbor matching. One extreme is a fully blocked randomized experimental design, which matches individuals on certain covariates exactly prior to treatment; and the other extreme is a completely randomized experimental design that matches no individuals and just applies a constant probability to each individual of being assigned to the treatment. Treating the former as if it were the latter with a post hoc matching method is clearly inefficient. The same argument for *relative* inefficiency then applies to *partially* blocked designs. The upshot is that pair matching can be used to improve matching methods rather than just using propensity scores: see Imai *et al.* (2009) and Iacus *et al.* (2011). Apart from complex survey design, theory or 'priors' might provide a basis for matching pairs according to certain characteristics. This motivation leads to methods referred to as 'coarsened exact matching', implemented by Blackwell *et al.* (2009). The idea is to coarsen the covariates of outcomes and match exactly on those coarsened covariates, but use the actual values of the covariates when evaluating treatment effects of matched observations. The process of coarsening is akin to how histograms are generated from continuous or multi-valued observations: automatic algorithms can be applied (e.g., choose 10 equally spaced bins to span the range of data) or priors used to define cutpoints (e.g., bin those under 18 together, then those between 18 and 30, etc.).

observational data to be transformed in a way that allows them to approximate a randomized experimental evaluation. The tension between those advocating the use of controlled experiments to answer behavioral questions and those advocating the use of non-experimental or observational data to answer behavioral questions is, perhaps sadly for some, entirely contrived.[14] Each has a methodological role, strengths and limitations.

Conclusion

Scale effects matter, and I welcome the effort by Al-Ubaydli *et al.* to be formal about it and to draw on insights from other fields.[15] I am not sure that much more is really needed in terms of formalism beyond understanding the

[14] It is important, then, to understand that sometimes this tension can be contrived for unethical reasons. Coller *et al.* (2002) used a database to calculate a smoking-attributable fraction (SAF) for health expenditures in the USA, on behalf of plaintiffs in tobacco litigation. Working on behalf of tobacco companies, Rubin (2000, 2001b) argues that the database does not exhibit covariate balance when one considers current or former smokers compared to never-smokers (2000, table 2, p. 338) or when one considers male current smokers, female current smokers, male former smokers and female former smokers, each compared to male or female never-smokers (2001b, Table 2, p. 179). In each case, he considers a binary treatment; in neither case does he consider whether these claims of covariate imbalance have any effect on the estimated SAF for these binary groups, or whether one could constructively adjust for them (e.g., by retaining "the more detailed smoking information for regression adjustment," as he suggests; 2000, p. 338). In fact, Rubin (2001b, p. 169) makes a point of arguing that one should decide on how to analyze data prior to seeing information on outcomes in a remarkable statement for a scholar working in a litigation context to make: "Arguably, the most important feature of experiments is that we must decide on the way data will be collected before observing the outcome data. *If we could try hundreds of designs and for each see the resultant answer, we could capitalize on random variation in answers and choose the design that generated the answer we wanted!* The lack of availability of outcome data when designing experiments is a tremendous stimulus for 'honesty' in experiments and can be in well-designed observational studies as well" (emphasis added). One can see here the idea of trying to parallel a practice of registering the design and hypotheses of a randomized controlled trial before running the trial. Whatever the merits of this practice – and that is a debate for another day – recommending this practice is arguably disingenuous *if one is going to then argue* that because the "raw data" in an observational study are unbalanced for measuring *some* treatment effects, it will measure *any* effects unreliably, those data cannot be easily corrected to provide reliable measurements in terms of recovering covariate balance or one should even just discard the dataset entirely. These are all positions taken by defendants in recent tobacco litigation. Rubin (2001a, p. 1409) reports calculations with what he admits is a "highly artificial example," and he concludes that there could be a *negative* SAF if one corrected for covariate balance. But as he clearly emphasizes, these were completely contrived numbers to illustrate a logical point. Presumably, the most telling evidence for unreliability would have been a subsequent demonstration that standard applications of the procedures advocated lead to qualitatively and quantitatively different answers with real data, and no such demonstrations have ever seen the light of day (one reason for this: they do not generate different answers at all). This substantive issue is, of course, central to one of the most significant public health issues ever.

[15] As long as there is no suggestion that 'gold standards' in one field should automatically be accepted in other fields: see Harrison (2011, §1.4) for elaboration.

production function for cognition,[16] in this case behavioral policy insights. But we are not ready to scale anything, and I fear for the welfare consequences of doing so on the basis of what passes for field experiments today.

Acknowledgments

I am grateful to Karlijn Morsink for valuable comments.

References

Angrist, J. D. and J.-S. Pischke (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.

Balakrishnan, U., J. Haushofer, and P. Jakiela (2020), 'How Soon Is Now? Evidence of Present Bias from Convex Time Budget Experiments', *Experimental Economics*, **23**: 294–321.

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

Blackwell, M., S. Iacus G. King, and G. Porro (2009), 'cem: Coarsened Exact Matching in Stata', *Stata Journal*, **9**(4): 524–546.

Bottai, M. and N. Orsini, (2019), 'qmodel: A Command for Fitting Parametric Quantile Models', *Stata Journal*, **19**(2): 261–293.

Camerer, C. F. and R. Hogarth (1999), 'The Effects of Financial Incentives in Experiments: A Review and Capital-Labor Framework', *Journal of Risk and Uncertainty*, **19**: 7–42.

Carter, M. R., E. Tjernström, and P. Toledo (2019), 'Heterogeneous Impact Dynamics of a Rural Business Development Program in Nicaragua', *Journal of Development Economics*, **138**: 77–98.

Coller, M., G. W. Harrison, and M. M. McInnes (2002), 'Evaluating the Tobacco Settlement: Are the Damages Awards Too Much or Not Enough?" *American Journal of Public Health*, **92**(6): 984–989.

de Haan, T. and J. Lind (2018), "Good Nudge Lullaby": Choice Architecture and Default Bias Reinforcement', *Economic Journal*, **128**(610): 1180–1206.

Ferber, R. and W. Z. Hirsch (1978), 'Social Experimentation and Economic Policy: A Survey', *Journal of Economic Literature*, **16**(4): 1379–1414.

Ferber, R. and W. Z. Hirsch (1982), *Social Experimentation and Economic Policy*, New York: Cambridge University Press.

Fisman, R., P. Jakiela, and S. Kariv (2017), 'Distributional Preferences and Political Behavior', *Journal of Public Economics*, **155**: 1–10.

Friedlander, D., and G. Burtless (1995), *Five Years After: The Long-Term Effects of Welfare-to-Work Programs*, New York: Russell Sage Foundation.

Frumento, P., and M. Bottai (2016), 'Parametric Modeling of Quantile Regression Coefficient Functions', *Biometrics*, **72**: 74–84.

Frumento, P., and M. Bottai (2017), 'Parametric Modeling of Quantile Regression Coefficient Functions with Censored and Truncated Data', *Biometrics*, **73**: 1179–1188.

Goldman, M. and D. M. Kaplan (2018), 'Comparing Distributions by Multiple Testing Across Quantiles or CDF Values', *Journal of Econometrics*, **206**(1): 143–166.

[16] See Camerer and Hogarth (1999).

Harrison, G. W. (2011), 'Randomisation and Its Discontents', *Journal of African Economies*, **20**(4): 626–652.

Harrison, G. W. (2013), 'Field Experiments and Methodological Intolerance', *Journal of Economic Methodology*, **20**(2): 103–117.

Harrison, G. W. (2019), 'The Behavioral Welfare Economics of Insurance', *Geneva Risk & Insurance Review*, **44**(2): 137–175.

Harrison, G. W. and J. M. Ng (2016), 'Evaluating the Expected Welfare Gain from Insurance', *Journal of Risk and Insurance*, **83**(1): 91–120.

Harrison, G. W., and J. M. Ng (2018), 'Welfare Effects of Insurance Contract Non-Performance', *Geneva Risk and Insurance Review*, **43**(1): 39–76.

Harrison, G. W., and J. M. Ng (2019), 'Behavioral Insurance and Economic Theory: A Literature Review', *Risk Management & Insurance Review*, **22**: 133–182.

Harrison, G. W., M. I. Lau, and E. Rutström (2009), 'Risk Attitudes, Randomization to Treatment, and Self-Selection into Experiments', *Journal of Economic Behavior and Organization*, **70**(3): 498–507.

Harrison, G. W., M. I. Lau, and H. Il Yoo (2020), 'Risk Attitudes, Sample Selection and Attrition in a Longitudinal Field Experiment', *Review of Economics & Statistics*, **102**(3): 552–568.

Hausman, J. A. and D. A. Wise (1985), *Social Experimentation*, Chicago: University of Chicago Press.

Heckman, J. J. (2010), 'Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy', *Journal of Economic Literature*, **48**(2): 356–398.

Iacus, S., G. King, and G. Porro (2011), 'Multivariate Matching Methods That Are Monotonic Imbalance Bounding', *Journal of the American Statistical Association*, **106**(493): 345–361.

Imai, K., G. King, and C. Nall (2009), 'The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation', *Statistical Science*, **24**(1): 29–53.

Imbens, G. W. and D. B. Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press.

Jakiela, P., and O. Ozier (2019), 'The Impact of Violence on Individual Risk Preferences: Evidence from a Natural Experiment', *Review of Economics & Statistics*, **101**(3): 547–559.

Kaplan, D. M. (2019), 'distcomp: Comparing Distributions', *Stata Journal*, **19**(4): 832–848.

Moffitt, R. (1986), 'Review of Social Experimentation', *Journal of Political Economy*, **94**(5), : 1121–1126.

Moffitt, R. (1998), 'Review of *Five-Years After: The Long-Term Effects of Welfare-to-Work Programs*', *Industrial Labor Relations Review*, **51**(2): 327–329.

Rubin, D. B. (2000), "Statistical Issues in the Estimation of the Causal Effects of Smoking Due to the Conduct of the Tobacco Industry," in J.L. Gastwirth (ed.), *Statistical Science in the Courtroom*, New York: Springer-Verlag.

Rubin, D. B. (2001a), 'Estimating the Causal Effects of Smoking', *Statistics in Medicine*, **20**: 1395–1414.

Rubin, D. B. (2001b), 'Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation', *Health Services & Outcome Research Methodology*, **2**: 169–188.