Finding Our Balance?
Revisiting the Randomization Revolution in Development Economics Ten Years Further On

Christopher B. Barrett, Cornell University[*]
and
Michael R. Carter, University of California at Davis

Abstract: Ten years ago, we offered some reflections on the power and pitfalls of randomized controlled trials in development economics, arguing that the research community had lost its balance between theory, observational data and randomized experiments. We remain convinced of both the importance and the limits of RCTs for development economics research. But with another decade of RCTs under our collective belts, three issues now strike us as having become increasingly important. First, ethical risks still loom large. Second, increasing evidence that many interventions have highly heterogeneous impacts, places a premium on reintegrating *ex ante* theorizing with RCT methods to understand the heterogeneity. In some cases, heterogeneity may imply RCTs are less desirable than other research methods. Third, the increasing use of RCTs to study informational, behavioral, and other perceptions-mediated interventions creates an opportunity for non-classical measurement error problems that undercut the statistical power of seemingly well-designed studies in ways that remain underappreciated.

A decade ago we explained why the randomized controlled trials (RCT) revolution in development economics was generally a good thing, that "[t]he rise of RCTs and their associated methods has indisputably raised the bar in applied economic analysis, and appropriately enhanced the scrutiny of identification strategies for inferring causal effects. The most sophisticated proponents never trumpeted these methods as a panacea, merely as progress" (Barrett and Carter 2010). We also emphasized the pitfalls and significant limitations of RCTs for addressing core development questions, advocating in particular for "greater attention … to the ethics of employing RCTs, the greater utilization of behavioral economic experiments to help resolve some of the same fundamental identification problems that motivate reliance on RCTs, and the more thoughtful use of nonexperimental findings to structure the behavioral models that should underpin good RCT design," and developed those thoughts further in a follow-on paper (Barrett and Carter 2014). We remain convinced of both the importance and the limits of RCTs for rigorous and impactful development economics research.

We offer three brief additional thoughts, building on the benefit of a decade more direct experience of conducting and observing RCTs. First, we emphasize again the ethical burdens

[*] Corresponding author at cbb2@cornell.edu, 340D Warren Hall, Cornell University, Ithaca, NY 14853-7801 USA.

intrinsic to RCTs. The enhanced agency an experimenter enjoys to randomize one or more feature(s) of subjects' world inextricably brings correspondingly enhanced opportunity to harm subjects. In our view, the social science community relies excessively on institutional review boards (IRBs) to prevent bad behavior. IRBs are important but insufficient, not least of which because they approve designs ex ante of research implementation and rely on full and frank reporting by investigators, including of post-approval design modifications. In our experience, such reporting is rare indeed, while modifications are near-ubiquitous. Many researchers seem not to recognize that subjects' informed consent in no way absolves researchers of responsibility for any injury subjects, or third parties, suffer as a direct result of an intervention. Moreover, we have been struck that heightened concerns about research ethics concerning reproducibility, RCT registry, pre-analysis plans, etc. do not seem to have been accompanied by similar consternation about the injuries too often directly done in intervention-based research. We aspire that journal reviewers and editors will more actively screen for ethically dubious research ex post, refusing to publish such studies. To be clear, ethical transgressions are by no means unique to RCTs; but the likelihood of harm increases with the degree of intervention of the researcher.

Second, in our earlier work, we stressed that when the impact of a given program treatment (*e.g*., a liquidity injection) is structurally or essentially heterogeneous (because, say, some recipients are liquidity constrained and others are not), then even unbiased, internally valid treatment estimates are merely data-weighted averages of heterogenous impacts. Of course, heterogeneity can be informatively identified in a purely empirical fashion (Bandiera *et al*. 2017; Carter *et al*. 2019).[1] But while the conditional quantile estimation methods used by these authors can help us identify heterogeneity *ex post*, the enduring challenge is to use theory and critical thinking *ex ante* to design studies sufficiently powered to measure and detect the structural sources of heterogeneity across distinct sub-populations.

In the case of the studies like that of Bandiera *et al*. that estimate the impact of programs intended to reduce extreme poverty, theoretical work on multiple equilibrium poverty traps offers precise insights into the likely sources of structural heterogeneity (see the discussion in Barrett *et al.* 2019). Failure to measure those dimensions (*e.g*., beneficiaries' mental health and sense of agency or self-efficacy), and to power a study adequately for inference on the different, identifiable sub-populations who experience different program impacts, implies that even a well-implemented RCT will be less powerful and useful than it might seem. More generally, a more careful integration of theory with empirical work not only opens the door to a more balanced epistemology – results are credible not just because they are statistically significant, internally unbiased parameter estimates, but because they conform with a compelling model of behavior – but it also allows a more reasoned approach to the question of inference to populations different from that of the particular study, as Deaton (2019) suggests.

[1] While the Bandiera *et al*. (2017) study of BRAC's multi-dimensional ultra-poor poverty reduction program in Bangladesh estimates quite substantial average treatment effects, they find using conditional quantile analysis that for key indicators such as consumption and asset accumulation, the program had zero impacts for at least a third of the beneficiary population (and much larger than average impacts for another sub-population). Carter *et al.* (2019) find a similar distribution of impacts in their study of an asset and skills transfer program in rural Nicaragua.

Development research typically aims to understand the structures that render people poor and what interventions might help people surmount those obstacles. For example, many agencies champion connecting semi-subsistence farmers to commercial value chains. But we know that agri-food companies typically intentionally select the supplier farmers with whom they contract (Barrett *et al.* 2012). The resulting selection bias confounds causal inference of the impact of contracting on smallholders' wellbeing in observational data (Bellemare and Bloem 2018). One way to obviate that problem would be to randomize contract offers among farmers, thereby breaking the selection mechanism, if one could convince a firm to do so. But that naïve, brute force approach would almost surely yield downwardly biased estimates of the true causal effects of contract farming on farmers ever likely to elicit a contract because firms offer contracts based on information that leads them to believe that particular farmers will do at least as well – for the firm and for the farmer – under contract as would unselected farmers. So a superior research design might be to work with the firm to establish their actual selection mechanism(s), and then control explicitly for – or exploit the discontinuities created by – the actual, known process(es) that generate(s) structural heterogeneity in the impacts of contracting between farms offered and not offered contracts. One can then deal with the farmer's endogenous selection to accept or reject the contract offer using an intent-to-treat estimator based on the selection-corrected contract offer. If we take theory seriously in identifying structural heterogeneity ex ante of empirical research design, RCTs may be statistically second-best approaches if we can expressly identify and control for the true selection mechanisms that underpin actual outcomes.

Third, in the last 10 years, we have not only witnessed increasing recognition of heterogeneous treatment effects, but also the increasing heterogeneity of treatments themselves, which is closely related to the 'faux exogeneity' point we raised a decade ago. The rise of informational and behavioral treatments – such as offering farmers a contract or information on a new technology – carries a non-classical measurement error (NCME) risk not commonly recognized. An RCT that involves tangible, unmalleable interventions – e.g., giving all treated children the same deworming medication, providing all farmers with exactly the same fertilizer blend – has a known, randomized treatment that not only resolves selection bias problem but also is invariant across treated research subjects. But many RCT interventions – *e.g.*, those that provide information, that create new social connections, that change a rule, that offer a contract, etc. – are designed around treatments that are uniform only from the researcher's perspective: a standardized video, written rule, loan agreement, *etc*.

The oft-overlooked danger is that treated subjects' perceptions of the intervention – the essential mediators that the experimenter posits may affect behaviors and thereby outcomes – are likely to vary among treated subjects. For example, one entrepreneur, uncertain of his own enterprise's viability but suddenly extended a randomized offer of credit, may perceive the bank decision as expert approval of a business model and plan, while another randomized credit recipient with greater confidence in her strategy interprets the same treatment merely as a loan offer. Or some farmers may interpret an NGO training as a directive to implement the new method as a quid pro quo for other services in the future, while others do not (Usmani *et* *al.* 2018). Moreover, variation in perceptions is almost surely driven by observable and unobservable characteristics. This implies not only measurement error in the treatment – subjects perceive different things in the intervention – but also that the measurement error is non-classical, likely correlated with other control variables or with unobservables also correlated with the outcome(s) of interest. We typically cannot identify whether the bias generated by NCME is positive or negative, and in the presence of multiple variables subject to NCME, as is true of much experimental and survey data, correction for the error in just one measure can aggravate rather than reduce the bias (Abay *et al.* 2019). RCTs that rely on perceptions-mediated treatments typically lack the statistical rigor that their most ardent proponents claim.

The challenge and privilege of doing development research stems from the importance of the structures that empower some of us to study the impediments to other people's progress. RCTs are a valuable tool to help us more rigorously identify what truly causes progress by obviating structural obstacles. But RCTs are not always the tool best suited to the task, they demand more careful ex ante theorizing, and carry with them added ethical burdens to which researchers are obliged to attend more thoughtfully.

References

Abay, Kibrom A., Gashaw T. Abate, Christopher B. Barrett, and Tanguy Bernard (2019), "Correlated Non-Classical Measurement Errors, 'Second Best' Policy Inference and the Inverse Size-Productivity Relationship in Agriculture," *Journal of Development Economics* 139: 171-184.

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul and Munshi Sulaiman (2017). "Labor Markets and Poverty in Village Economies," *Quarterly Journal of Economics* 132(2): 811-870.

Barrett, Christopher B., Maren E. Bachke, Marc F. Bellemare, Hope C. Michelson, Sudha Narayanan, and Thomas F. Walker (2012). "Smallholder participation in contract farming: comparative evidence from five countries." *World Development* 40(4): 715-730.

Barrett, Christopher B. and Michael R. Carter (2010), "The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections," *Applied Economic Perspectives and Policy* 32 (4): 515-548.

Barrett, Christopher B. and Michael R. Carter (2014), "Retreat from Radical Skepticism: Rebalancing Theory, Observational Data and Randomization in Development Economics," chapter 3 in Dawn L. Teele, editor, *Field Experiments and Their Critics* (New Haven, CT: Yale University Press).

Barrett, C.B., M.R. Carter and J.-P. Chavas (2019). "The Economics of Poverty Traps," in C. Barrett, M.R. Carter and J.-P. Chavas (eds) *The Economics of Poverty Traps* (University of

Chicago Press & NBER).

Bellemare, Marc F., and Jeffrey R. Bloem (2018). "Does contract farming improve welfare? A review." *World Development* 112: 259-271.

Carter, Michael R., Emilia Tjernström, and Patricia Toledo (2019). "Heterogeneous impact dynamics of a rural business development program in Nicaragua." *Journal of Development Economics* 138: 77-98.

Deaton, A. (forthcoming). “Randomization in the tropics revisited: a theme and eleven variations,” in Florent Bédécarrats, Isabelle Guérin and François Roubaud, eds., *Randomized controlled trials in the field of development: A critical perspective*, Oxford University Press.

Usmani, Faraz, Marc Jeuland, and Subhrendu K. Pattanayak (2018). “NGOs and the effectiveness of interventions”. UNU-WIDER Working Paper № 2018/59.