Randomized interventions and “real” treatment effects: A cautionary tale and an example

Bulte, E., Di Falco, S., & Lensink, R.

Randomized interventions and "real" treatment effects: A cautionary tale and an example

Erwin Bulte[1]*, Salvatore Di Falco [2] and Robert Lensink[3]

[1] Development Economics Group, Wageningen University, 6700 KN Wageningen, Netherlands, E-mail: Erwin.bulte@wur.nl (* corresponding author)

[2] Institute of Economics and Econometrics, Geneva School of Economics and Management, University of Geneva, Switzerland, Email: Salvatore.DiFalco@unige.ch

[3] Faculty of Economics and Business, University of Groningen, 9700AV Groningen, Netherlands, and Development Economics Group, Wageningen University, 6700 KN Wageningen, Netherlands, E-mail: B.W.Lensink@rug.nl

Abstract: The experimental approach has revolutionized development economics. Nonetheless, randomization cannot do everything. We discuss challenges to RCTs, paying special attention to internal validity. Randomized interventions in social sciences are not double-blind and do not, in general, hold all relevant covariates constant. Treated and untreated subjects adjust their behavior in response to treatment status. Disentangling the treatment effect into its behavioral component and the direct effect of the intervention is difficult, and implies a return to the toolkit of observational studies. This is illustrated using improved seed distribution in African farming. While standard RCTs found large treatment effects, double-blind RCTs revealed that a large share of this impact is due to farmers allocating extra effort and their best plots to the cultivation of new seeds.

Acknowledgments: We thank Arun Agrawal, Michael Carter, Gunnar Köhlin, Travis Lybbert, and jean-Philippe Platteau for comments and suggestions. The usual disclaimer applies.

The experimental approach has revolutionized development economics and catapulted it from the periphery to the mainstream in economics. Such transformation meets resistance, and the literature contains several critical discussions of the RCT's rise to dominance (e.g. Barrett and Carter, 2010, and Deacon, 2019). Two well-known challenges to the experimental method are its inability to address several first-order issues (e.g. trade or monetary policy) because not all problems can be broken up into randomizable sub-problems, and the fact that the external validity of individual RCTs may be limited. What works here may not work there, and we need many RCTs – each case study generating one observation – to learn something about the generalizability[1] of the insights obtained in any particular setting.

In this piece we focus on a less-known challenge, namely the lack of internal validity of many RCTs. In the early days of the experimental revolution in economics, some 10-15 years ago, RCTs were hailed as the 'gold standard', reflecting its undisputed status in modern medicine. However, there is a large difference between RCTs in medicine and in economics. The former are double-blind so that neither the subject nor the analyst knows the treatment status of the subject. Double-blinding is the preferred approach to eliminate placebo effects, or other effects that would bias the average treatment effect. Needless to say that in economics double-blinding is not the norm – participants in an experiment typically know whether they receive the treatment (e.g. participate in a training, receive a transfer), or not.

Herein lies a challenge. If subjects know their treatment status, they may adjust their behavior. In fact, we often *expect* them to change their behavior in order to fully benefit from the treatment. While the analyst randomizes subjects into treatment, she therefore cannot assume that all relevant covariates are held constant across treatment arms. Disentangling the overall treatment effect that is measured into its behavioral component and the direct effect of the intervention is difficult (see Barrett and Carter 2010, Chassang et al. 2015).

For example, Beaman et al. (2013) describe a randomized intervention where free fertilizer was distributed to a sample of female farmers in Mali. Treated farmers "re-optimized" farm management, and started to use more hired labor and herbicides. As a result, yields adjusted. Beaman et al (2013) conclude such behavioral responses are very common, and "plague most empirical studies in this literature". Is it problematic to conflate the direct effect of the intervention and the behavioral change induced by the intervention?

In what follows we argue that this may, indeed, be problematic, and may lead to biased welfare assessments. The case we consider is the provision of improved ('modern') seed varieties to smallholders in Tanzania. Specifically, we organized a study where we varied two dimensions:

- Half of the farmers participated in *a standard RCT*, where half of these farmers randomly received improved seed, and the remaining farmers received a traditional variety.
- The other half of the farmers participated in *a double-blind RCT* where they received either improved or traditional seed, but were not fully informed about the seed they received. They were told there was a 50% probability that they received the improved seed, and a 50% probability that they received the traditional variety.

[1] For an analysis on the issue of replicability refer to Butera and List (2019).

To disguise the nature of the seed in the double-blind study we had to make the different seed look the same. This entailed dusting all seed with purple fungicide (as was already done with the improved variety). In short, we have four groups of farmers: modern-informed, modern-uninformed, traditional-informed and traditional-uninformed. This design enables us to probe the relevance of behavioral responses in the context of crop-based agriculture.[2] We organized the same experiment twice at different points in time: once with maize, and once with cowpeas. The results were similar.

For the current purpose we restrict the analysis to the subsample of farmers receiving the modern seed, and consider whether uncertainty about seed type affects harvest levels. We obtained a measure of output immediately after the growing season, and try to explain variation in observed harvest levels by a treatment dummy (i.e. whether the smallholder in question was certain about his seed type) and a vector of demographic controls and village fixed effects.

For both the maize and cowpea study we find that harvest levels are consistently lower in the double-blind experiment, or when smallholders are not fully informed about the seed they received. Uncertainty reduces the average harvest of improved cowpeas seed by 10-20% (Bulte et al. 2014) and the harvest of improved maize seeds by 25% (Bulte et al. 2019). About one quarter of the increment in harvest levels that was measured in the standard RCT should not be attributed to the improved seed – it is the result of farmers changing their behavior.

What do farmers do differently if they know they received the modern seed? This is a difficult issue, because smallholders often exploit multiple margins. They may allocate more (or 'better') labor to the preferred plot, or adjust the quantities of other complementary inputs – fertilizer, irrigation, herbicides, and so on. They may also decide to grow the modern seed on their best plot (in terms of soil quality, say). We tried to capture the smallholders' response along these dimensions, and more, and were indeed able to pick up some of the channels through which farmers adjust. A key channel is the supply of (family) labor to the plot in question. We were unable to explain most of the variation due to behavioral responses—perhaps because our measures of behavior were too crude, or perhaps because part of the adjustments extend beyond the dimensions we readily observe and measure in the field.

Where does this leave us? One could argue that economic RCTs should not be double-blind, and that behavioral responses, such as the 'crowding in' of labor, are part and parcel of what we try to measure. A double-blind experiment would prevent farmers from responding optimally, and would therefore generate an *under-estimate* of the true welfare gains associated with the intervention. We believe this line of argument is valid. But it is only part of the story, because a naïve comparison of harvest levels in a standard RCT would *over-estimate* the true gains. This is where the challenge to internal validity originates.

We believe there are two plausible ways forwards. First, we may try to keep track of *all the behavioral adjustments* along all possible dimensions, and properly account for the (opportunity) cost of these adjustments. This enables computation of the true welfare gain. This is the high road, in theory, but it is complicated by the fact that behavioral adjustments may be many and sometimes hard to predict or measure (as we learned the hard way).

[2] For an alternative approach to disentangle the direct treatment effect, the behavioral response, and a possible interaction effect, refer to Chassang et al. (2015).

Second, we may try to move up from studying the specific activity that is 'treated' (maize cropping) to a higher level of aggregation where the various costs and benefits of all behaviors come together. For example, rather than studying maize harvests or profits, we may look at differences in household income across experimental arms. This variable captures not only the returns to maize cultivation, but also accounts for the extra labor that went into it (because off-farm income will be lower) or the fact that the best plot was allocated to maize (which lowers the productivity of the next best crop, now grown on another field).

If farmers adjust optimally to the treatment, then focusing on household income would produce a measure of benefit that is *greater*, in relative terms, than the estimated gain in a double-blind experiment, but *smaller* than the measure of benefit eventuating in a standard RCT focusing on maize harvests (profits). However, household income depends on many factors and varies for many different reasons – mostly unrelated to the experiment. The signal-to-noise ratio is therefore low, and (much) larger samples are necessary to reduce the risk of being under-powered. This further adds to the costs of experimenting in the field.

References

Barrett, C.B. and M.R. Carter. 2010. The Power and Pitfalls of Experiments in Development Economics: Some Non-Random Reflections. *Applied Economic Perspectives and Policy* 32:515-548

Beaman, L., D. Karlan, B. Thuysbaert & C. Udry (2013). Profitability of fertilizer: Experimental evidence from female rice farmers in Mali. *American Economic Review* (P&P) 103(3), 381-386

Bulte, E., G. Beekman, S. Di Falco, J. Hella, and L. Pan (2014). Behavioral responses and the impact of new agricultural technologies: Evidence from a double-blind field experiment in Tanzania. *American Journal of Agricultural Economics* 96, 813-830

Bulte E., S. Di Falco, M. Kassie, and X. Vollenweider. 2019. Counterfeit Seeds, Labor Supply and Economic Returns: Experimental Evidence from Tanzania, mimeo

Butera, L., and J. A. List. 2017. *An economic approach to alleviate the crises of confidence in science: With an application to the public goods game*. NBER No. w23335. *National Bureau of Economic Research*

Chassang, S., E. Snowberg, B. Seymour & C. Bowles (2015). Accounting for behavior in treatment effects: New applications for blind trials. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0127227

Deacon, A. 2019. Randomization in the tropics revisited: a theme and eleven variations. Woodrow Wilson School, Princeton University Schaeffer Center for Health Policy and Economics, University of Southern California National Bureau of Economic Research