World Development 127 (2020) 104824

Contents lists available at ScienceDirect

World Development

journal homepage: www.elsevier.com/locate/worlddev

Commentary

Highly prized experiments

Martin Ravallion [1]

Department of Economics, Georgetown University, Washington DC 20057, USA

ARTICLE INFO

Article history:
Accepted 1 December 2019
Available online 14 December 2019

JEL:
B23
H43
O22

Keywords:
Randomized controlled trials
Bias
Sample selection

ABSTRACT

The new Nobel prize winners have expertly popularized randomized controlled trials (RCTs) as the "tool-of-choice" for empirical research. The award is a good opportunity to reflect on the role of RCTs in development-policy evaluation. Unbiasedness is the tool's main virtue; transparency is another. Practitioners should also be aware of some limitations. First, an RCT assigns the treatment in a different way to most real-world policies, which use purposive selection; given heterogeneous impacts, one is evaluating a different intervention. Second, the tool may only be feasible for non-random subsets of both the relevant populations and the policy options, biasing assessments of overall development effectiveness. Third, given budget-constraints and a bias-variance trade-off, a non-RCT may allow a larger sample size, making its trials often closer to the truth. There is a continuing need for a broad range of research methods for addressing pressing knowledge gaps in fighting poverty.

The three winners of the 2019 Sveriges Riksbank Prize in Economic Sciences (in Memory of Alfred Nobel)—Abhijit Banerjee, Esther Duflo and Michael Kremer—have contributed greatly to development economics, both directly and by attracting others to the field. There have been many contributions, but the prize winners are best known for their randomized experiments, also called randomized controlled trials (RCTs). As said in the headline of the announcement, the prize was awarded "for their experimental approach to alleviating poverty."

The use of RCTs in social-policy evaluation goes back to the 1960s, with applications to development policy starting in the 1970s. The new Millennium saw a major expansion in their use, notably with the creation of the *Abdul Latif Jameel Poverty Action Lab* (J-PAL), founded by two of the prize winners (Banerjee and Duflo, along with Sendhil Mullainathan). At the time of writing, J-PAL has almost 1000 completed and ongoing RCTs in over 80 countries. Countless other experiments have been inspired by them. Many interesting, and increasingly nuanced, lessons have emerged (as reviewed in Banerjee, Duflo, & Kremer, 2019). This has come with a marked impact on research methods. One often hears now that an RCT is the "gold standard," and even uniquely placed as the rigorous method of impact evaluation.

This short article cannot do justice to either the contributions of the experimental approach or the debates it has generated.[2] I focus on what many economists and others have seen as the main virtue of an RCT, namely that, under ideal conditions, it delivers an unbiased estimate of a specific parameter, namely the mean impact of an assigned intervention across a population. "Unbiased" means that, with repeated trials, we get estimates that tend to get closer on average to the true value of that parameter. There will be an experimental error in any one trial, but the mean error will eventually go to zero. (This also delivers an estimable variance, to facilitate calculation of the confidence interval.) There are other merits of RCTs (such as their transparency), but unbiasedness would clearly be at the top of the list for advocates.

The prize winners responded well to the preference for unbiasedness among economists, and the limitations of some of the arguments made for identifying causal effects without randomization. RCT's have largely ended all those econ-seminar debates and referees' comments about latent confounders clouding causal inferences about mean impacts.

However, practitioners should be aware of the limitations of prioritizing unbiasedness, with RCTs as the *a priori* tool-of-choice. This is not to question the contributions of the Nobel prize winners. Rather it is a plea for assuring that the "tool-of-choice" should

[1] The author thanks Francois Bourguignon, Dan Cao, Denis Cogneau, Sylvie Lambert, Rachael Meager, Franco Peracchi, John Rust, Daniel Valderrama-Gonzalez and Dominique van de Walle for their comments.

E-mail address: mr1185@georgetown.edu

[2] An important "pre-J-PAL" discussion is Heckman and Smith (1995). More recently, see Deaton and Cartwright (2018) (and the accompanying comments) and Ravallion (2020), which provides further references to the literature.

https://doi.org/10.1016/j.worlddev.2019.104824
0305-750X/

always be the best method for addressing our most pressing knowledge gaps in fighting poverty.

Imagine that there exists both an RCT and a non-randomized (observational) design for a pilot to assess the likely impact of an antipoverty program at scale. We can safely assume that the program has latent heterogeneous impacts. For the non-RCT, people select into the program, or are selected, and we take random samples of those who do and those that do not. Such a study need not be biased, and many things can go wrong with RCTs in practice, creating bias. However, for the sake of the argument, let us suppose that the proposed RCT will be implemented well.

Can we rank the rival methods *ex ante* according to how close their trial estimates are likely to be to the (unknown) true mean impact of the program when operating at scale? An estimate can be said to be "close to the truth" if it is within an interval of width δ centered on the true mean impact.

There are reasons to question any presumption that the RCT option should be preferred. For one thing, an RCT is a rather artificial construction, unlike almost any imaginable real-world policy. The program at scale is unlikely to be randomly assigned; it is hard to imagine any government forgoing its power to set eligibility criteria at scale in favor of randomly withholding treatment, or randomly assigning it whether it is needed or not. Given latent impact-heterogeneity, the mean impact from the RCT may be very different to a purposive assignment at scale in which those who benefit more tend to select into the program (Heckman & Smith, 1995). By contrast, observational studies need not change the program for the purpose of learning. Even a biased non-RCT may then be more revealing about impact at scale.

Another concern stems from the fact that RCTs are often easier to do with a non-governmental organization (NGO). Academic "randomistas," looking for local partners, appreciate the attractions of working with a compliant NGO rather than a politically sensitive and demanding government. Thus, the RCT is confined to what NGO's can do, which is only a subset of what matters to development. Also, the desire to randomize may only allow an unbiased impact estimate for a non-randomly-selected sub-population—the catchment area of the NGO. And the selection process for that sub-sample may be far from clear. Often we do not even know what "universe" is represented by the RCT sample. Again, with heterogeneous impacts, the biased non-RCT may be closer to the truth for the whole population than the RCT, which is (at best) only unbiased for the NGO's catchment area.

Here it is notable that the prize winners have made progress in persuading governments to do RCTs. That helps alleviate my concerns. But it raises another issue: that RCTs are only feasible for a subset of what governments do in the interests of development—and it is a selected (non-random) subset, namely relatively small, short-term, assignable interventions, such as cash transfers. The marked shift toward RCTs in evaluations at the World Bank has come with a seemingly poor fit of the evaluation portfolio to the Bank's work program (World Bank, 2012). Insistence on doing RCTs creates such knowledge gaps.

If we are really concerned about obtaining unbiased estimates of the impact of the portfolio of development policies it would be better to carefully choose a representative sample of those policies for evaluation, and then find the best method for each of the selected programs, with an RCT as only one of many possible options.

Nor is bias all that matters. Drawing only one randomly-chosen observation of the outcomes for treated and controlled units will give an unbiased estimate of impact, but it would almost never be considered reliable. Recognizing this, a popular decision rule in statistics is to minimize the mean-squared error (MSE), i.e., the expected value of the squared deviation between the estimate and its true value. As is easily verified (and well-known in statistics), the MSE is the estimator's squared bias plus its variance. A larger sample size reduces the variance of an estimate. If removing all bias comes at the expense of a lower sample size then the MSE may be higher than for a biased non-RCT.

The economics of research design comes into play. Observational studies can sometimes draw on data from administrative records (including "big-data") and existing surveys. RCTs typically require new special-purpose surveys, which can be costly to do well. Even when the observational study requires a new survey, it avoids the extra costs generated by an RCT. Thus, for a given budget, insisting on an RCT can lead to lower sample sizes and (hence) higher variances. Indeed, one often hears complaints about underpowered RCTs (see, for example, White, 2014) and sampling variability has been found to account for a large share of the variance in impact estimates for RCTs (Meager, 2019, with reference to microcredit schemes). Cost comparisons in World Bank (2012) suggest higher costs for RCTs.

The implication is illustrated by a numerical example (using computer-generated data) found in Ravallion (2020). Each trial was drawn from one of two normal distributions, one for an RCT and one for a non-RCT. The mean of the RCT distribution of trial results was taken to be the true mean, while it was not for the non-RCT. With (say) twice the sample size, the variance of the estimates from the (biased) non-RCT is still low enough to assure that the method yields a lower MSE and higher share of its trials that are close to the truth than the RCT. Indeed, the non-RCT was more often closer to the truth for all δ!

Of course, this is only one example. More research is needed on the performance of alternative methods—including on the distribution of the biases in observational studies. All this example illustrates is that we should not presume that the RCT option delivers more reliable impact estimates.

The task of evidence-based policy making against poverty will require openness to multiple options for measuring and describing patterns in the data, and for identifying and understanding causal impacts. No one method should dominate.

References

Banerjee, A., Duflo, E., & Kremer, M. (2019). The influence of randomized controlled trials on development economics research and on development policy. In K. Basu, D. Rosenblatt, & C. Sepúlveda (Eds.), *The State of Economics, The State of the World*. Cambridge Mass: MIT Press.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine, 210*, 2–21.

Heckman, J., & Smith, J. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives, 9*(2), 85–110.

Meager, R. (2019). Understanding the average impact of microcredit expansion: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics, 11*(1), 57–91.

Ravallion, M. (2020). Should the randomistas (continue to) rule? In F. Bédécarrats, I. Guérin, & F. Roubaud (Eds.), *Randomized Control Trials in the Field of Development: A Critical Perspective*. Oxford: Oxford University Press.

White, H. (2014). *Ten things that can go wrong with randomised controlled trials*. Evidence Matters Blog, International Initiative for Impact Evaluation.

World Bank (2012). *World Bank Group Impact Evaluations: Relevance and Effectiveness*. World Bank: Independent Evaluation Group.