NBER WORKING PAPER SERIES

SHOULD THE RANDOMISTAS (CONTINUE TO) RULE?

Martin Ravallion

Working Paper 27554
http://www.nber.org/papers/w27554

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2020

François Roubaud encouraged the author to write this paper. The author thanks Jason Abaluck, Sarah Baird, Radu Ban, Mary Ann Bronson, Caitlin Brown, Sylvain Chabé-Ferret, Kevin Donovan, Ryan Edwards, Markus Goldstein, Miguel Hernan, Emmanuel Jimenez, Max Kasy, Madhulika Khanna, Nishtha Kochhar, Agnès Labrousse, Andrew Leigh, David McKenzie, Rachael Meager, Berk Özler, Dina Pomeranz, Lant Pritchett, Milan Thomas, Vinod Thomas, Eva Vivalt, Dominique van de Walle, Andrew Zeitlin, and participants at a workshop in Paris, March 2019. The staff of the International Initiative for Impact Evaluation kindly provided an update to their database on published impact evaluations and helped with the author's questions. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Should the Randomistas (Continue to) Rule?
Martin Ravallion
NBER Working Paper No. 27554
July 2020
JEL No. B41,C93,O22

ABSTRACT

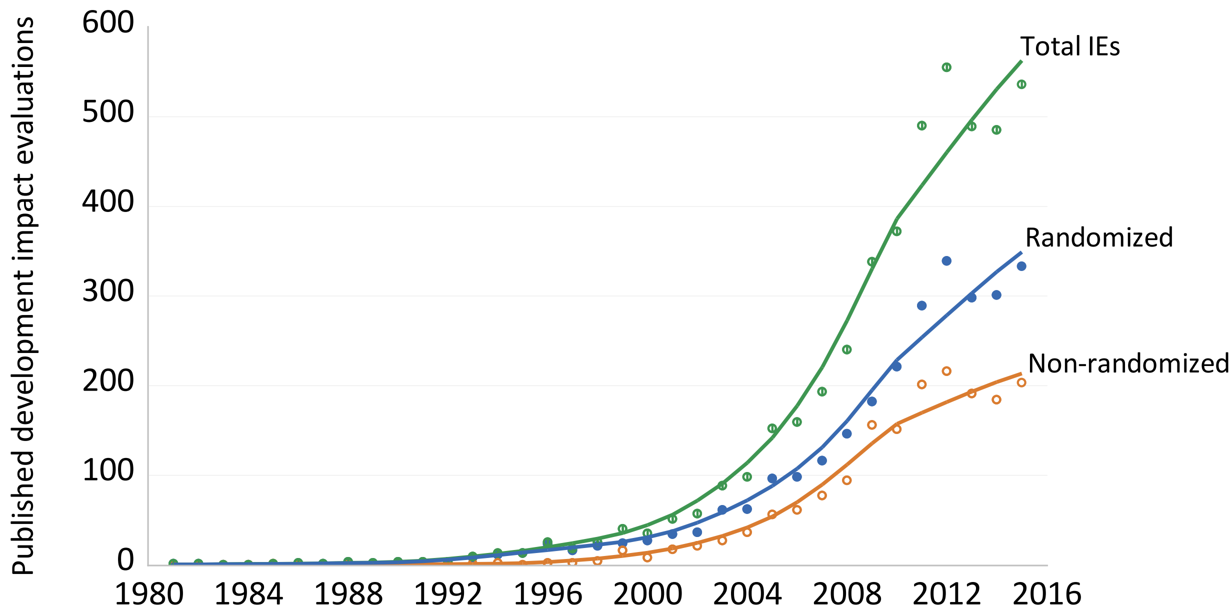
The rising popularity of randomized controlled trials (RCTs) in development applications has come with continuing debates about the merits of this approach. The paper takes stock of the issues. It argues that an unconditional preference for RCTs is questionable on three main counts. First, the case for such a preference is unclear on a priori grounds. For example, with a given budget, even a biased observational study can come closer to the truth than a costly RCT. Second, the ethical objections to RCTs have not been properly addressed by advocates. Third, there is a risk of distorting the evidence-base for informing policymaking, given that an insistence on RCTs generates selection bias in what gets evaluated. Going forward, pressing knowledge gaps should drive the questions asked and how they are answered, not the methodological preferences of some researchers. The gold standard is the best method for the question at hand.

Martin Ravallion
Department of Economics
Georgetown University
ICC 580
Washington, DC 20057
and NBER
mr1185@georgetown.edu

1. Introduction

The new millennium has seen a huge increase in the application of impact evaluations to developing countries, typically with the aim of improving policy-making. The International Initiative for Impact Evaluation (3ie) has compiled metadata on such evaluations, as reproduced in Figure 1.[2] We see a remarkable 30-fold increase in the annual production of published IEs since 2000, compared to the 19 years prior to 2000.[3]

Figure 1: Annual counts of published impact evaluations for developing countries



Note: Fitted lines are nearest neighbor smoothed scatter plots. See footnote 4 in the main text on likely under-counting of non-randomized evaluations in earlier years. Source of primary data: International Initiative for Impact Evaluation.

There are two broad groups of methods, as also identified in Figure 1.[4] In the first, access to a program (the "treatment") is randomly assigned to some units, with others randomly set aside as controls. To measure the program's impact one then compares mean outcomes for these

---

[2] See Cameron et al. (2016) and Sabet and Brown (2018). The numbers here span 1981-2015.
[3] 4,501 impact evaluations are recorded in the 3ie database, covering the period 1981-2015, of which 4,338 were published in 2000-15. The annual rates are 271 since 2000 and 9 for 1981-1999.
[4] The 3ie series is constructed by searching for selected keywords in digitized texts. 3ie staff warned me (in correspondence) that their old search protocols were probably less effective in picking up observational studies relative to RCTs prior to 2000. So the earlier, lower, counts of non-randomized evaluations in Figure 1 may be deceptive. The 3ie counts pick up many more RCTs than reported in Bouguen et al. (2019) (also noting that the latter give cumulative totals not annual flows).

two samples. This is the simplest version of a randomized controlled trial (RCT). A second group of methods does not use randomization. These include purely "observational studies" in which the assignment of treatment is taken as data, and understood to be purposive rather than random. The second group also includes deterministic assignments, such as based on priors about likely benefits from the treatment. While some non-RCTs that help inform policy making are purely descriptive, others attempt to control for the pre-treatment differences between treated and un-treated units based on what can be observed in data, with the aim of drawing credible causal inferences about impact.

While the use of RCTs in development applications began around 1980, a rapid expansion in their use emerged some 20 years later. About 60% of the impact evaluations since 2000 have used randomization. The latest 3ie count has 333 papers using this tool for 2015.[5] The growth rate is striking. Fitting an exponential trend (and the fit is good) to the counts of RCTs in Figure 1 yields an annual growth rate of around 20%—more than double the growth rate for all scientific publishing post-WW2.[6] As a further indication, if one enters "RCT" or "randomized controlled trial" in the *Google Ngram Viewer* one finds that the incidence of these words (as a share of all ngrams in digitized texts) has tended to rise over time and is higher at the end of the available time series (2008) than ever before.

The fact that so much of the growth evident in Figure 1 has been for RCTs would surely not have been anticipated prior to 2000. After all, RCTs are not feasible for many of the things that governments and others do in the name of development. Nor had RCTs been historically popular, given the often-heard concerns about withholding a program from some people who need it, while providing it to some who do not, for the purpose of research. Development RCTs used to be a hard sell. Something changed. How did RCTs become so popular? And is their popularity justified?

Advocates of RCTs have been dubbed the "randomistas."[7] They proffer RCTs as the "gold standard" for impact evaluation—the most "scientific" or "rigorous" approach, promising

---

[5] To put this in perspective for economists, this is about the same as the total number of papers (in all fields) published per year in the *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*, *Econometrica* and the *Review of Economic Studies* (Card and DellaVigna, 2013).

[6] Regressing the log RCT count (dropping three zeros in the 1980s) on time gives a coefficient of 0.20 (s.e.=0.01; n=32; R^2=0.96) or 0.18 (0.01; n=16; R^2=0.96) if one only uses the series from 2000 onwards. In modern times (post-WW2), the growth rate of scientific publications is estimated to be 8-9% per annum (Bornmann and Mutz, 2015).

[7] That term "randomistas" is not pejorative; indeed, RCT advocates also use it approvingly, such as Leigh (2018).

to deliver largely atheoretical and assumption-free, yet reliable, IE.[8] This view has come from prominent academic economists, and it has permeated the popular discourse, with discernable influence in the media, development agencies and donors, as well as among researchers and their employers.[9] This is an unconditional preference for RCTs. While there are a great many contexts for an impact evaluation(types of interventions, sectors of the economy, countries, communities, social/ethnic groups), the gold-standard claim is typically made independently of context.

There have been pushbacks. RCTs in social-policy applications have raised many concerns.[10] Critics have argued that (*inter alia*): the assumptions required for a reliable impact estimate using an RCT need not hold in reality; RCTs are ethically questionable; and the "black box" nature of RCTs limits their usefulness for policymaking, including both scaling up and learning about likely impact in other contexts. There have been defenses against the critics.[11] And passions have run high at times, with one commentator dismissing ethical criticisms of RCTs as "garbage" (Fiennes, 2018) while one critic has called the RCT revolution "madness" (indeed, "much worse than madness" at one point) (Pritchett, 2020, p.26).

In the light of the rising prominence of development RCTs, and the continuing debates, this paper returns, 10 years later, to the question posed in Ravallion (2009a), "Should the randomistas rule?" The sense in which randomistas "rule" is in their claimed hierarchy of methods, which is the foundation of their intellectual authority and power to persuade.[12] That hierarchy is the main focus of the paper. While recognizing the attractions of RCTs for some purposes, the paper argues that the supportive public narrative on RCTs that has emerged is not

[8] For example, Banerjee (2006) writes that: "Randomized trials like these—that is, trials in which the intervention is assigned randomly—are the simplest and best way of assessing the impact of a program." Similarly, Imbens (2010, p.407) claims that "Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top." And Duflo (2017, p.3) refers to RCTs the "tool of choice."

[9] An example of the broader influence of the "gold standard" view is the Wikipedia entry on IE, which states that "Randomized field experiments are the strongest research designs for assessing program impact… as it allows for a fair and accurate estimate of the program's actual effects." In another example, Keating (2014) writes that "Randomistas, proponents of randomized controlled trials, have recently been transforming the way we think about economic development and aid to poor countries." Similarly, Leigh's (2018) volume is entitled "Randomistas: How Radical Researchers Changed our World."

[10] See Heckman and Smith (1995), Grossman and Mackenzie (2005), Cartwright (2007), Ravallion (2009a,b; 2012), Rodrik (2009), Barrett and Carter (2010), Deaton (2010), Keane (2010), Baele (2013), Basu (2014), Mulligan (2014), Pritchett and Sandefur (2015), Favereau (2016), Ziliak and Teather-Posadas (2016), Hammer (2017), Deaton and Cartwright (2018), Gibson (2019), Pritchett (2020) and Young (2018).

[11] Including Banerjee and Duflo (2009), Goldberg (2014), Imbens (2010, 2018), Glennerster and Powers (2016) and McKenzie (2019).

[12] Thus, McKenzie's (2019) observation that only 10% of all papers in development economics (any field, in 14 journals) are RCTs does not refute the claim that the randomistas do indeed rule in the sense used here.

well grounded in an appreciation of the limitations of this research tool. The paper's intended audience is not the experts on either side, but the broader community of economists and other social scientists, donors, policymakers and their advisors, students and young researchers.

The paper begins with an overview of the theory of impact evaluation, as relevant to the choice of methods (Section 2.2). It then discusses the randomistas' influence on development research (Section 2.3), the concerns about the ethical validity of their preferred method (Section 2.4), and the relevance of their research to policy (Section 2.5). Section 2.6 concludes.

2. Foundations of impact evaluation

The focus is on assigned programs, in that some units (the "treated") in a well-defined population get the program and some do not. Imagine drawing two random samples from the population, one from those treated and one from those not, and then measuring relevant outcomes for both. This constitutes a single experimental trial.[13] The difference in mean outcomes is the trial's estimate of the true mean impact for that population, also called the average treatment effect (ATE). That estimate can differ from the true value due to measurement errors, sampling variability, spillover effects ("contamination") between the two groups, monitoring effects, and/or systematic bias arising from any confounding variables that jointly alter outcomes and treatment status. Each trial's sampled pair gives a different estimate, sometimes too high, sometimes too low, though we never know by how much since we do not (of course) know the true value. Every trial has some error.

The ideal RCT is the special case of the above setup in which the trial's treatment status is also assigned randomly (in addition to drawing random samples from the two populations, one treated and one not) and the only error is due to sampling variability. This ideal can be illusive in practice, especially with human subjects; the discussion will return to the real-world departures from the ideal, but for now an ideal RCT is assumed. In this special case, as the sample sizes increase, the trial estimates gets closer to the true mean impact. This is the sense in which an ideal RCT is said to be unbiased, namely that the sampling error is driven to zero in expectation. This property also allows us to estimate the variance of the estimates. Thus, using both random

[13] The word "experiment" is sometimes defined as any situation in which the evaluator controls everything, and this is deemed to be the case for an RCT; see, for example, Cox and Reid (2000). However, it is almost never the case that the evaluator controls everything in RCTs with human subjects, as used to evaluate social policies. Here I use the broader definition of "experiment," not assuming full control. It may or may not be an RCT.

treatment and random sampling facilitates calculation of the standard error of the impact estimate from an RCT, to establish a statistical confidence interval.[14]

Prominent randomistas have sometimes left out the "in expectation" qualifier, or ignored its implications for the existence of experimental errors (as noted by Deaton and Cartwright, 2018). These advocates of RCTs attribute any difference in mean outcomes between the treatment and control samples to the intervention.[15] This common mistake might be thought of as little more than a minor expository simplification.[16] However, the simplification is now embedded in much of the public narrative. Beyond the experts (putting aside their unguarded statements), many people in the development community now think that any measured difference between the treatment and control groups in an RCT is attributable to the treatment. It is not; even the ideal RCT has some unknown experimental error.

A rare but instructive case is when there is no treatment. Absent any other effects of assignment (such as from monitoring), the impact is zero. Yet the random error in one trial can still yield a non-zero mean impact from an RCT. An example is an RCT in Denmark in which 860 elderly people were randomly and unknowingly divided into treatment and control groups prior to an 18-month period without any actual intervention (Vass, 2010). A statistically significant (prob.=0.003) difference in mortality rates emerged at the end of the period.

In the light of these observations, consider the choice of methods. Suppose that, with a given budget, we can implement either an RCT or an observational study. For the latter, people select into the program, and we take random samples of those who do and those that do not. We want to rank the methods *ex ante* according to how close their trial estimates are likely to be to the true value. Let us say that an estimate is "close to the truth" if it is within some fixed interval

---

[14] Current practices in this respect can be questioned. Young (2018) points to a number of concerns in past impact estimates of standard errors when using RCTs with regression controls and shows that many published economics papers have over-estimated the statistical significance of their impact estimates. Also see the discussions in Deaton and Cartwright (2018) and Imbens (2018).

[15] For example, with reference to RCTs, Banerjee and Duflo (2017) write that "any difference between the treatment group and the comparison group can be confidently attributed to the treatment," and Banerjee et al. (2019) claim that an RCT ensures that "any difference between treatment and control units reflects the impact of the treatment." One finds a similarly unguarded claim in the "Introduction to Evaluation" on the website of the *Abdul Latif Jameel Poverty Action Lab* (J-PAL) (which Section 2.3 returns to); having described a stylized RCT for a water purification project, with treatment and control groups, J-PAL says that: "any differences seen later on can be attributed to one having been given the water purification program, and the other not." Another example is found in a technical manual on impact evaluation by the Inter-American Development Bank and the World Bank (Gertler et al. 2016).

[16] As Imbens (2018) suggests, in his comments on Deaton and Cartwright (2018).

centered on the true value. (The focus here is on the "internal validity" of each estimator—its accuracy for the population in hand; Section 2.5 turns to "external validity.")

The reason one hears most often for the "gold-standard" ranking is the unbiasedness of an ideal RCT. Economists have focused a lot on one particular source of bias, namely any difference between the mathematical expectation of a parameter estimate and its (unknown) true value. (In some of the literature this is called "systematic bias," as distinct from the, potentially many, sources of trial-specific errors.[17]) Even by this narrow definition, an observational study need not be biased. One typically adjusts for covariate imbalance, including in an RCT. Bias is removed when the treatment status is conditionally exogenous, i.e., uncorrelated with the error term conditional on the covariates (though this is clearly a stronger assumption than for an RCT). That assumption may or may not be acceptable, depending on the context (the program and the data available). Whether or not the treatment is exogeneous given the control variables depends on whether those variables adequately reflect the determinants of treatment placement; that must be judged in each setting. A good understanding of the economic and social determinants of program placement—the decision problems facing the various stakeholders in the specific context—can help in determining what data one needs. Omitted confounders will often remain, although that need not mean large biases on adjusting for the observed confounders.

If unmeasured confounders are a serious concern then the bias can be removed if one can find a source of exogenous variation in treatment status that is not also a determinant of outcomes given treatment. This is an instrumental variable (IV). A valid IV must be correlated with treatment status <u>and</u> uncorrelated with outcomes, given treatment and the control variables. In a regression, this requires that the IV is uncorrelated with the error term—giving what is called the "exclusion restriction." This condition must ultimately be judged on theoretical grounds, though close study of the factors determining treatment status in the specific setting can be valuable in finding theoretically plausible IVs, as well as potential confounders. For example, consider a program for which the assignment to treatment depends on whether an eligibility score is above some critical threshold, along with other factors adding fuzziness to the assignment. As long as the threshold is arbitrary (namely that mean counterfactual outcomes do not change at the threshold), whether the score is above or below this critical value is a

[17] On the multiple sources of "bias" see Hernan and Robins (2018).

theoretically defensible IV.[18] Though less familiar to economists, bias in an observational study due to unmeasured confounders can also be eliminated if there is an intermediate variable that links treatment to outcomes but does not depend on the confounders.[19]

Even if we agree that an RCT is better at removing bias in a specific setting, that does not clinch the ranking. There are two main reasons. First, given the constraints faced on RCTs in practice, it may not be feasible to properly represent the population of interest. At least when there is a free media, governments are likely to see a political risk in supporting ethically-questionable research. While RCTs are sometimes done with governments, more benign observational studies are often easier to accept. Thus, academic randomistas looking for local partners see the attractions of working instead with local non-governmental organizations (NGOs). The desire to randomize may thus deliver (under ideal conditions) an unbiased impact estimate for a non-randomly selected sub-population, such as those in the catchment area of a cooperative local NGO. Furthermore, the selection process for the compliant sub-sample may be far from clear (indeed, without even a mention in the paper on how it was chosen). It is unclear what can be learnt from an unbiased estimate for a non-randomly selected subsample of the population of interest. Given the likely heterogeneity in impacts, the biased observational study for a random sample from the whole population may be closer to the truth.

Second, bias is not the only thing that matters. The appropriate decision rule for choosing an estimator (and designing a research study more generally) depends on the application. A popular statistical decision rule 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 well-known in statistics, the MSE is the estimator's squared bias plus its variance.[20] Thus, this decision rule does not tell us that an unbiased estimator is always best.[21] MSE is not the only defensible decision rule—for example, one might ask how often the trials are within some absolute distance of the true value—but the point here is that unbiasedness is not all that matters.

---

[18] This is an example of regression-discontinuity design; for a formal treatment see Hahn et al. (2001).

[19] This is sometimes called "front-door adjustment" as distinct from "back-door adjustment" using an IV (Pearl and Mackenzie, 2018, Chapters 4 and 7). An example of front-door adjustment can be found in Glynn and Kashin (2018). For a more formal treatment see Pearl (2009, Chapter 3).

[20] If β is the true value and $\hat{\beta}$ its estimator then (by definition) MSE $\equiv E\left(\left(\hat{\beta}-\beta\right)^2\right) = \left(E(\hat{\beta})-\beta\right)^2 + E\left(\left(\hat{\beta}-E(\hat{\beta})\right)^2\right)$. The first term on the RHS is the squared bias of $\hat{\beta}$ while the second term is its variance.

[21] This is well recognized in introductory econometrics texts. For example, Johnston (1984, p.28) writes that "…on the mean-squared-error criterion a biased estimator may be preferred to one with smaller or zero bias if its variance is sufficiently small to offset the larger bias." Also see the discussion in Greene (1991, pp.97-99).

The economics of impact evaluation comes into play here. Larger sample sizes reduce the variance of estimates. Many observational studies use existing data, from administrative records ("big-data"), as well as existing surveys. RCTs typically require new special-purpose surveys. Thus, for a given budget, RCTs will often have lower sample sizes and (hence) higher variances.

Nor is the outcome clear when a non-RCT requires new surveys. A good way to reduce bias is with better data. Longer survey questionnaires will probably entail smaller sample sizes for a given budget. But the data requirements for an RCT are unlikely to be different, noting that one wants baseline data to test for covariate balance in an RCT.[22] The additional randomization (for treatment) in an RCT is unlikely to be costless, and re-randomization may well be needed to assure covariate balance (Morgan and Rubin, 2012). In medical applications, RCTs are widely thought to be more costly than observational studies.[23] I have not seen systematic cost data for development impact evaluations, though one often hears concerns about underpowered RCTs.[24] Cost comparisons for evaluations at the World Bank suggest higher costs for RCTs (though the comparisons are crude).[25] Development RCTs often encounter difficulties of implementation in the field that are not found for observational studies.

To see what this can mean for the choice of method, suppose that each trial is drawn from one of two normal distributions, one for an RCT and one for a non-RCT. The parameters (its mean and variance) depend on the chosen method. The mean of the RCT distribution of trial results is taken to be the true mean, while it is not for the non-RCT. Even so, despite the bias, the variance of a non-RCT could be low enough to assure that it yields a higher share of its trials that are close to the truth than the RCT. Figure 2 illustrates a hypothetical case, showing that even a biased observational study can be closer to the truth than an (unbiased) RCT.[26] Two densities are shown for impact estimates from both RCT and non-RCT designs, both drawn from normal distributions. (The densities may or may not be conditional on covariates.) The true impact is

[22] *Ex post* balancing tests and retrospective adjustments are often recommended for RCTs (Cox and Reid, 2000; Hinkelmann and Kempthorne, 2008; Bruhn and McKenzie, 2009; Hernán and Robins, 2018).

[23] See, for example, Hannan (2008) and Frieden (2017).

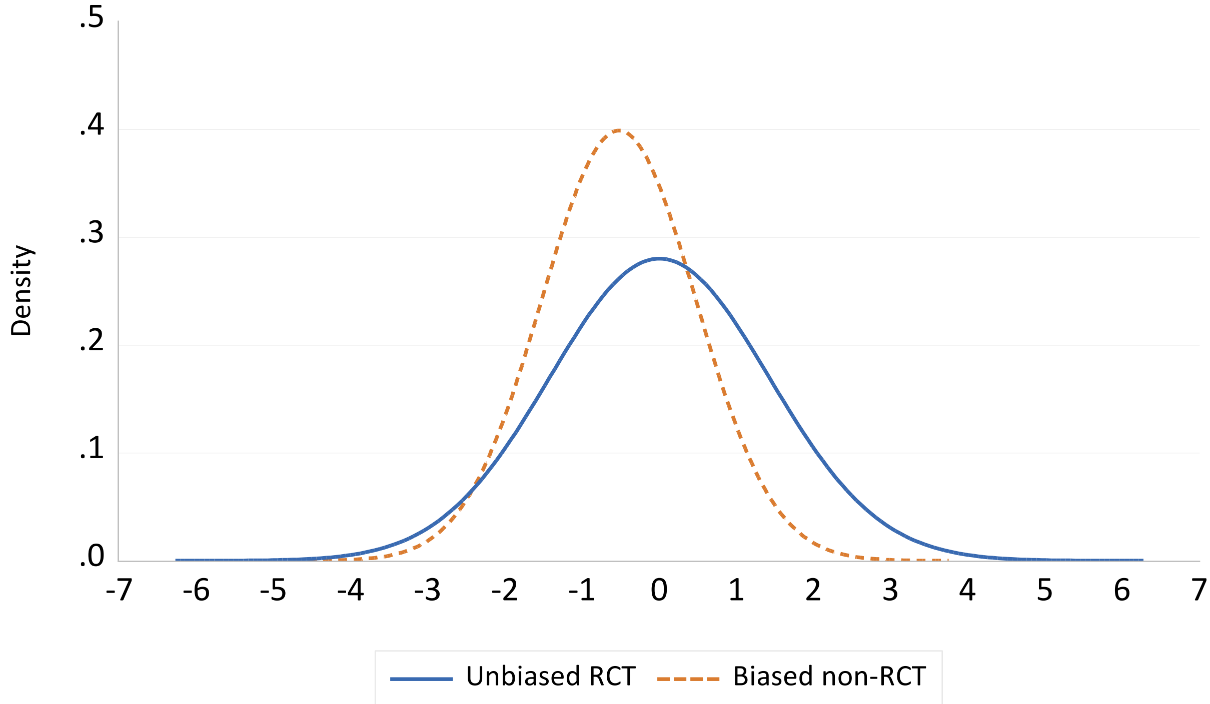
[24] For example, in reference to development RCTs, White (2014) says that "..the actual power of many RCTs is only around 50 per cent. So, an RCT is no better than tossing a coin for correctly finding out if an intervention works." Sampling variability appears to account for half or more of the variability in impact estimates from RCTs; see Meager (2018), with reference to microcredit schemes.

[25] The World Bank's impact evaluations in recent times have tended to be RCTs with considerably higher average cost than the evaluations done in the International Financial Cooperation (within the World Bank Group), where observational studies are more common (World Bank, 2012). This is at best suggestive since the comparison is not properly controlled.

[26] Greene (1991, Section 4 .3) uses a similar example to show that a more biased estimator can have lower MSE.

zero, which is the mean of the distribution from which the RCT trials are drawn. The non-RCT trials are drawn instead from a distribution with a mean of -0.5, which is their systematic bias. The other difference is that the RCT trials are drawn from a distribution with a variance of 2, while for the observational study it is 1. This can be interpreted as saying that, for a given budget, the non-RCT allows double the sample size in each trial.

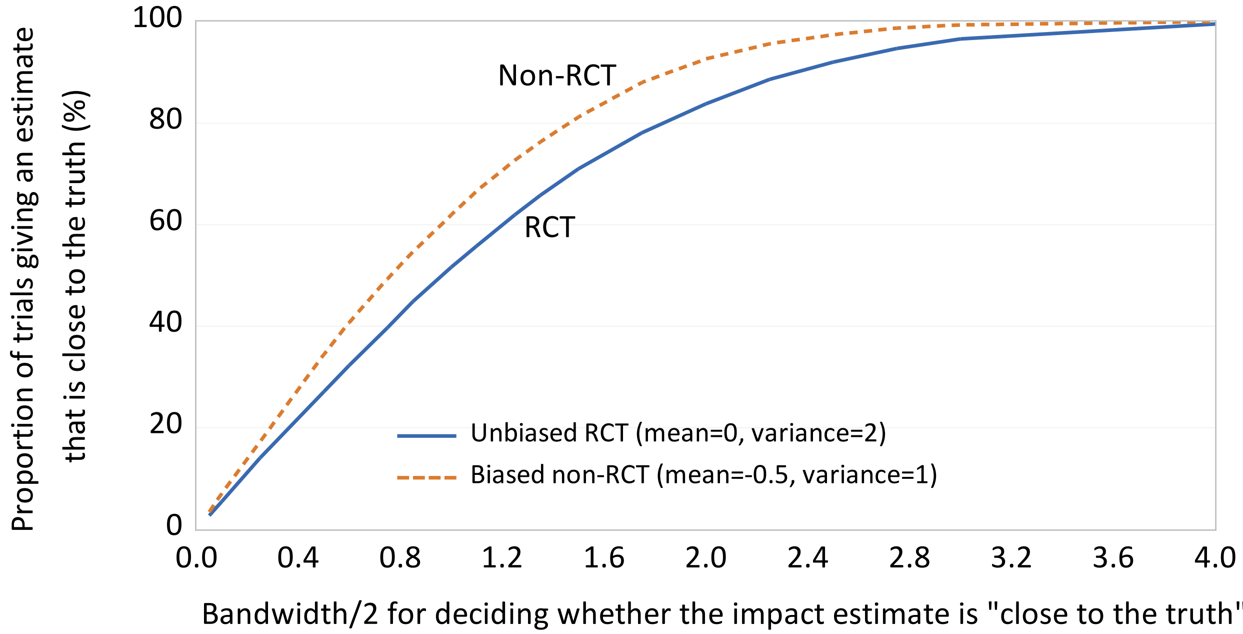
Figure 2: Density functions for the estimates of mean impact from two hypothetical designs for impact evaluations



Which method does better, in that its trial estimates tend to be closer to the truth? Define "closer to the truth" as being more likely to be within a fixed interval centered on the true value—in this case, an interval $(-\delta, \delta)$ (for some $\delta > 0$).[27] Figure 3 gives the percentage of trials close to the truth for each method. Suppose we define "closer to the truth" as an impact estimate in the interval (-0.5, 0.5). We find that the RCT gets an estimate that is within this interval for 27% of its trials, but this is so for 34% of the non-RCT trials. If instead we define "closer to the truth" as an estimate in the interval (-1, 1) then this is so for 52% of RCT trials versus 62% using the observational study. In this example, the observational study is closer to the truth for all δ!

[27] In some applications the interval need not be symmetric around the true value, i.e., errors in one direction are more costly to the agreed objective.

Figure 3: Proportion of trials giving an impact estimate that is close to the truth, comparing an unbiased RCT with a biased non-RCT on a larger sample



Of course, this is only one of many possibilities, and one can readily construct examples where the RCT does better. Figures 2 and 3 only illustrate that the less biased impact evaluation need not get us closer to the truth. That is an open question as it depends crucially on the power of the trials that can be afforded given the budget. The key point is that we cannot rule out the possibility that, for a given budget, the RCT ends up with less reliable impact estimates, that are often further from the truth than even a biased observational study. We lack a theoretical justification for the claimed (unconditional) “gold standard” hierarchy of methods.

In defense of RCTs, we currently know rather little about the distribution of the biases in observational studies, while (as noted) the unbiasedness of an RCT comes with an estimable variance, to facilitate calculation of its confidence interval. This points to the need for more research on the distribution of estimates from observational studies, such as by comparing estimates with those from RCTs for the same setting.[28] It remains, however, that if we insist on only doing RCTs (when feasible) then we may be forgoing observational studies that are more often close to the truth.

[28] See, for example, Chabé-Ferret (2018).

Greater clarity may well emerge when we know the context. If one knows the setting and program well enough to identify the relevant confounders—the model of how the program works—and can collect data on them, or find measurable deconfounders, then one may well obtain a very reliable impact estimate by observation alone. On the other hand, if there is little scope for collecting baseline data on the relevant confounders, and the unit cost of randomized assignment is not too high (so that reasonably large sample sizes are feasible with the available budget), then an RCT has much appeal. It is the widely-heard "gold-standard" generalization that is at issue here.

We can go further and ask what design is optimal in the sense of minimizing (say) the MSE, while recognizing our uncertainty about the true model.[29] Let us assume that at least some of the baseline data are continuous covariates and that we have Bayesian priors on the model uncertainty. Then we can appeal to a result in Kasy (2016), namely that there exists a deterministic (non-random) assignment of treatment status based on the covariates that minimizes the expected MSE.[30] Then there is no gain from randomizing the assignment given the covariates (and there will be efficiency gains from taking account of observables in RCTs). By implication, to justify a strong preference for an RCT one needs to attach some intrinsic value to randomization as an end in itself, and be willing to forgo accuracy in getting closer to the true mean impact. Advocates have sometimes fallen back on such methodological preferences, independently of precision in estimating impact; for example, Banerjee et al. (2018) show that an RCT can still dominate as long as one puts a high enough weight on the welfare of those who prefer RCTs.

The influence of the randomistas has stemmed in part from the (much-heard) belief that RCTs are the preferred statistical tool when feasible. This review of the underlying theory has cast considerable doubt on that belief. As we will see, other sources of their influence are no less questionable.

---

[29] Following Kasy (2016) this can be recognized as a problem in statistical decision theory, i.e., the choice of an estimation method to minimize a loss function based on the data actually available.

[30] This holds for any Bayesian risk function and for a minimax rule for the worst-case (Kasy, 2016). Kasy provides software to implement the optimal assignment of treatment for minimizing the MSE. Note that a continuous covariate assures a unique optimum assignment of treatment status. With discrete covariates an RCT may do as well, but no better.

3. The influence of the randomistas on development research

Early examples of the use of RCTs in social policy contexts include the various experiments on US social policies starting in the 1960s.[31] With regard to development applications, the 3ie database has 133 published RCTs over the period 1981-1999. The earliest RCT in the database is from a World Bank research project on education interventions (textbooks and radio lessons) to improve the math scores of students in Nicaragua, namely Jamison et al. (1981).[32] Among the pre-2000 RCTs, that done by the Government of Mexico for the *Progresa* evaluation, which started in 1997, is an especially notable example. The (generally positive) results in the literature generated by the data from that RCT were influential in the expansion of Conditional Cash Transfers to over 50 countries today.[33]

So, at the beginning of the new Millennium, there was nothing new to the idea of RCTs in development applications. What changed is the popularity of that idea. The annual production of RCTs has been far higher since 2000 (Figure 1). Numerous individual academics and groups have contributed, but one group stands out, the *Abdul Latif Jameel Poverty Action Lab* (J-PAL).[34] This was founded in 2003 (as the *Poverty Action Lab*) and has been based in the Department of Economics at the Massachusetts Institute of Technology (MIT). The founders were Abhijit Banerjee, Esther Duflo and Sendhil Mullainathan. At the time of writing, J-PAL's website reports that they have over 1,000 completed and ongoing RCTs in 84 countries. For an academic research group to get that far in just 15 years is nothing short of amazing. On top of its own RCTs, J-PAL has clearly influenced the shift in emphasis in empirical development economics more broadly toward RCTs. Indeed, J-PAL's huge RCT output is unlikely to be even a majority of the total count of ongoing RCTs today.

The effort (centered on J-PAL but going further) came with a new prestige for this type of empirical research on development economics, as indicated by the award of the 2019 Sveriges

[31] On the history of RCTs in US social policy see the discussions in Burtless (1995) and List and Rasul (2011). Other commentaries on the history of RCTs more generally can be found in Ziliak (2014) and Leigh (2018).
[32] Banerjee et al. (2019) claim that the use of RCTs in development was "kick-started" in 1994 in a study by one of the authors (Kremer). However, published development RCTs had existed for at least 13 years prior to that.
[33] On the *Progresa* impact evaluation and its influence see Skoufias and Parker (2001) and Fiszbein and Schady (2010).
[34] Another prominent group doing and promoting RCTs is the non-profit organization, *Innovations for Poverty Action* (IPA), founded in 2002 by Dean Karlan (then at Yale). IPA and J-PAL often work together, and clearly have close links. Within international organizations, the most prominent group doing RCTs is the *Development Impact Evaluation* (DIME) group at the World Bank; three-quarters of DIME's evaluations have used this method (World Bank, 2016).

Riksbank Prize in Economic Sciences (in Memory of Alfred Nobel) to Abhijit Banerjee, Esther Duflo and Michael Kremer. As said in the headline of the announcement, the prize was awarded "for their experimental approach to alleviating poverty."

The discussion in this section looks first at the reasons for the influence of the randomistas on development research. It then asks if their influence has been justified.

Why have the randomistas had so much influence? Propagating the view that (when feasible) RCTs dominate purely observational studies in attempting to infer causation has clearly held sway. The landing page of J-PAL's website tells us that: "Our mission is to reduce poverty by ensuring that policy is informed by scientific evidence." Toward that aim, J-PAL only does RCTs. Strictly that does not imply that J-PAL's researchers think that observational studies are unscientific (and, independently of J-PAL, many J-PAL-affiliated researchers have used methods of observational study). However, in this context, the phrases "scientific evidence" and (another favorite, including on J-PAL's website) "rigorous evidence" are code for RCTs in the eyes of most readers, and that is plainly intentional. The implication is even stronger than the "gold standard" claim: for some advocates, RCTs are not just top of the menu of approved methods, nothing else is on the menu!

The appeal of RCTs reflects in part the challenges faced in identifying causal impacts. Since the 1990s, we have seen a welcome rise in the attention given to identification problems in economics,[35] though it has been argued that this partially displaced attention from other important issues, such as measurement errors (Gibson, 2019). More critical attention has been given to the validity of IV estimators. It is easy to show that a failure of either of the aforementioned conditions for a valid IV can bias the estimate—possibly more so than for Ordinary Least Squares (OLS), which treats placement as exogenous. It was not hard for researchers to find exogenous variables that are correlated with selected treatment status (though they still needed to pass the appropriate tests). Accepting the exclusion restriction (that the IV is irrelevant to outcomes given treatment status and the control variables) on theoretical grounds was often far more challenging. There were some cases in which the IV could be readily accepted, but this was not always so. From the mid-1990s, seminar audiences and referees were regularly pointing to reasons why the IV in specific papers could have an effect on outcomes that

[35] In common usage, a parameter is said to be "identified" when its value can, in principle, be derived from the data under certain assumptions.

is independent of the endogenous variable. In due course, some economists started to reject almost any attempt at establishing causality without random assignment.

If one only wants to know the difference in mean outcomes between those assigned the option for treatment and those not—which is called the Intent-to-Treat (ITT) parameter—then randomization side-steps these concerns about IV estimates. Given randomization, the treatment assignment is exogenous, uncorrelated with the regression error term. However, ITT can be a rather limited parameter. It is sometimes defended as "policy-relevant" in that the policy is often the assignment of the <u>option</u> for treatment. Yet how would you react to finding that the mean impact is (say) zero among those offered treatment, but positive among those who took it up? Such a finding would surely be of interest to policymakers and citizens. In learning from an RCT, a prospective adopter of the treatment will want to know mean impact for those treated.

Take-up of a randomly-assigned treatment with human subjects is never assured, and compliance is typically endogenous. So the econometric problem often returns in practice. The randomistas have a solution: use randomized assignment as the IV for actual treatment. Clearly, take-up requires assignment, so this IV is correlated with treatment status. Since it is random, the IV is also uncorrelated (in expectation) with the error term when the treatment effect is common across the population. (The discussion will return to the complications that can arise when impacts vary in a way that is unknown to the researcher but known to each participant, who responds accordingly.)

Beyond these econometric arguments, a number of other factors contributed to the randomistas' growing influence from the early 2000s. Researchers who did not use randomization started to be criticized by the randomistas, and their papers started to get ignored in citations to the relevant literature. Some of this took the form of referees' comments on journal articles, which are not public. Journal editors do not need to accept such critiques, though the leading randomistas appear to have been influential and in due course became quite prominent among the editors and editorial boards of economics journals. At times, the critiques also took a public form, such as the study by Finkelstein and Taubman (2015), which questioned the fact that observational and other non-random methods are often used in evaluating health-care delivery policies. This finding was then reported in the *New York Times* under the heading "Few Health System Studies use *Top* Method, Report Says" (Tavernise, 2015; my emphasis), where the "top" is explicitly taken to be an RCT. The message here is clear, though it is less clear that it is right. Some public health specialists have argued that there has been too much attention to evaluations for individual treatments at the expense of research on health systems.[36]

The leading randomistas also did a good job in teaching others how to use their preferred method.[37] Development economists got up to speed quickly, as did some NGOs. They have also been steadily raising the bar on what constitutes a good RCT, though the observation of Heckman and Smith (1995) that RCTs get less critical scrutiny than other methods still seems true today.

Another factor enhancing their influence is that J-PAL's founders professed their desire to make the world a better place through evidence-based policymaking. This was J-PAL's declared motivation from the outset. By this view, doing many RCTs lets us figure out what works and what does not, to scale up the former and scale down the latter (Banerjee, 2006). An analogy is drawn with RCT's in clinical trials, as used to find out what drug works best on average (Favereau, 2016).

Some followers have clearly been attracted by the zeal of the leading randomistas. By this view, "… the experimental ethic has been proposed as the way to change the spirit of development" (Donovan, 2018, p.27). The randomistas can be seen in part as an epistemic movement that attracts its "true believers."[38] The movement's faith in RCTs promises its followers a "quiet revolution" (Banerjee and Duflo, 2011, p.265).

Supporters (including donors) have also been attracted by the simplicity of RCTs—that they are "more transparent and easier to explain" (Duflo, 2017, p.17). It is easier for non-economists to understand an RCT than the methods often favored in observational studies, which were also getting increasingly sophisticated, and technically demanding, by the time J-PAL was founded.

Is the randomistas' influence justified? As Section 2.2 argued, the statistical foundations do not tell us that (when feasible) RCTs are invariably more reliable, whatever the context, and so sit at the top of the hierarchy of methods. This is more a matter of faith than science. The rejection of methods using non-random assignment in some quarters has clearly been an over-reaction to the challenges faced in identifying causal effects this way.

---

[36] See, for example, Rutter et al. (2017).

[37] An example is the excellent "RCT toolkit" produced by Duflo et al. (2011). The World Bank's Development Impact blog has provided a great deal of useful methodological support for doing RCTs.

[38] A reviewer of Leigh (2018) describes the author as a "true believer" and then recounts the various personal choices that Leigh makes in his life based on the results of RCTs (Wydick, 2018).

Nor is the analogy to clinical trials persuasive. It is unclear that the idea of using black-box RCTs to figure out what works and what does not in development is feasible given the dimensionality in both interventions and contexts. Too often, the arguments made for RCTs lack a clear economic rationale for the intervention, or a coherent structure for understanding why it may or may not work (Heckman and Smith, 1995).

While the development randomistas were pointing to clinical trials as the model, medical researchers were taking a more nuanced view.[39] On the one hand, some of the recent literature suggests that past concerns about bias in causal observational health and medical studies have been exaggerated. On the other, it now seems to be accepted that gains from removing systematic bias need to be weighed against the costs and risks of clinical RCTs.

Yet, putting these points to one side, it must be recognized that the medical context is different. Economists (and other social scientists) are dealing with people (as individuals and groups) in social and/or economic contexts in which they can be expected to exhibit greater heterogeneity, and almost certainly greater agency, than is likely in clinical trials. We may often know rather little about the specific setting *a priori*.

Some deeper inferential issues lie under the surface of the randomistas' claims—issues that are known to the experts on both sides but poorly understood more broadly. There is almost certainly some unobserved heterogeneity in the impacts of treatment. There are many sources, including both the circumstances of the individual (such as past experience with the type of intervention) and the effort made by agents (reflecting their beliefs about the impact).[40] Such heterogeneity raises the question of "impact for whom?" This was answered by Angrist et al. (1996), who showed that the IV estimator is giving the mean impact for a subset of the treated, namely the "compliers," induced to switch their treatment status by the randomized assignment.[41]

When estimating the mean impact on those treated, the validity of randomized assignment as the IV to address selective take-up can be questioned in the presence of behavioral responses to such unobserved heterogeneity in the impacts of treatment (Heckman and Vytlacil, 2005; Heckman et al., 2006). The differing impacts must then be relegated to the regression error

[39] Examples of the following points are found in Concato et al. (2000), Silverman (2009), Bothwell et al. (2016) and Frieden (2017).
[40] On the latter source see Chassang et al. (2012), who study the implications for the external validity of RCTs.
[41] Also see the discussion in Pearl (2009, Chapter 8).

term, interacting with the selective take-up of the randomized assignment. Those units with high returns to treatment will be more likely to take it up. Then the interaction effect that has now surfaced in the error term must be correlated with the randomized assignment. The exclusion restriction fails. (Of course, this does not matter if one only wants ITT.)

Identifying the impacts of social programs is rarely easy, with or without randomized assignment. Suppose that the latent characteristics that enhance impact at the individual level also matter to the counterfactual outcomes in an RCT with selective compliance. The choice of estimation method then depends crucially on what impact parameter one is interested in, the type of program one is evaluating and the behavioral responses to that program (as shown in Ravallion, 2014). If the latent factors leading to higher returns to treatment are associated with lower counterfactual outcomes then the "IV cure" for endogenous treatment can be worse than the disease. Indeed, the OLS estimator may even be unbiased, despite the selective take-up. The key point is that practitioners need to think carefully about the likely behavioral responses to heterogeneous impacts in each application—similarly to any observational study.

The design of RCTs in practice can also pose threats to identification. The randomized assignment is sometimes done across clusters of individuals, such as villages. Some clusters get the treatment and some do not. Those within a selected treatment cluster are left free to take up the treatment as they see fit. This is a now classic design in development applications.[42] It runs into a problem whenever there is interference within the clusters whereby non-participants in the selected treatment clusters are impacted by the program. For example, the cluster RCT in Ravallion et al. (2015) used an entertaining movie to teach people their rights under India's *National Rural Employment Guarantee Act*. It was impossible to enforce ticket assignments within villages; the movie had to be shown in public places—often open areas of the village. So access to the movie was randomly assigned across villages, with people free to choose whether to watch it. Some did not, but (of course) they can talk with others who did, and this turned out to be an important channel of impact on knowledge. The cluster randomization had to be combined with a behavioral model of why some people watched the movie (Alik-Lagrange and Ravallion, 2019). Only then could the direct treatment effect (watching the movie) be isolated from the indirect effect (living in a village with access to the movie). In this example, the

[42] Of course, if one can use double randomization—randomizing within villages as well as between them—then one can readily address this type of interference (Baird et al., 2017). Cluster randomizations are designed for situations in which within-cluster randomization if not feasible. Such situations are common in development applications.

spillover effects within clusters violate the exclusion restriction, so the use of cluster assignment as the IV for individual take-up performs poorly.

The generic point is that—contrary to the claims about clean identification of the mean causal impact using randomized assignment—assumptions and models are often required in practice. It does not help that the behavioral assumptions underlying studies using randomization are not always explicit (Keane, 2010). Structural approaches, in contrast, force this to happen.

Some concerns have received less attention in the literature than they merit. One example is Hawthorne effects, whereby monitoring changes behavior. (For example, if you know you are in the control group you may be inclined to seek a substitutable treatment. Or some in the treatment group may try to please the experimenter.[43]) RCTs in economics do not often have the double-blind feature common to clinical trials, so biases associated with monitoring are more likely, and they merit more attention in development applications.[44] A second example is the topic of the next section.

4. Taking ethical objections seriously

Ethical concerns are never far removed from policymaking. There are two dangers of not taking the ethics of evaluation seriously. First, morally unacceptable evaluations may end up being done, and possibly more often in poor places with vulnerable populations and weak institutions for protecting their rights. Second, socially valuable evaluations may be blocked as too risky politically, largely in ignorance of the benefits.

RCTs have been criticized on the grounds that "randomizers are willing to sacrifice the well-being of study participants in order to 'learn'" (Ziliak and Teather-Posadas, 2016).[45] Critics have often pointed out that in an RCT some people who need the treatment are not getting it, while others receive a treatment they do not need. The criticism is also heard that RCTs in poor countries do not get the same ethical scrutiny that is expected (though by no means assured) in rich countries.[46] In using RCTs for clinical trials of potentially hazardous treatments, there have

[43] One RCT randomly assigned knowledge about the experimenter's intent, but did not find any significant effect (Mummolo and Peterson, 2018). Further tests of this sort are needed.
[44] This aspect of the difference between economic RCTs and clinical RCTs is discussed further in Favereau (2016). For a useful overview of the Hawthorne effect in the health field see Friedman and Gokul (2014).
[45] Also see the comments in Barrett and Carter (2010), Baele (2013) and Mulligan (2014).
[46] In the US, the ethics of using RCTs for the evaluation of Federal social policies has not received the same attention as for clinical trials. Blustein (2005) discusses the reasons.

been some well documented cases in which participants in developing countries were largely unware of the health risks they faced if they end up being treated.[47] Baele (2013) argues that the development randomistas have not paid enough attention to the ethics of their RCTs. Glennerster and Powers (2016) offer a cautious ethical defense of RCTs against their critics.

Ethical validity is not a serious issue for all evaluations. Sometimes an impact evaluation is built onto an existing program such that nothing changes about how the program works. The evaluation takes as given the way the program assigns its benefits. So if the program is deemed to be ethically acceptable then this can be presumed to hold for the evaluation. We can dub these "ethically-benign evaluations."

Other impact evaluations deliberately alter the program's (known or likely) assignment mechanism—who gets the program and who does not. Then the ethical acceptability of the intervention, as it normally works at scale, does not imply that the evaluation is ethically acceptable. Call these "ethically-contestable evaluations." The main examples in practice are RCTs. Scaled-up programs almost never use randomized assignment, so the RCT has a different assignment mechanism, with potentially large differences in the benefits, given the likely heterogeneity in impacts. An RCT can be contested ethically even when the real program is fine.

It is surely a rather extreme position (not often associated with economists) to say that good ends can never justify bad means. It is ethically defensible to judge processes in part by their outcomes; indeed, there is a long and respected view in moral philosophy that consequences often trump processes—with utilitarianism as the leading example. It is not inherently "unethical" to do an RCT as long as this is deemed to be justified by the expected benefits from new knowledge. However, the consequential benefits do need to be carefully weighed against the process concerns. This is especially so in the (many) instances in which a feasible, and ethically benign, observational study is an option.

Ethics has been much discussed in medical research where the principle of equipoise requires that there should be no decisive prior case for believing that the treatment has impact.[48] Only if we are sufficiently ignorant about whether it is better to be in the treatment group or the

---

[47] See, for example, Sathyamala (2019) on an RCT used to study the health risks of a contraceptive drug in Africa.
[48] There is a good discussion in Freedman (1987), which introduced the principle of equipoise in clinical trials. In the context of development impact evaluations, see Baele (2013) and McKenzie (2013).

control should we randomize at all, or continue with an RCT.[49] If evaluators are to take ethical validity seriously then some development RCTs will have to be ruled out as unacceptable given that we are already reasonably confident of the outcomes—that the gain from knowledge is not likely to be large enough to justify the ethically-contestable research.[50]

The principle of equipoise is rarely applied to RCTs for development and social policies. Indeed, there may well be a tendency in the opposite direction. A recent call-for-proposals from a prominent philanthropic funder gave explicit preference to any RCT proposal "That is backed by highly-promising prior evidence, suggesting it could produce sizable impacts on outcomes…" (Arnold Foundation, 2018, p.2). At one level, one can understand the funder's preference, given that RCTs are costly and there is a desire to have impact with limited resources. Some *ex ante* filters of this sort make sense. (One would not want to fund an RCT for an intervention that is unlikely to turn out to be feasible on the ground.) However, the above example points to a tension between donor objectives and ethical concerns. *Ex ante* confidence of "sizeable impacts on outcomes" leaves one worried about withholding a treatment from those who need it (and wasting treatment on those who do not). This also points to a concern about the funding processes determining what gets evaluated. Section 2.5 returns to this topic.

There have been some ethical defenses of RCTs. One view is that RCTs are justified whenever rationing is required; when there is not enough money to cover everyone, it is argued that randomized assignment is a fair solution.[51] This makes sense when information is very poor. In some development applications, we may know very little *ex ante* about how best to assign participation to maximize impact. Nevertheless, when alternative allocations are feasible and one does have prior information about who is likely to benefit, it is surely fairer to use that information, and not randomize, at least unconditionally.

It has also been argued that the method of conditional randomization (also called "blocked" or "stratified" randomization) can relieve ethical concerns. The idea here is that one first selects eligible types of participants based on prior knowledge about likely gains, and only then randomly assigns the intervention, given that not all can be covered. For example, if one is evaluating a training program or a program that requires skills for maximum impact, one would

[49] The "we" here is best thought of as a set of people with sound knowledge of the relevant literature and experience. This is sometimes called "community equipoise."
[50] See the examples discussed in Barrett and Carter (2010), Ziliak and Teather-Posadas (2016) and Narita (2018).
[51] See, for example, Goldberg's (2014) comments on Mulligan (2014). The same point is made by Fiennes (2018).

reasonably assume (backed up by some evidence) that prior education and/or experience would enhance impact, and then design the evaluation accordingly. This has ethical advantages over pure randomization when there are priors about likely impacts.

There is a catch. The set of things observable to the evaluator is typically only a subset of what is seen on the ground. At (say) village level, there will often be more information than is available to the evaluator—information revealing locally that the program is being assigned to some who do not need it, and withheld from some who do. But whose information should decide the matter? Pleading ignorance seems a lame excuse for an evaluator when other stakeholders do in fact know very well who is in need and who is not.

It has also been argued that encouragement designs are less contentious ethically. The idea is that nobody is prevented from accessing the primary service of interest but the experiment instead randomizes access to some form of incentive or information. This does not remove the ethical concern—it merely displaces it from the primary service of interest to another space. Ethical validity still looms as a concern when the encouragement is being deliberately withheld from some people who would benefit and given to some who would not.

Consider, for example, the RCT in Bertrand et al. (2007). One treatment arm provided a large financial reward to those participants who could quickly obtain a driver's license in Delhi India, which facilitated bribes to licensing officials. The RCT did not pay bribes directly or give out licenses to people who did not verifiably know how to drive, but these were predictable outcomes. The expected gain from this RCT was a clean verification of the proposition that corruption happens in India and has real effects. However, there does not seem to have been any serious prior doubt about the truth of that claim.

RCTs can be designed to help address ethical concerns. One option is to use an "equivalence trial" for which the control group gets what is thought to be the next best treatment.[52] Possibly in contrast to biomedical settings, there may be little agreement on the <u>best</u> option in each specific development application. Nonetheless, it seems unlikely that the common use of the "do-nothing" or placebo control would pass close ethical scrutiny in most development applications. There is usually some option. (Nor is "doing nothing" likely to be a particularly relevant counterfactual for most policy makers.)

---

[52] This idea has been a much debated in biomedical applications, notably in the context of the revisions done in 2000 to the World Medical Association's 1964 Helsinki Declaration. For further discussion see Levine (2006).

Another option is adaptive randomization. This is feasible when there is a sequencing of assignment, with observed responses at each step. Adaptive randomizations change the assignment along the way, in the light of the accumulated evidence on impacts.[53] Narita (2018) has proposed an interesting market-like adaptive design for social experiments, whereby one takes account of each participant's willingness-to-pay for the chance of treatment, given prior knowledge about impacts.[54] Unlike a classical RCT, one ends up with a Pareto efficient experiment, though with similar statistical properties for the impact estimates. At the time of writing, this idea does not appear to have been implemented in the field.

In the US and elsewhere, Institutional Review Boards (IRBs) have become common for proposed studies with human subjects. There is a designated IRB for most research institutions. They are largely self-regulating. Beyond occasional anecdotes, there does not appear to have been a systematic assessment of how well IRB processes have worked for development RCTs. One thing seems clear: IRBs need to give more attention to assessing the expected benefits of an ethically-contestable evaluation given prevailing knowledge. Syntheses of current knowledge can help and these are becoming more common.[55]

If pressed, many randomistas acknowledge the ethical concerns reviewed above, though they rarely give them more than scant attention in their papers. They assume (more often implicitly) that their RCTs generate benefits that outweigh such concerns. Whether that is true is rarely obvious, and merits more attention.

We should also ask how well research efforts match the knowledge gaps. Imbalances of this sort raise further ethical concerns, given pressing development challenges and limited resources for research. The next section takes up these issues.

5. Relevance to policymaking

While there is clearly a lot more to good policymaking than good evidence, policymakers increasingly turn to evidence, hoping to inform their choices, and win political debates. The policy-relevance of evaluative research matters.

---

[53] These are getting serious attention in biomedical research. For example, the US Food and Drug Administration (2010) has issued guidelines for adaptive evaluations. Also see Cox and Reid (2000, Chapter 3).
[54] Also see Chassang et al. (2012) and the discussion in Özler (2018).
[55] These are sometimes referred to as systematic reviews; see for example, the 3ie searchable database and the Campbell Collaboration on such reviews.

To my knowledge, there has not yet been a comprehensive and objective assessment of the influence on development policy of all those RCTs. Nonetheless, one can point to examples of policy-relevant research using RCTs. To give just one example, Banerjee et al. (2014) used RCTs in six countries (Ethiopia, Ghana, Honduras, India, Pakistan, Peru) to evaluate the long-established approach taken against poverty by BRAC using a combination of transfers (assets and cash) targeted to the poorest with literacy and skill training.[56] The researchers found economic gains from adopting BRAC's approach some three years after the initial asset transfer, and one year after the disbursements finished. If one is willing to extrapolate the earnings gains into the distant future—although that is clearly a strong assumption—then their present value often exceeds the cost of the BRAC-type program (Banerjee et al. 2014).

Without aiming to provide a comprehensive assessment, this discussion points to some limitations of RCTs for informing development policy, drawing on the literature.

Policy-relevant parameters: Even under ideal conditions, an RCT is only well-suited to estimating a rather narrow subset of the parameters of interest to policymakers. In reality, one expects that there will be both gainers and losers, depending on the context and the characteristics of participating units (and, as noted, some of those characteristics are unobserved to the analyst, though still motivators of behavior, including whether or not to take up the treatment). There is a distribution of impacts. Policymakers may want to know what proportion of the population benefit, and what proportion lose, or what types of people gain and what types lose. Identifying these policy-relevant parameters will typically require more data and more structural-econometric methods. A full-blown structural model need not be essential for addressing the question of interest, but (at the other extreme) an RCT will rarely deliver what is needed.

There are ways of reliably learning more about individual impacts than simply their mean. For example, the Local Instrumental Variables estimator proposed by Heckman et al. (2006) aims to identify the marginal treatment effects (MTEs) at all values of the empirical probability of being treated. Unlike a standard RCT, "selective trials" allow one to identify the MTEs by basing the probability of assignment to treatment (rather than control) on agents'

[56] BRAC now stands for *Building Resources Across Communities*. The NGO started in Bangladesh (where it was once called the *Bangladesh Rural Advancement Committee*) but now works in many countries.

expressed willingness to pay (Chassang et al., 2012). One can then aggregate up to get the mean impact, as would be identified by an RCT. But one learns a lot more than the average impact.

Sometimes it is also possible to reliably ask counterfactual questions in surveys. This is done in Murgai et al. (2016), who asked participants in a workfare program what they think they would be earning otherwise (with observational checks against local labor markets). Then one can learn more about the distribution of impacts, though (of course) there are measurement errors in survey responses, so some averaging will almost certainly be required.

An aspect of performance that is often of interest to policymakers is who benefits from the program, as determined (in part) by the assignment mechanism implied by its design. If it is demand driven, what are the characteristics of those choosing to take it up? If it is rationed, to whom? Such questions come at the first stage in an important class of observational methods using matching that start with a statistical model of who gets the program and who does not.[57] Of course, if it is an RCT then, in expectation, the assignment is not predictable, and if there is full compliance then nothing can be learnt about the types of people likely to participate when the program is scaled up.

With imperfect compliance, we can learn about this from the first stage of the aforementioned IV estimator. Indeed, as Heckman and Pinto (2019) argue, once we recognize that take-up of the randomized assignment is the outcome of rational choice, we can use it to study both the determinants of participation and identify a wider range of casual parameters. For example, by varying the incentives in the classic RCT, and invoking the weak axiom of revealed preference, the results in Heckman and Pinto (2019) can be applied to the problem of low take-up of social policies, which is often common among poor and/or socially excluded people. Instead of thinking about selective compliance by human subjects as a statistical nuisance we can learn from it.

An RCT might also be used to assess likely impact *ex ante*, and then later do a separate evaluation of the actual program at scale, using an observational estimator. This sounds promising but it should be understood that, given selective take-up and heterogeneous impacts, one has essentially evaluated two different programs, only one of which is actually implemented

[57] This refers to propensity-score matching. The predicted values of that model are the “propensity scores” used in selecting observationally-balanced treatment and comparison groups (Rosenbaum and Rubin, 1983).

by the government. It is not hard to guess which will be of greater interest to policymakers. Will the second evaluation be done? Possibly not if one takes the "gold standard" view.

At the heart of the problem of learning about policy effectiveness is that an RCT is a rather artificial construction, unlike almost any imaginable real-world policy.

External validity: Policymakers naturally want to learn from such an experimental trial about how the same intervention might perform in another setting. This is a question about external validity. This can be in doubt for a number of reasons, including monitoring effects, general equilibrium effects, sampling problems and specific care in providing the treatment in the RCT (Duflo et al., 2008).

Such issues are often ignored in papers documenting development RCTs, or the issues are only given a superficial treatment. For the majority of the 54 development RCTs published in eight economics journals (2009-14), Peters et al. (2018) find that the sources of external invalidity are not addressed and the information to address them is not provided. If different RCTs on a given intervention tended to agree then we can be more confident about external validity. But that is not the case. Vivalt (2017) has documented the variance found in the impact estimates for a given program across settings (and even types of evaluators). Her findings warn against generalizations. As Vivalt also notes, poor documentation of contextual factors does not help. Pritchett and Sandefur (2015) provide examples (for microcredit schemes) in which a (presumed) internally-valid RCT done in one context is inferior to an observational study for predicting impact in another context. Not all of this variability in estimates is due to heterogeneity in the true impacts; an estimate for seven microcredit RCTs found that 60% of the variability is due to sampling variation (Meager, 2018). In practice, policy makers will not be able to easily distinguish sampling variation from true impact variability.

The advantages of working with NGOs in doing RCTs (Section 2.2) have also raised questions about external validity. An example is found in the RCT on schooling in Kenya by Duflo et al. (2015). Randomly chosen schools were given the resources to hire an extra teacher working on a short-term contract. Children with the contract teachers were found to do significantly better in test scores than those with regular civil-service teachers. This experiment was implemented by a local NGO. However, Bold et al. (2018) attempted to replicate this at scale, using a follow-up RCT, but this time with an arm implemented by the government (as well as one by the NGO). This revealed that it was NGO-implementation that led to test-score gains, not the type of teacher. The teacher-effect found by Duflo et al. (2015) had vanished.

A "black box" reduced-form estimate (whether from an RCT or not) is not very informative for many purposes of policymaking. Learning from RCTs poses specific problems. Consider how we might learn about scaling up from an RCT (which is surely an important aim). An RCT randomly mixes low-impact people (for whom the program has little benefit) with high-impact people, based on latent attributes. It is plausible that the scaled-up program will have higher representation from the high-impact types, who will be attracted to the program.[58] Given this purposive selection based on the (heterogeneous) impacts, the national program is fundamentally different to the RCT, which may contain little useful information for scaling up.

This reflects a more general point made by Moffitt (2006) that many things can change—inputs and even the program itself—on scaling up a pilot. An NGO keen to demonstrate its worthiness to attract funders will have an incentive to show impact from a trial that is not typical of its normal operations. Young researchers doing a field trial may apply greater effort than the government officials implementing the scaled up version. External validity imposes constraints on the design and execution of pilots that are not given sufficient attention in practice.

One approach to learning about external validity is to repeat the evaluation in different contexts. For example, using an observational method, Galasso and Ravallion (2005) studied the performance of Bangladesh's *Food-for-Education* program in each of 100 villages and correlated the results with characteristics of those villages. The differences in performance were partly explicable in terms of observable village characteristics, such as intra-village land inequality (with more unequal villages being less effective in reaching their poor). Not allowing for such differences has been seen as a serious weakness in past evaluations.[59]

Looking inside the black box of an impact evaluation can throw useful light on its external validity and policy implications. This will often require information external to the original evaluation design. An example is the *Proempleo* RCT by Galasso et al. (2004). Vouchers for a wage subsidy were randomly assigned across people currently in a workfare program, with a randomized control group. The theory is that the wage subsidy will reduce labor costs to the firm and so make hiring the worker more attractive. Consistently with the predictions

[58] This is an instance of what Heckman and Smith (1995) dubbed "randomization bias."
[59] See for example the comments by Moffitt (2004) on trials of welfare reforms in the US.

of the theory, the RCT found a significant impact on employment. However, subsequent checks against administrative records revealed a very low take-up of the wage subsidy by firms. So *Proempleo* did not work the way the theory had assumed. Follow-up qualitative interviews with firms and workers indicated that the vouchers had credential value to workers—a "letter of introduction" that few people had (and the fact that it was allocated randomly was a secret locally in this RCT). This could not be known from the RCT, but required supplementary observational data. (And this had not been anticipated by the researchers *ex ante*, so rigid adherence to a pre-analysis plan would have missed a crucial, policy-relevant, aspect of why the program had impact.) The extra data also revealed the importance of providing information about how to get a job, which carried implications for scaling up. However, scaling up the wage subsidy based on the RCT would have been a mistake.

Some researchers have been using randomization (either of the intervention or of some key determinant of its placement) to throw light on deeper structural parameters. For example, Todd and Wolpin (2006) use the aforementioned RCT for *Progresa* in Mexico to model the dynamic behavioral responses to the schooling incentive provided by that scheme. Such research can help us understand a program's impacts and facilitate simulations of alternative policy designs. Todd and Wolpin show that a switch of the *Progresa* subsidy to higher levels of schooling would enhance overall impacts. In a similar vein, there is scope for using an RCT to test one or more key links in the "theory of change" underlying a program's rationale, even if the tool is not applicable to the program itself. This echoes the arguments of Heckman (1992) and Heckman and Pinto (2019) on the scope for more ambitious experiments informed by theory.

Knowledge gaps: To help antipoverty policymaking, researchers should ideally be filling the gaps between what we know about the effectiveness of policies and what policymakers need to know. This is clearly not happening as well as we might hope. For example, Kapur (2018) recounts an interview with a former Chief Economic Advisor of the Government of India (GOI): "When asked how many of these expensive RCTs had moved the policy needle in India, Arvind Subramanian, Chief Economic Advisor, GOI, was hard pressed to find a single one that had been helpful to him in addressing the dozens of pressing policy questions that came across his table."[60]

[60] Also see the comments in Basu (2014) (another ex Chief Economic Advisor of GOI.)

Why do these knowledge gaps exist? There are random factors but there are also more systematic "knowledge market failures" (Ravallion, 2009b). One source is the existence of externalities in evaluations. There is evidence that having an impact evaluation in place for an ongoing development project can help improve some aspects of its implementation, such as its speed of disbursement (Legovini et al., 2015). However, the knowledge gains from an evaluation also bring benefits to future projects, which (hopefully) draw on the lessons learnt from prior evaluations. Current project managers cannot be expected to take proper account of these external benefits when deciding how much to spend on evaluating their own project. There are clearly larger externalities for some types of evaluations, such as those that are more innovative—the first of their kind. The externalities in evaluation also play a role in the "myopia bias" that has been noted in development applications, such that long-term evaluations are rare (Ravallion, 2009b; Bouguen et al., 2019).

Knowledge market failures also stem from publication biases originating in both the selection processes of journal editors and the behavior of authors, including in documentation. Null results are less likely to be published or even written up.[61] Subsequent replications of experiments in economics often find less strong effects.[62] In some cases, the prior results have been adequately replicated but in the process have been found to be highly sensitive to questionable aspects of the data analysis that had not been obvious in the original paper.[63]

The dynamics of publication processes are a further source of persistence in knowledge gaps. Errors occur in the literature and it can take time to correct them. In recognition of its originality, the first paper on a topic may well be published prominently. Subsequent papers will tend to be relegated to lesser journals, cited less often, or may even have a hard time being published at all. The author of the original paper becomes the gatekeeper of knowledge on the topic. The gatekeeper is sometimes passable, but still has considerable influence. However, the

[61] Among 221 social science studies it was found that "Strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up" (Franco et al., 2014, p.1502). The distribution of reported p-values in papers published in the AER, QJE and *Journal of Political Economy* suggests that researchers tend to make specification choices that inflate the significance of their results to get over a "5% significance" hurdle (Brodeaur et al., 2016). Christensen and Miguel (2018) survey the evidence on these biases in published economic research and discuss how the biases might be reduced.

[62] Camerer et al. (2016) replicated 18 laboratory experiments published in the *American Economic Review* (AER) and *Quarterly Journal of Economics* (QJE). On average, the replicated effect size was one third lower than the original.

[63] See, for example, Bédécarrats et al. (2019), which casts doubt on both the internal and external validity of the original RCT by Crépon et al. (2015).

first paper may not have got it right. On top of this, the incentives for effort at replication appear to be weak in economics.[64] (Yet in the sciences, failures to replicate have been common; see Ioannidis, 2005.) Thus, the first draw from the distribution of impacts can have a lasting distortionary effect on accepted knowledge.

External invalidity also raises concerns about the process of knowledge accumulation. Even if the first paper came close to the truth in the specific context, it may have limited validity in different circumstances. When the topic concerns the impact of a policy, or an issue that is very relevant to that impact, policy knowledge will tend to be skewed accordingly.

These are generic concerns, not confined to RCTs. However, the "gold standard" method-hierarchy could well make things worse, as we will now see.

Matching research efforts with policy challenges: Knowledge gaps also stem from miss-alignments of evaluative effort. One aspect is that development evaluators too often ignore the scope for fungibility. Recipients (governmental or not) can re-allocate their own efforts in response to new funds, such as development aid. Donors are often implicitly funding something else. A less well-known implication is that donors and higher levels of government may well be evaluating the wrong thing from the point of view of assessing their own impact—they evaluate the project that the aid recipient put up for funding rather than the project that was actually funded, given the scope for fungibility. Then evaluative efforts are miss-aligned with development efforts.

RCTs are not to be blamed for this. However, strong methodological preferences on the part of evaluators can readily reinforce the miss-alignment. The development randomistas have had both output and substitution effects on knowledge. There is at least the suggestion of a positive output effect in the fact that we have seen a great many more RCTs since 2000 (Figure 1). However, as discussed already, neither the internal nor the external validity of these development RCTs is fully evident. We do not know the counterfactual—what we would have learnt if those resources (financial and human capital) had been deployed elsewhere.

The substitution effect relates to the methods used. Take, for example, the World Bank. While the earliest RCT in the 3ie database is by the Bank, until the early 2000s the tool was seen as only one of many credible options for IE. Since then there has been a marked switch in favor

[64] See the discussion in Rodrik (2009). Since then, 3ie has supported replication efforts for development impact evaluations through its Replication Window and its *Journal of Development Effectiveness*.

of RCTs within the Bank, which has been applauded by some observers; for example, an editorial in *The Lancet* declared (in ignorance of more than the history) that "The World Bank is finally embracing science." (Lancet, 2004, p.731).[65] The Bank's Independent Evaluation Group (IEG) reports that over 80% of the impact evaluations starting in 2007-10 used randomization, as compared to 57% in 2005-06 and only 19% in prior years (World Bank, 2012).

Even if we presume that all those RCTs had a positive output effect on knowledge, the substitution effect could well work in the opposite direction. There are three aspects of the substitution effect. First, the emphasis on identifying causal impacts using RCTs has deflected attention from other methods of empirical investigation, including descriptive research, which is surely undervalued in development research today. Some of the policy lessons emerging from RCT research papers could have been derived from good "thick" descriptions (using qualitative and/or quantitative methods) of the real-world processes linking interventions to outcomes.

Second, there is a concern that the emphasis on assigned individualized programs has deflected attention from systemic research, typically using structural models. In economics more broadly, the decline in attention to structural work in teaching and research has been noted by Keane (2010) and others. This has also been a raised as a specific concern for research on public health (Rutter et al., 2017).

Third, a problem in evaluating the impact of the portfolio of development policies is that randomization is only feasible for a non-random sub-set of policies and settings. The implication is that we lose our ability to make inferences about a broad range of policies if we rely solely on RCTs. As a generalization, randomization tends to be better suited to programs with clearly identified participants and non-participants, relatively short time horizons, that do not require imposing charges/taxes, and for which there is little scope for the costs or benefits to spillover to the group of non-participants. Thus RCTs make more sense for private goods, which are easy to assign across individual households, than public goods with benefits shared across many people (Hammer, 2017). There are exceptions (such as certain local public goods). However, it is generally far more difficult to randomize the location of medium- to large-scale infrastructure projects and seemingly impossible to randomize sectoral and economy-wide reforms. This makes the tool of limited use for some core activities in any country's development strategy.

[65] On the influence of RCTs at the World Bank see Webber and Prouse (2018).

Evaluations of the impacts of providing private goods beg for an economic rationale for the "policy." Would not markets provide the private good efficiently, eliminating the need for any impact evaluation? There may be good reasons why an evaluation for a private good is needed in specific contexts, but more often it seems that the randomistas are simply chasing opportunities for randomization. Granted, redistributive goals are mentioned at times, but in a rather casual way. Distributional impacts (such as on poverty) are rarely addressed with any rigor, or even identified as explicit outcomes. In short, the public economics is often missing.

To give an example of how an insistence on using RCTs distorts knowledge for policy making, consider deforestation in developing countries. A common scenario is that forest-owning households cutting down their trees do not take account of the external cost of their contribution to global warming. A solution has long been known, namely a Pigouvian tax. But this would be hard to implement as an RCT, since the power to tax mostly lies with governments, who would (understandably) be resistant. Instead, one can randomize payments to those who choose not to cut down their trees, as in the RCT for Uganda by Jayachandran et al. (2017). This policy can be implemented by a local NGO, bypassing the government. Here there is a public-economic rationale, but the use of an RCT constrains the policy options evaluated. And the tax policy will probably have different impacts (if only because the payment policy gives extra value to the stock of trees, generating an income effect, separately to the price effect).

Of course, no single tool can cover all applications. The question here is whether we have a reasonable balance today between research effort and policy challenges. The (questionable) hierarchy of methods advocated by the randomistas makes it harder to attain that balance. Indeed, even for private goods, the very idea of randomized assignment is antithetical to the goals of many development programs, which typically aim to reach certain types of people or places. In delivering cash transfers to poor people—a favorite intervention for development RCTs—governments will hopefully be able to do better than a random assignment.

The aforementioned IEG report documents the unbalanced assignment of World Bank impact evaluations across the sectors of its operations, and the seemingly poor fit of the evaluation portfolio to the Bank's sectoral and development priorities (World Bank, 2012). Though I have not seen evidence, I suspect that there is also an imbalance in the assignment of evaluative effort according to the likely duration of project benefits. Long-term evaluations of World Bank development projects are rare, despite the claims made about longer-term impacts. I can testify from personal experience how hard it is to organize and implement long-term evaluations at the World Bank.[66] It is plausible that favoring RCTs exacerbates a myopia bias in development knowledge.

This is not just happening in the World Bank. The sectoral bias in the use of RCTs more broadly is evident from the results of Cameron et al. (2016) who provide a cross-tab of over 2,200 published impact evaluations (in the aforementioned 3ie database) by method and sector.[67] Overall, about two-thirds of these evaluations use RCTs, but the RCTs tend to be concentrated in certain sectors, notably education (58% used an RCT), health, nutrition and population (83%; 93% in health alone), information and communications technology (67%), and water and sanitation (72%). Observational studies are more common—with under one-third using an RCT—in agriculture and rural development, economic policy, energy, environment and disaster management, private sector development, transportation, and urban development. The production of impact evaluations has also been uneven geographically (even allowing for population). India has had the largest absolute number but Kenya has had the most per capita.[68] The geography of RCT placement is influenced by researcher connections with local NGOs.

There are both supply and demand sides to this bias. On the supply side of evaluations, the reality today is that, enamored by the promise of cleanly identifying a causal effect, many economists and other social and political scientists have been searching for something to randomize. If randomization is not feasible, they turn to ask another question.

On the demand side, governments (and development agencies) are largely free to choose what is evaluated. One concern here is that they do not always know what evidence they need (Duflo, 2017). Politics also plays a role. They may be drawn to pick programs for which there is little risk that a negative appraisal will hurt politically, or to pick those that do matter but for which there are good reasons to be confident of a politically acceptable result (again raising ethical concerns). Other important programs will not be evaluated. The risks are plain.

Addressing these concerns calls for more strategic evaluation agendas, not driven by the methodological preferences of researchers. We have started to see more strategic agendas for RCTs. This is welcome, though the strategies are still led by academic researchers, based on their

[66] This largely based on the study reported in Chen et al. (2009).
[67] In addition to RCTs the methods identified are difference-in-differences, instrumental variables, regression discontinuity and matching. Multiple methods are allowed in the counts.
[68] For details see Cameron et al. (2016) and Sabet and Brown (2018).

interests and devoted to one tool. If we are really concerned about obtaining unbiased estimates of the impact of the portfolio of development policies it would surely be better to carefully choose (or maybe even randomly choose!) what gets evaluated, and then find the best method for the selected programs, with an RCT as only one option. That is what is called for if we take seriously the goal of obtaining an unbiased assessment of overall development impact. Research can serve that goal, but it is unlikely to happen automatically.

6. Conclusions

We are seeing a welcome shift toward a culture of experimentation in fighting poverty, and addressing other development challenges. RCTs have a place on the menu of tools for this purpose. However, they do not deserve the special status that advocates have given them, and which has so influenced researchers, development agencies, donors and the development community as a whole. To justify a confident ranking of two evaluation designs, we need to know a lot more than the fact that only one of them uses randomization.

The popularity of RCTs has rested on a claimed hierarchy of methods, with RCTs at the top, as the "gold standard." This hierarchy does not survive close scrutiny. Despite frequent claims to the contrary, an RCT does not equate counterfactual outcomes between treated and control units. The absence of systematic bias does not imply that the experimental error in a one-off RCT is less than the error in some alternative non-random method. We cannot know that. Among the feasible methods in any application (with a given budget for evaluation), the RCT option need not come closer to the truth. Indeed, if the sample size for an observational study is sufficiently greater than for an RCT in the same setting, then the trials by observational study can be more often close to the truth even if they are biased.

There is still ample scope for useful observational and other non-random studies (such as deterministic experimental assignments), informed by theory. Yes, there is model uncertainty, though generally not as much as the randomistas assume. Moreover, when we look at RCTs in practice, we see them confronting problems of miss-measurement, selective compliance and contamination. Then it becomes clear that the tool cannot address the questions we ask about poverty, and policies for fighting it, without making the same type of assumptions found in observational studies—assumptions that the randomistas promised to avoid.

RCTs are also ethically contestable in a way that observational studies are not. The ethical case against RCTs cannot be judged properly without assessing the expected benefits from new knowledge, given what is already known. Review boards need to give more attention to the *ex-ante* case for deliberately withholding an intervention from those who need it, and deliberately giving it to some who do not, for the purpose of learning. There may be a good case in specific contexts, based on the limitations of existing knowledge, but the case does need to be made in a credible way and not just taken for granted.

The questionable claims made about the superiority of RCTs as the "gold standard" have had a distorting influence on the use of impact evaluations to inform development policymaking. The bias stems from the fact that randomization is only feasible for a non-random subset of policies. When a program is community- or economy-wide or there are pervasive spillover effects from those treated to those not, an RCT will be of little help, and may well be deceptive. The tool is only well suited to a rather narrow range of development policies, and even then it will not address many of the questions that policymakers ask. Advocating RCTs as the best, or even only, scientific method for impact evaluation risks distorting our knowledge base for fighting poverty. That risk was one of the main concerns in Ravallion (2009a), and the experience since then has reinforced that concern.

While we have seen much progress over the last 10 years, there are still grounds for doubting whether evaluative research on development fits well with the policy challenges now faced. This paper has argued that a better alignment requires:

- Abandoning claims about an unconditional hierarchy of methods, with RCTs at the top, and making clear that "scientific" and "rigorous" evidence is not confined to RCTs.
- Demanding a clear and well-researched *ex ante* statement of the expected benefits from an RCT, to be weighed against the troubling ethics.
- Making explicit the behavioral assumptions underlying randomized evaluations, similarly to the standards of structural approaches.
- Going beyond mean causal impacts, to include other parameters of policy interest and better understanding the mechanisms linking interventions to outcomes.
- Viewing RCTs as only one element of a tool kit for addressing the knowledge gaps relevant to the portfolio of development policies.

References

Alik Lagrange, Arthur, and Martin Ravallion, 2019. "Estimating Within-Group Spillover Effects Using a Cluster Randomization: Knowledge Diffusion in Rural India," *Journal of Applied Econometrics* 34: 110-128.

Angrist, Joshua, Guido Imbens and Donald Rubin, 1996. "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, XCI: 444-455.

Baele, Stéphane, 2013. "The Ethics of New Development Economics: is the Experimental Approach to Development Economics Morally Wrong?," *Journal of Philosophical Economics* 7(1): 1-42.

Baird, Sarah, Aislinn Bohren, Craig McIntosh and Berk Özler, 2017. "Optimal Design of Experiments in the Presence of Interference," *Review of Economics and Statistics*, forthcoming.

Banerjee, Abhijit, 2006. "Making Aid Work. How to Fight Global Poverty—Effectively," *Boston Review*, July/August.

Banerjee, Abhijit, and Esther Duflo, 2009. "The Experimental Approach to Development Economics," *Annual Review of Economics* 1: 151-178.

______________ and ___________, 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, New York: Public Affairs.

______________ and ___________, 2017. "Pushing Evidence-Based Policymaking for the Poor," Livemint, October 16.

Banerjee, Abhijit, Sylvain Chassang, Sergio Monero, and Erik Snowberg, 2018. "A Theory of Experimenters." NBER Working Paper 23867.

Banerjee, Abhijit, Esther Duflo and Michael Kremer, 2019. "The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy," In *State of Economics, State of the World*. Edited by Kaushik Basu, David Rosenblatt and Claudia Paz Sepulveda. Cambridge, Mass.: MIT Press, forthcoming.

Banerjee, Abhijit, Dean Karlan, and Johnathan Zinman, 2014."Six Randomized Evaluations of Microcredit: Introduction and Further Steps," *American Economic Journal: Applied Economics* 7(1): 1-21.

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

Basu, Kaushik, 2014."Randomization, Causality and the Role of Reasoned Intuition," *Oxford Development Studies* 42(4): 455-472.

Bédécarrats, Florent, Isabelle Guérin, Solène Morvant-Roux, and François Roubaud, 2019. "Estimating Microcredit Impact with Low Take-up, Contamination and Inconsistent Data. A Replication Study of Crépon, Devoto, Duflo, and Pariente (American Economic Journal: Applied Economics, 2015)" *International Journal for Re-Views in Empirical Economics* 3.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan, 2007."Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption," *Quarterly Journal of Economics* 122(4): 1639-1676.

Bethlehem, Jelke, 2009. *Applied Survey Methods: A Statistical Perspective.* New York: Wiley.

Blustein, Jan, 2005. "Toward a More Public Discussion of the Ethics of Federal Social Program Evaluation," *Journal of Policy Analysis and Management* 24(4): 824-852.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, 2018. "Experimental Evidence on Scaling Up Education Reforms in Kenya," *Journal of Public Economics* 168: 1-20.

Bornmann, Lutz, and Ruediger Mutz, 2014."Growth Rates of Modern Science: A Bibliometric Analysis based on the Number of Publications and Cited References," *Journal of the Association of Information Science and Technology* 66: 2215-2222

Bothwell, Laura, Jeremy Greene, Scott Podolsky, and David Jones, 2016."Assessing the Gold Standard—Lessons from the History of RCTs," *New England Journal of Medicine* 374(22): 2175-2181.

Bouguen, Adrien, Yue Huang, Michael Kremer and Edward Miguel, 2019. "Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics," *Annual Review of Economics* 11:523–61.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, 2016. "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics* 8(1):1–32.

Brown, Caitlin, Martin Ravallion and Dominique van de Walle, 2018. "A Poor Means Test? Econometric Targeting in Africa," *Journal of Development Economics* 134: 109-124.

Bruhn, Miriam, and David McKenzie, 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics* 1(4): 200-232.

Burtles, Gary, 1995. "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives* 9(2): 63-84.

Cameron, Drew B., Anjini Mishra, and Annette N. Brown, 2016. "The Growth of Impact Evaluation for International Development: How Much Have We Learned?" *Journal of Development Effectiveness* 8 (1): 1–21.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, and T. Chan, 2016. "Evaluating Replicability of Laboratory Experiments in Economics," *Science* 351(6280):1433–1436.

Card, David, and Stefano DellaVigna, 2013. "Nine Facts about Top Journals in Economics," NBER Working Paper 18665.

Cartwright, Nancy, 2007. "Are RCTs the Gold Standard?" *BioSocieties* 2(1): 11-20.

Chabé-Ferret, Sylvain, 2018. "An Approach Combining Theory, Simulations and Empirics Provides Evidence of Regularities in the Bias of Observational Methods," Toulouse School of Economics.

Chassang, Sylvain, Gerard Padró i Miquel and Erik Snowberg, 2012. "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review* 102(4): 1279–1309.

Chen, Shaohua, Ren Mu and Martin Ravallion, 2009. "Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China," *Journal of Public Economics* 93: 512-528.

Christensen, Garret, and Edward Miguel, 2018. "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature* 56(3): 920–980.

Concato, John, Nirav Shah and Ralph Horwitz, 2000. "Randomized Controllled Trials, Observational Studies, and the Hierarchy of Research Design," *New England Journal of Medicine* 342(25): 1887–92.

Cox, D.R., and N. Reid, 2000. *The Theory of the Design of Experiments*, Monographs on Statistics and Applied Probability 86, Chapman and Hall, New York.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté, 2015. "Estimating the Impact of Microcredit on Those Who Take It up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7(1): 123–50.

Deaton, Angus, 2010. "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature* 48(2): 424-455.

Deaton, Angus, and Nancy Cartwright, 2018. "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science and Medicine* 210: 2-21.

Donovan, Kevin, 2018. "The Rise of the Randomistas: On the Experimental Turn in International Aid," *Economy and Society* 47(1): 27-58.

Duflo, Esther, 2017. "The Economist as Plumber," *American Economic Review: Papers and Proceedings* 107(5): 1-26.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, 2015. "School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools," *Journal of Public Economics* 123: 92-110.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, 2011. "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics, Volume 4*, Amsterdam: North-Holland.

Favereau, Judith, 2016. "On the Analogy between Field Experiments in Economics and Clinical Trials in Medicine," *Journal of Economic Methodology* 23(2): 203-222.

Fiennes, Caroline, 2018. "Funders Start Assessing Their Own Performance. To Understand What a Charity is Achieving, You Must Understand What Good Research Looks Like." *Financial Times*, November 27.

Finkelstein, Amy, and Sarah Taubman, 2015. "Randomize Evaluations to Improve Health Care Delivery," *Science* 347(6223): 720-722.

Fiszbein, Ariel, and Norbert Schady, 2010. *Conditional Cash Transfers for Attacking Present and Future Poverty,* World Bank, Washington DC.

Food and Drug Administration, 2010. *Adaptive Design Clinical Trials for Drugs and Biologics*, Food and Drug Administration, US Government. Washington DC.

Franco, Annie, Neil Malhotra, and Gabor Simonovits, 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science* 345(6203): 1502-1505.

Freedman, Benjamin, 1987. “Equipoise and the Ethics of Clinical Research,” *The New England Journal of Medicine* 317(3):141–145.

Frieden, Thomas, 2017. “Evidence for Health Decision Making—Beyond Randomized Controlled Trials,” *New England Journal of Medicine* 377(5): 465-475.

Friedman, Jed, and Brindal Gokul, 2014. “Quantifying the Hawthorne Effect,” Development Impact Blog, World Bank.

Galasso, Emanuela, and Martin Ravallion, 2005. “Decentralized Targeting of an Anti-Poverty Program,” *Journal of Public Economics* 89(4): 705-727.

Galasso, Emanuela, Martin Ravallion, and Agustin Salvia, 2004. “Assisting the Transition from Workfare to Work: Argentina’s *Proempleo* Experiment,” *Industrial and Labor Relations Review* 57(5): 128-142.

Gertler, P. J., S. Martinez, P. Premand, L. Rawlings, and C.M.J. Vermeersch, 2016. *Impact Evaluation in Practice* (2nd edition), Washington, DC: Inter-American Development Bank and World Bank.

Gibson, John, 2019, “Are we Estimating the Right Thing? An Editor Reflects,” *Applied Economic Perspectives and Policy* 41(3): 329-350.

Glennerster, Rachel, and Shawn Powers, 2016. “Balancing Risk and Benefit. Ethical Tradeoffs in Running Randomized Evaluations.” *Oxford Handbook on Professional Economic Ethics*. Edited by George DeMartino and Deirdre McCloskey. Oxford: Oxford University Press.

Glynn, Adam, and Konstantin Kashin, 2018. “Front-door Versus Back-door Adjustment with Unmeasured Confounding: Bias Formulas for Front-door and Hybrid Adjustments with Application to a Job Training Program,” *Journal of the American Statistical Association*, forthcoming.

Goldberg, Jessica, 2014. “The R-Word Is Not Dirty,” Blog Post, Center for Global Development, Washington DC.

Greene, William, 1991. *Econometric Analysis.* New York : Macmillan.

Grossman, Jason, and Fiona Mackenzie, 2005. “The Randomized Controlled Trial: Gold Standard, or Merely Standard?” *Perspectives in Biology and Medicine* 48(4): 516-534.

Hahn, Jinyong, Petra Todd and Wilbert Van der Klaauw, 2001. “Identification and Estimation of Treatment Effects with a Regression-Discontinutiy Design,” *Econometrica* 69(1):

201-209.

Hammer, Jeffrey, 2017. "Randomized Control Trials for Development? Three Problems," Brookings Institution Blog Post, May 11.

Hannan, Edward, 2008. "Randomized Clinical Trials and Observational Studies: Guidelines for Assessing Respective Strengths and Limitations," *JACC: Cardiovascular Interventions* 2(3): 211-217.

Heckman James, 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*. Edited by C. Manski and I. Garfinkel. Cambridge, Mass.: Harvard University Press.

Heckman James and Rodrigo Pinto, 2019. "Exploiting Noncompliance to Enhance Causal Inference of Randomized Controlled Trials." Mimeo, University of Chicago.

Heckman James and Jeffrey Smith, 1995. "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9(2): 85-110.

Heckman, James, and Sergio Urzúa, 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics* 156: 27-37.

Heckman James, Sergio Urzua and Edward Vytlacil, 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics* 88(3): 389-432.

Heckman, James, and Edward Vytlacil, 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica* 73(3): 669-738.

_____________, and ______________, 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics Volume 6B*. Edited by James Heckman and Edward Leamer. Amsterdam: Elsevier.

Hernán Miguel A, and Jamie M. Robins, 2018. *Causal Inference.* Boca Raton: Chapman & Hall/CRC, forthcoming.

Hinkelmann, Klaus, and Oscar Kemthorne, 2008. *Design and Analysis of Experiments*, New York: John Wiley.

Imbens, Guido, 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature* 48(2): 399–423.

_____________, 2018. "Comments on Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Deaton and Cartwright," *Social Science and Medicine* 210: 50-52.

Ioannidis, John, 2005. "Why Most Published Research Findings are False," *PLoS Medicine* 2(8): 1-6.

Jamison, Dean, Barbara Searle, Klaus Galda, and Stephen P. Heyneman, 1981. "Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement," *Journal of Educational Psychology* 73(4): 556-567.

Jayachandran, Seema, Joost de Laat, Eric Lambin, Charlotta Stanton, Robin Audy, and Nancy Thomas, 2017, "Cash for Carbon: A Randomized Trial of Payments for Ecosystem Services to Reduce Deforestation," *Science* 357(6348): 267-273.

Johnston, Jack, 1984, *Econometric Methods*, New York: McGraw Hill.

Kapur, Davesh, 2018. "Academic Research on India in the US: For Whom does the Bell Toll?" *India in Transition*, Center for the Advanced Study of India, University of Pennsylvania, June 29.

Kasy, Maximilian, 2016. "Why Experimenters might not Always want to Randomize, and what they should do Instead," *Political Analysis* 24: 324-338.

Keane, Michael, 2010. "Structural vs. Atheoretic Approaches to Econometrics," *Journal of Econometrics* 156(1): 3-20.

Keating, Joshua, 2014. "Random Acts. What Happens when you Approach Global Poverty as a Science Experiment?" *Slate*, March 26.

Lancet, The, 2004. "The World Bank is Finally Embracing Science," *The Lancet* 364: 731-732.

Laura and John Arnold Foundation, 2018. "Request for Proposals: Randomized Controlled Trials to Evaluate Social Programs Whose Delivery Will Be Funded by Government or Other Entities," Laura and John Arnold Foundation.

Legovini, Arianna, Vincenzo Di Maro, and Caio Piza, 2015. "Impact Evaluation Helps Deliver Development Projects," Policy Research Working Paper 7157, World Bank.

Leigh, Andrew, 2018. *Randomistas. How Radical Researchers Changed our World.* New Haven: Yale University Press.

Levine, Robert, 2006. "Some Recent Developments in the International Guidelines on the Ethics of Research Involving Human Subjects," *Annals of the New York Academy of Science* 918: 170-178.

List, John A., and Imran Rasul, 2011. "Field Experiments in Labor Economics," in *Handbook of Labor Economics, Volume 4, Part A*, pp.103-228.

McKenzie, David, 2013. "How Should we Understand 'Clinical Equipoise' When Doing RCTs in Development?" Development Impact Blog, World Bank.

_______________, 2019. "Discussant's Comments." In *State of Economics, State of the World*. Edited by Kaushik Basu, David Rosenblatt and Claudia Paz Sepulveda. Cambridge, Mass.: MIT Press, forthcoming.

Meager, Rachael, 2018. "Understanding the Average Impact of Microcredit Expansion: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied* 11(1): 57-91.

Moffitt, Robert, 2004. "The Role of Randomized Field Trials in Social Science Research," *American Behavioral Scientist* 47(5): 506-540.

_____________, 2006. "Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective." In *Scale-Up in Education: Ideas in Principle*. Edited by Barbara Schneider and Sarah-Kathryn McDonald. Lanham, Maryland: Rowman and Littlefield.

Morgan, Kari Lock, and Donald B. Rubin, 2012. "Rerandomization To Improve Covariate Balance In Experiments," *Annals of Statistics* 40(2): 1263–1282.

Mulligan, Casey, 2014. "The Economics of Randomized Experiments," *Economix Blog, New York Times*, March 5.

Mummolo, Jonathan, and Erik Peterson, 2018. "Demand Effects in Survey Experiments: An Empirical Assessment," *American Political Science Review*, forthcoming.

Murgai, Rinku, Martin Ravallion and Dominique van de Walle, 2015. "Is Workfare Cost Effective against Poverty in a Poor Labor-Surplus Economy?" *World Bank Economic Review* 30(3): 413-445.

Narita, Yusuke, 2018. "Toward an Ethical Experiment," Cowles Foundation Discussion Paper No. 2127, Yale University.

Özler, Berk, 2018. "Incorporating Participants Welfare and Ethics into RCTs," Development Impact Blog Post, World Bank.

Pearl. Judea. 2009. *Causality: Models, Reasoning and Inference* (2nd edition). New York: Cambridge University Press.

Pearl, Judea, and Dana Mackenzie, 2018. *The Book of Why. The New Science of Cause and Effect.* New York: Basic Books.

Peters, Jörg, Jörg Langbein and Gareth Roberts, 2018. "Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity," *World Bank Research Observer* 33(1): 34–64.

Pritchett, Lant, 2020. "Randomizing Development: Method or Madness?" In *Randomized Control Trials in the Field of Development: A Critical Perspective*. Edited by Florent Bédécarrats, Isabelle Guérin and François Roubaud (eds). Oxford: Oxford University Press.

Pritchett, Lant, and Justin Sandefur, 2015. "Learning from Experiments when Context Matters," *American Economic Review: Papers and Proceedings* 105(5): 471–475.

Ravallion, Martin, 2009a. "Should the Randomistas Rule?" *Economists' Voice* 6(2): 1-5.

_______________, 2009b. "Evaluation in the Practice of Development," *World Bank Research Observer* 24(1): 29-54.

_______________, 2012. "Fighting Poverty one Experiment at a Time: A Review Essay on Abhijit Banerjee and Esther Duflo, *Poor Economics*," *Journal of Economic Literature* 50(1): 103-114.

_______________, 2014. "On the Implications of Essential Heterogeneity for Estimating Causal Impacts Using Social Experiments," *Journal of Econometric Methods* 4(1): 145-151.

_______________, 2016. *The Economics of Poverty: History, Measurement and Policy.* New York: Oxford University Press.

Ravallion, Martin, Dominique van de Walle, Puja Dutta, and Rinku Murgai, 2015. "Empowering Poor People through Public Information? Lessons from a Movie in Rural India," *Journal of Public Economics* 132(December): 13-22.

Rodrik, Dani, 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In *What Works in Development? Thinking Big and Thinking Small*. Edited by Jessica Cohen and William Easterl, Washington, D.C.: Brookings Institution Press).

Rosenbaum, Paul, and Donald Rubin, 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70: 41-55.

Rutter, Harry, Natalie Savona, Ketevan Glonti, Jo Bibby, Steven Cummins, Diane Finegood, Felix Greaves, Felix, Laura Harper, Penelope Hawe, Laurence Moore, Mark Petticrew, Eva Rehfuess, Alan Shiell, James Thomas, and Martin White, 2017. "The Need for a Complex Systems Model of Evidence for Public Health," *Lancet* 390(10112): 2602-04.

Sabet, Shayda Mae, and Annette Brown, 2018. "Is Impact Evaluation Still on the Rise? The New Trends 2010-2015," *Journal of Development Effectiveness* 10(3): 291-304.

Sathyamala, C. 2019. "In the name of science: Ethical violations in the ECHO randomised trial," *Global Public Health*, forthcoming.

Silverman, Stuart, 2009. "From Randomized Controlled Trials to Observational Studies," *American Journal of Medicine* 122(2): 114-120.

Skoufias, Emmanuel, and Susan Parker, 2001. "Conditional Cash Transfers and Their Impact on Child Work and Schooling: Evidence from the PROGRESA Program in Mexico," *Economía* 2(1): 45-86.

Tavernise, Sabrina, 2015. "Few Health System Studies use Top Method, Report Says," *New York Times* February 12.

Todd, Petra, and Kenneth Wolpin, 2006. "Assessing the Impact of a School Subsidy Program in Mexico using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling," *American Economic Review* 96(5): 1384-1417.

Vass, Mikkel, 2010. *Prevention of Functional Decline in Older People: The Danish Randomised Intervention Trial on Preventative Home Visits.* Doctoral Dissertation, Faculty of Health Science, University of Copenhagen, Copenhagen, Denmark.

Vivalt, Eva, 2017. "How Much Can We Generalize from Impact Evaluations?" Australian National University.

Webber, Sophie, and Carolyn Prouse, 2018. "The New Gold Standard: The Rise of Randomized Control Trials and Experimental Development," *Economic Geography* 94(2): 166-187.

White, Howard, 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.* Independent Evaluation Group, World Bank.

__________, 2016. *Transforming Development through Impact Evaluation*. I2i DIME Annual Report, World Bank, Washington DC.

Wydick, Bruce, 2018. "Review of Randomistas: How Radical Researchers Changed Our World," Development Impact Blog, World Bank.

Young, Alwyn, 2018. "Hanneling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *Quarterly Journal of Economics*, forthcoming.

Ziliak, Stephen T., 2014. "Balanced versus Randomized Field Experiments in Economics: Why W. S. Gosset aka "Student" Matters," *Review of Behavioral Economics* 1: 167–208.

Ziliak, Stephen T. and Edward R. Teather-Posadas, 2016. "The Unprincipled Randomization Principle in Economics and Medicine." In *Oxford Handbook on Professional Economic Ethics* Edited by George DeMartino and Deirdre McCloskey, Oxford: Oxford University Press.