Understanding and misunderstanding randomized controlled trials

Angus Deaton and Nancy Cartwright

Princeton University, NBER, and University of Southern California

Durham University and UC San Diego

We acknowledge helpful discussions with many people over the several years this paper has been in preparation. We would particularly like to note comments from seminar participants at Princeton, Columbia, and Chicago, the CHESS research group at Durham, as well as discussions with Orley Ashenfelter, Anne Case, Nick Cowen, Hank Farber, Jim Heckman, Bo Honoré, Chuck Manski, and Julian Reiss. Ulrich Mueller had a major influence on shaping Section 1. We have benefited from generous comments on an earlier version by Christopher Adams, Tim Besley, Chris Blattman, Sylvain Chassang, Jishnu Das, Jean Drèze, William Easterly, Jonathan Fuller, Lars Hansen, Jeff Hammer, Glenn Harrison, Macartan Humphreys, Michal Kolesár, Helen Milner, Tamlyn Munslow, Suresh Naidu, Lant Pritchett, Dani Rodrik, Burt Singer, Richard Williams, Richard Zeckhauser, and Steve Ziliak. We are also grateful for editorial assistance from Donal Khosrowski, Cheryl Lancaster, and Tamlyn Munslow. Cartwright's research for this paper has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 667526 K4U), the Spencer Foundation, and the National Science Foundation (award 1632471). Deaton acknowledges financial support from the National Institute on Aging through the National Bureau of Economic Research, Grants 5R01AG040629-02 and P01AG05842-14 and through Princeton University's Roybal Center, Grant P30 AG024928.

ABSTRACT

Randomized Controlled Trials (RCTs) are increasingly popular in the social sciences, not only in medicine. We argue that the lay public, and sometimes researchers, put too much trust in RCTs over other methods of investigation. Contrary to frequent claims in the applied literature, randomization does *not* equalize everything other than the treatment in the treatment and control groups, it does not automatically deliver a precise estimate of the average treatment effect (ATE), and it does not relieve us of the need to think about (observed or unobserved) covariates. Finding out whether an estimate was generated by chance is more difficult than commonly believed. At best, an RCT yields an unbiased estimate, but this property is of limited practical value. Even then, estimates apply only to the sample selected for the trial, often no more than a convenience sample, and justification is required to extend the results to other groups, including any population to which the trial sample belongs, or to any individual, including an individual in the trial. Demanding 'external validity' is unhelpful because it expects too much of an RCT while undervaluing its potential contribution. RCTs do indeed require minimal assumptions and can operate with little prior knowledge. This is an advantage when persuading distrustful audiences, but it is a disadvantage for cumulative scientific progress, where prior knowledge should be built upon, not discarded. RCTs can play a role in building scientific knowledge and useful predictions but they can only do so as part of a cumulative program, combining with other methods, including conceptual and theoretical development, to discover not 'what works', but 'why things work'.

Introduction

Randomized controlled trials (RCTs) are widely encouraged as the ideal methodology for causal inference. This has long been true in medicine (e.g. for drug trials by the FDA. A notable exception is the recent paper by Frieden (2017), ex-director of the U.S. Centers for Disease Control and Prevention, who lists key limitations of RCTs as well as a range of contexts where RCTs, even when feasible, are dominated by other methods. Earlier critiques in medicine include Feinstein and Horwitz (1997), Concato, Shah, and Horwitz (2000), Rawlins (2008), and Concato (2013).) It is also increasingly true in other health sciences and across the social sciences, including psychology, economics, education, political science, and sociology. Among both researchers and the general public, RCTs are perceived to yield causal inferences and estimates of average treatment effects (ATEs) that are more reliable and more credible than those from any other empirical method. They are taken to be largely exempt from the myriad problems that characterize observational studies, to require minimal substantive assumptions, little or no prior information, and to be largely independent of 'expert' knowledge that is often regarded as manipulable, politically biased, or otherwise suspect. They are also sometimes felt to be more resistant to researcher and publisher degrees of freedom (for example through *p*-hacking, selective analyses, or publication bias) than non-randomized studies given that trial registration and pre-specified analysis plans are mandatory or at least the norm.

We argue that any special status for RCTs is unwarranted. Which method is most likely to yield a good causal inference depends on what we are trying to discover as well as on what is already known. When little prior knowledge is available, no method is likely to yield well-supported conclusions. This paper is not a criticism of RCTs in and of themselves, nor does it propose any hierarchy of evidence, nor attempt to identify good and bad studies. Instead, we will argue that, depending on what we want to discover, why we want to discover it, and what we already know, there will often be superior routes of investigation and, for a great many questions

where RCTs can help, a great deal of other work—empirical, theoretical, and conceptual—needs to be done to make the results of an RCT serviceable.

Our arguments are intended not only for those who are innocent of the technicalities of causal inference but also aim to offer something to those who are well versed with the field. Most of what is in the paper is known to someone in some subject. But what epidemiology knows is not what is known by economics, or political science, or sociology, or philosophy—and the reverse. The literatures on RCTs in these areas are overlapping but often quite different; each uses its own language and different understandings and misunderstandings characterize different fields and different kinds of projects. We highlight issues arising across a range of disciplines where we have observed misunderstanding among serious researchers and research users, even if not shared by all experts in those fields. Although we aim for a broad cross-disciplinary perspective, we will, given our own disciplinary backgrounds, be most at home with how these issues arise in economics and how they have been treated by philosophers.

We present two sets of arguments. The first is an enquiry into the idea that ATEs estimated from RCTS are likely to be closer to the truth than those estimated in other ways. The second explores how to use the results of RCTs once we have them.

In the first section, our discussion runs in familiar statistical terms of bias and precision, or efficiency, or expected loss. Unbiasedness means being right on average, where the average is taken over an infinite number of repetitions using the same set of subjects in the trial, but with no limits on how far any one estimate is from the truth, while precision means being close to the truth on average; an estimator that is far from the truth in one direction half of the time and equally far from the truth in the other direction half of the time is unbiased, but it is imprecise. We review the difference between balance of covariates in expectation versus balance in a single run of the experiment (sometimes called 'random confounding' or `realized confounding' in epidemiology, see for instance Greenland and Mansournia (2015) or VanderWeele (2012)) and the related distinction between precision and unbiasedness. These distinctions should be well known wherever RCTs are conducted or RCT

results are used, though much of the discussion is, if not confused, unhelpfully imprecise. Even less recognized are problems with statistical inference, and especially the threat to significance testing posed when there is an asymmetric distribution of individual treatment effects in the study population.

The second section describes several different ways to use the evidence from RCTs. The types of use we identify have analogues, with different labels, across disciplines. This section stresses the importance for using RCT results of being clear about the hypothesis at stake and the purpose of the investigation. It argues that in the usual literature, which stresses extrapolation and generalization, RCTs are both under- and over-sold. Oversold because extrapolating or generalizing RCT results requires a great deal of additional information that cannot come from RCTs; undersold, because RCTs can serve many more purposes than predicting that results obtained in a trial population will hold elsewhere.

One might be tempted to label the two sections 'Internal validity' and 'External validity'. We resist this, especially in the way that external validity is often characterized. RCTs are under-sold when external validity means that the 'the same ATE holds in this new setting', or 'the ATE from the trial holds generally', or even that the ATE in a new setting can be calculated in some reasonable way from that in the study population. RCT results can be useful much more broadly. RCTs are oversold when their non-parametric and theory-free nature, which is arguably an advantage in estimation or internal validity, is used as an argument for their usefulness. The lack of structure is often a disadvantage when we try to use the results outside of the context in which the results were obtained; credibility in estimation can lead to incredibility in use. You cannot know how to use trial results without first understanding how the results from RCTs relate to the knowledge that you already possess about the world, and much of this knowledge is obtained by other methods. Once RCTs are located within this broader structure of knowledge and inference, and when they are designed to enhance it, they can be enormously useful, not just for warranting claims of effectiveness but for scientific progress more generally. Cumulative science is difficult; so too is reliable prediction about what will happen when we act.

Nothing we say in the paper should be taken as a general argument against RCTs; we simply try to challenge unjustifiable claims and expose misunderstandings. We are not against RCTs, only magical thinking about them. The misunderstandings are important because they contribute to the common perception that RCTs always provide the strongest evidence for causality and for effectiveness and because they detract from the usefulness of RCT evidence as part of more general scientific projects. In particular, we do not try to rank RCTs versus other methods. What methods are best to use and in what combinations depends on the exact question at stake, the kind of background assumptions that can be acceptably employed, and what the costs are of different kinds of mistakes. By getting clear in Section 1 just what an RCT, qua RCT, can and cannot deliver, and laying out in Section 2 a variety of ways in which the information secured in an RCT can be used, we hope to expose how unavailing is the 'head-to-head between methods' discourse that often surrounds evidence-ranking schemes.

Section 1: Do RCTs give good estimates of Average Treatment Effects

We start from a *trial sample*, a collection of subjects that will be allocated randomly to either the treatment or control arm of the trial. This 'sample' might be, but rarely is, a random sample from some population of interest. More frequently, it is selected in some way, for example to those willing to participate, or is simply a convenience sample that is available to the those conducting the trial. Given random allocation to treatments and controls, the data from the trial allow the identification of the two (marginal) distributions, $F_1(Y_1)$ and $F_0(Y_0)$, of outcomes Y_1 and Y_0 in the treated and untreated cases within the trial sample. The ATE estimate is the difference in means of the two distributions and is the focus of much of the literature in social science and medicine.

Policy makers and researchers may be interested in features of the two marginal distributions and not simply the ATE, which is our main focus here. For example, if Y is disease burden, measured perhaps in QALYs, public health officials may be interested in whether a treatment reduced inequality in disease burden, or in what it did to the 10^{th} or 90^{th} percentiles of the distribution, even though different

people occupy those percentiles in the treatment and control distributions. Economists are routinely concerned with the 90/10 ratio in the income distribution, and in how a policy might affect it (see Bitler et al. (2006) for a related example in US welfare policy). Cancer trials standardly use the median difference in survival, which compares the times until half the patients have died in each arm. More comprehensively, policy makers may wish to compare expected utilities for treated and untreated under the two distributions and consider optimal expected-utility maximizing treatment rules conditional on the characteristics of subjects (see Manski (2004) and Manski and Tetenov (2016); Bhattacharya and Dupas (2012) give an application.) These other kinds of information are important, but we focus on ATEs and do not consider these other uses of RCTs further in this paper.

1.1 *Estimating average treatment effects*

A useful way to think about the estimation of treatment effects is to use a schematic linear causal model of the form:

$$Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij} \qquad (1)$$

where, Y_i is the outcome for unit *i* (which may be a person, a village, a hospital ward), T_i is a dichotomous (1,0) treatment dummy indicating whether or not *i* is treated, and β_i is the individual treatment effect of the treatment on *i*: it represents (or regulates) how much a value *t* of *T* contributes to the outcome *Y* for individual *i*. The *x*'s are observed or unobserved other linear causes of the outcome, and we suppose that (1) captures a minimal set of causes of Y_i sufficient to fix its value. *J* may be (very) large. The unrestricted heterogeneity of the individual treatment effects, β_i, allows the possibility that the treatment interacts with the *x*'s or other variables, so that the effects of *T* can depend on (be modified by) any other variables. Note that we do not need *i* subscripts on the γ's that control the effects of the other causes; if their effects differ across individuals, we include the interactions of individual characteristics with the original *x*'s as new *x*'s. Given that the *x*'s can be unobservable, this is not restrictive. Usage here differs across fields; we shall typically refer to factors other than *T* represented on the right-hand side of (1) by the term *covariates*,

while noting that these include both what are sometimes labelled the 'independently operating causes' (represented by the *x*'s) as well as 'effect modifiers' when they interact with the $\beta's$, a case we shall return to below. They may also capture the possibility that there are different baselines for different observations.

We can connect (1) with the counterfactual approach, often referred to as the Rubin Causal Model, now common in epidemiology and increasingly so in economics (see Rubin (2005), or Hernán (2004) for an exposition for epidemiologists, and Freedman (2006) for the history). To illustrate, suppose that *T* is dichotomous. For each unit *i* there will be two possible outcomes, typically labelled Y_{i0} and Y_{i1}, the former occurring if there is no treatment at the time in question, the latter if the unit is treated. By inspection of (1), the differences between the two outcomes, $Y_{i1} - Y_{i0}$, are the individual treatment effects, β_i, which are typically different for different units. No unit can be both treated and untreated at the same time, so only one or other of the outcomes occurs, but not both—the other is counterfactual so that individual treatment effects are in principle unobservable.

The basic theorem from this setup is a remarkable one. It states that the average treatment effect is the difference between the average outcome in the treatment group minus the average outcome in the control group so that, while we cannot observe the *individual* treatment effects, we can observe their mean. The estimate of the average treatment effect is simply the difference between the means in the two groups, and it has a standard error that can be estimated using the statistical theory that applies to the difference of two means, on which more below. The difference in means is an *unbiased* estimator of the mean treatment effect. The theorem is remarkable because it requires so few assumptions, although it relies on the fact that the mean is a linear operator, so that the difference in means is the mean of differences. No similar fact is true for other statistics, such as medians, percentiles, or variances of treatment effects, none of which can be identified from an RCT without substantive further assumptions, see Deaton (2010, 439) for a simple exposition. Otherwise, no model is required, no assumptions about covariates, confounders, or other causes are needed, the treatment effects can be heterogeneous, and

nothing is required about the shapes of statistical distributions other than the existence of the counterfactual outcome values.

Dawid (2000) argues that the existence of counterfactuals is a metaphysical assumption that cannot be confirmed (or refuted) by any empirical evidence and is controversial because, under some circumstances, there is an unresolvable arbitrariness to causal inference, something that is not true of (1), for example. See also the arguments by the empiricist philosopher, Reichenbach (1954), reissued as Reichenbach (1976).) In economics, the case for the counterfactual approach is eloquently made by Imbens and Wooldridge (2009, Introduction), who emphasize the benefits of a theory-free specification with almost unlimited heterogeneity in treatment effects. Heckman and Vytlacil (2007, Introduction) are equally eloquent on the drawbacks, noting that the counterfactual approach often leaves us in the dark about the exact nature of the treatment, so that the treatment effects can be difficult to link to invariant quantities that would be useful elsewhere (invariant in the sense of Hurwicz (1966)).

Consider an experiment that aims to tell us something about the treatment effects; this might or might not use randomization. Either way, we can represent the treatment group as having $T_i = 1$ and the control group as having $T_i = 0$. Given the study (or trial) sample, subtracting the average outcomes among the controls from the average outcomes among the treatments, we get

$$\bar{Y}_1 - \bar{Y}_0 = \bar{\beta}_1 + \sum_{j=1}^{J} \gamma_j (\bar{x}_{1ij} - \bar{x}_{0ij}) = \bar{\beta}_1 + (\bar{S}_1 - \bar{S}_0) \qquad (2)$$

The first term on the far-right-hand side of (2), which is the ATE in the treated population in the trial sample, is generally the quantity of interest in choosing to conduct an RCT, but the second term or error term, which is the sum of the net average balance of other causes across the two groups, will generally be non-zero and needs to be dealt with somehow. We get what we want when the means of all the other causes are identical in the two groups, or more precisely (and less onerously) when the sum of their net differences $\bar{S}_1 - \bar{S}_0$ is zero; this is the case of *perfect balance*. With perfect balance, the difference between the two means is *exactly* equal to the aver-

age of the treatment effects among the treated, so that we have the ultimate precision in that we know the truth in the trial sample, at least in this linear case. As always, the 'truth' here refers to the *trial sample,* and it is always important to be aware that the trial sample may not be representative of the population that is ultimately of interest, including the population from which the trial sample comes; any such extension requires further argument.

How do we get balance, or something close to it? In a laboratory experiment, where there is usually much prior knowledge of the other causes, the experimenter has a good chance of controlling (or subtracting away the effects of) the other causes, aiming to ensure that the last term in (1) is close to zero. Failing such knowledge and control, an alternative is *matching*, which is frequently used in non-randomized statistical, medical (case-control studies), and econometric studies, (see Heckman et al. (1997)). For each subject, a matched subject is found that is as close as possible on all suspected causes, so that, once again, the last term in (1) can be kept small. When we have a good idea of the causes, matching may also deliver a precise estimate. Of course, when there are unknown or unobservable causes that have important effects, neither laboratory control nor matching offers protection.

What does randomization do? Suppose that no correlations of the x's with Y are introduced post-randomization, for example by subjects not accepting their assignment, or by treatment protocols differing from those used for controls. With this assumption, randomization provides *orthogonality* of the treatment to the other causes represented in equation (1): Since the treatments and controls come from the same underlying distribution, randomization guarantees, by construction, that the last term on the right in (1) is zero *in expectation*. The expectation is taken over repeated randomizations on the trial sample, each with its own allocation of treatments and controls. Assuming that our caveat holds, the last term in (2) will be zero when averaged over this infinite number of (entirely hypothetical) replications, and the average of the estimated ATEs will be the true ATE in the trial sample. So $\bar{\beta}_1$ is an unbiased estimate of the ATE among the treated in the trial sample, and this is so whether or not the causes are observed. Unbiasedness does not require us to know anything about covariates, confounders, or other causes, though it does require that

they not change after randomization so as to make them correlated with the treatment, an important caveat to which we shall return.

In any one trial, the difference in means is the average treatment effect among those treated *plus* the term that reflects the randomly generated imbalance in the net effects of the other causes. We do not know the size of this error term, and there is nothing in randomization that limits its size though, as we discuss below, it will tend to be smaller in larger samples. In any single trial, the chance of randomization can over-represent an important excluded cause(s) in one arm over the other, in which case there will be a difference between the means of the two groups that is *not* caused by the treatment. In epidemiology, this is sometimes referred to as 'random confounding', or 'realized confounding', a phenomenon that was recognized by Fisher in his agricultural trials. (An instructive example of perfect random confounding is constructed by Greenland (1990).)

If we were to repeat the trial many times, the over-representation of the unbalanced causes will sometimes be in the treatments and sometimes in the controls. The imbalance will vary over replications of the trial, and although we cannot see this from our single trial, we should be able to capture its effects on our estimate of the ATE from an estimated standard error. This was Fisher's insight: not that randomization balanced covariates between treatments and controls but that, conditional on the caveat that no post-randomization correlation with covariates occurs, randomization provides the basis for calculating the size of the error. Getting the standard error and associated significance statements right are of the greatest importance; therein lies the virtue of randomization, not that it yields precise estimates through balance.

1.2 *Misunderstandings: claiming too much*

Exactly what randomization does is frequently lost in the practical and popular literature. There is often confusion between perfect control, on the one hand (as in a laboratory experiment or perfect matching with no unobservable causes), and control in expectation on the other, which is what randomization contributes. If we knew enough about the problem to be able to control well, that is what we would

(and should) do. Randomization is an alternative when we do not know enough to control, but is generally inferior to good control when we do. We suspect that at least some of the popular and professional enthusiasm for RCTs, as well as the belief that they are precise by construction, comes from misunderstandings about balance or, in epidemiological language, about random or realized confounding on the one hand and confounding in expectation on the other. These misunderstandings are not so much among the researchers who will usually give a correct account when pressed. They come from imprecise statements by researchers that are taken literally by the lay audience that the researchers are keen to reach, and increasingly successfully.

Such a misunderstanding is well captured by a quote from the second edition of the online manual on impact evaluation jointly issued by the Inter-American Development Bank and the World Bank (the first, 2011 edition is similar):

> We can be confident that our estimated impact constitutes the true impact of the program, since we have eliminated all observed and unobserved factors that might otherwise plausibly explain the difference in outcomes. Gertler et al. (2016, 69)

This statement is false, because it confuses *actual* balance in any single trial with balance in expectation over many (hypothetical) trials. If it were true, and if *all* factors were indeed controlled (and no imbalances were introduced post randomization), the difference would be an exact measure of the average treatment effect among the treated in the trial population (at least in the absence of measurement error). We should not only be confident of our estimate but, as the quote says, we would know that it is the truth. Note that the statement contains no reference to sample size; we get the truth by virtue of balance, not from a large number of observations.

There are many similar quotes in the economics literature. From the medical literature, here is one from a distinguished psychiatrist who is deeply skeptical of the use of evidence from RCTs:

> The beauty of a randomized trial is that the researcher does not need to understand all the factors that influence outcomes. Say that an undiscovered

> genetic variation makes certain people unresponsive to medication. The randomizing process will ensure—or make it highly probable—that the arms of the trial contain equal numbers of subjects with that variation. The result will be a fair test. Kramer (2016,18)

Claims are made that RCTs reveal knowledge without possibility of error. Judy Gueron, the long-time president of MDRC (originally known as the Manpower Development Research Corporation), which has been running RCTs on US government policy for 45 years, asks why federal and state officials were prepared to support randomization in spite of frequent difficulties and in spite of the availability of other methods and concludes that it was because "they wanted to learn the truth," Gueron and Rolston (2013, 429). There are many statements of the form "We *know* that [project X] worked because it was evaluated with a randomized trial," Dynarski (2015).

It is common to treat the ATE from an RCT as if it were the truth, not just in the trial sample but more generally. In economics, a famous example is Lalonde's (1986) study of labor market training programs, whose results were at odds with a number of previous non-randomized studies. The paper prompted a large-scale re-examination of the observational studies to try to bring them into line, though it now seems just as likely that the differences lie in the fact that the different study results apply to different populations (Heckman et al. (1999)). With heterogeneous treatment effects, the ATE is only as good as the study sample from which it was obtained. (See Longford and Nelder (1999) who are concerned with the same issue in regulating pharmaceuticals. (We return to this in discussing support factors and moderator variables in Section 2.2) In epidemiology, Davey-Smith and Ibrahim (2002) state that "observational studies propose, RCTs dispose." Another good example is the RCT of hormone replacement therapy (HRT) for post-menopausal women. HRT had previously been supported by positive results from a high-quality and long-running observational study, but the RCT was stopped in the face of excess deaths in the treatment group. The negative result of the RCT led to widespread abandonment of the therapy, which might (or might not) have been a mistake (see Vandenbroucke (2009) and Frieden (2017)). Yet the medical and popular literature

routinely states that the RCT was right and the earlier study wrong, simply *because* the earlier study was not randomized. The gold standard or `truth' view does harm when it undermines the obligation of science to reconcile RCTs results with other evidence in a process of cumulative understanding.

The false belief in automatic precision suggests that we need pay no attention to the other causes in (1) or (2). Indeed, Gerber and Green (2012, 5), in their standard text for RCTs in political science, note that RCTs are the successful resolution of investigators' need for "a research strategy that does not require them to identify, let alone measure, all potential confounders." But the RCT strategy is only successful if we are happy with estimates that are arbitrarily far from the truth, just so long as the errors cancel out over a series of imaginary experiments. In reality, the causality that is being attributed to the treatment might, in fact, be coming from an imbalance in some other cause in our particular trial; limiting this requires serious thought about possible covariates.

1.3 *Sample size, balance, and precision*

The literature on the precision of ATEs estimated from RCTs goes back to the very beginning. Gosset (writing as 'Student') never accepted Fisher's arguments for randomization in agricultural field trials and argued convincingly that his own non-random designs for the placement of treatment and controls yielded more precise estimates of treatment effects (see Student (1938) and Ziliak (2014)). Gosset worked for Guinness where inefficiency meant lost revenue, so he had reasons to care, as should we. Fisher won the argument in the end, not because Gosset was wrong about efficiency, but because, unlike Gosset's procedures, randomization provides a sound basis for statistical inference, and thus for judging whether an estimated ATE is different from zero by chance. Moreover, Fisher's blocking procedures can limit the inefficiency from randomization (see Yates (1939)). Gosset's reservations were echoed much later in Savage's (1962) comment that a Bayesian should not choose the allocation of treatments and controls at random but in such a way that, given what else is known about the topic and the subjects, their placement reveals the most to the researcher. We return to this below.

At the time of randomization and in the absence of post-randomization changes in other causes, a trial is more likely to be balanced when the sample size is large. As the sample size tends to infinity, the means of the *x*'s in the treatment and control groups will become arbitrarily close. Yet this is of little help in finite samples. As Fisher (1926) noted: "Most experimenters on carrying out a random assignment will be shocked to find how far from equally the plots distribute themselves," quoted in Morgan and Rubin (2012, 1263). Even with very large sample sizes, if there is a large number of causes, balance on *each* cause may be infeasible. Vandenbroucke (2004) notes that there are three billion base pairs in the human genome, many or all of which could be relevant prognostic factors for the biological outcome that we are seeking to influence. It is true, as (2) makes clear, that we do not need balance on each cause individually, only on their net effect, the term $\overline{S^1} - \overline{S^0}$. But consider the human genome base pairs. Out of all those billions, only one might be important, and if that one is unbalanced, the results of a single trial can be 'randomly confounded' and far from the truth. Statements about large samples guaranteeing balance are not useful without guidelines about how large is large enough, and such statements cannot be made without knowledge of other causes and how they affect outcomes. Of course, lack of balance in the net effect of either observables or non-observables in (2) does not compromise the inference in an RCT in the sense of obtaining a standard error for the unbiased ATE (see Senn (2013) for a particularly clear statement), although it does clarify the importance of having credible standard errors, on which more below.

Having run an RCT, it makes good sense to examine any available covariates for balance between the treatments and controls; if we suspect that an observed variable *x* is a possible cause, and its means in the two groups are very different, we should treat our results with appropriate suspicion. In practice, researchers often carry out a statistical test for balance after randomization but before analysis, presumably with the aim of taking some appropriate action if balance fails. The first table of the paper typically presents the sample means of observable covariates for the control and treatment groups, together with their differences, and tests for

whether or not they are significantly different from zero, either variable by variable, or jointly. These tests are appropriate for *unbiasedness* if we are concerned that the random number generator might have failed, or if we are worried that the randomization is undermined by non-blinded subjects who systematically undermine the allocation. Otherwise, supposing that no post-randomization correlations are introduced, unbiasedness is guaranteed by the randomization, whatever the test shows, and the test is not informative about the balance that would lead to *precision*; Begg (1990, 223) notes, "(I)t is a test of a null hypothesis that is known to be true. Therefore, if the test turns out to be significant it is, by definition, a false positive." The Consort 2010 updated statement, guideline 15 notes "Unfortunately significance tests of baseline differences are still common; they were reported in half of 50 RCTs trials published in leading general journals in 1997." We have not systematically examined the practice across other social sciences, but it is standard in economics, even in high-quality studies in leading journals, such as Banerjee et al. (2015), published in *Science*.

Of course, it is always good practice to look for imbalances between *observed* covariates in any single trial using some more appropriate distance measure, for example the normalized difference in means (Imbens and Wooldridge (2009, equation (3)). Similarly, it would have been good practice for Fisher to abandon a randomization in which there were clear patterns in the (random) distribution of plots across the field, even though the treatment and control plots were randomly selections that, by construction, could not differ 'significantly' using the standard (incorrect) balance test. Whether such imbalances should be seen as undermining the estimate of the ATE depends on our priors about which covariates are likely to be important, and *how* important, which is (not coincidentally) the same thought experiment that is routinely undertaken in observational studies when we worry about confounding.

One procedure to improve balance is to adapt the design *before* randomization, for example, by stratification. Fisher, who as the quote above illustrates, was well aware of the loss of precision from randomization argued for 'blocking' (stratification) in agricultural trials or for using Latin Squares, both of which restrict the amount of imbalance. Stratification, to be useful, requires some prior understanding

of the factors that are likely to be important, and so it takes us away from the 'no knowledge required' or 'no priors accepted' appeal of RCTs; it requires thinking about and measuring confounders. But as Scriven (1974, 69) notes: "(C)ause hunting, like lion hunting, is only likely to be successful if we have a considerable amount of relevant background knowledge." Cartwright (1994, Chapter 2) puts it even more strongly, "No causes in, no causes out." Stratification in RCTs, as in other forms of sampling, is a standard method for using background knowledge to increase the precision of an estimator. It has the further advantage that it allows for the exploration of different ATEs in different strata which can be useful in adapting or transporting the results to other locations (see Section 2).

Stratification is not possible if there are too many covariates, or if each has many values, so that there are more cells than can be filled given the sample size. With five covariates, and ten values on each, and no priors to limit the structure, we would have 100,000 possible strata. Filling these is well beyond the sample sizes in most trials. An alternative that works more generally is to *re-randomize*. If the randomization gives an obvious imbalance on known covariates—treatment plots all on one side of the field, all the treatment clinics in one region, too many rich and too few poor in the control group—we try again, and keep trying until we get a balance measured as a small enough distance between the means of the observed covariates in the two groups. Morgan and Rubin (2012) suggest the Mahalanobis *D*–statistic be used as a criterion and use Fisher's randomization inference (to be discussed further below) to calculate standard errors that take the re-randomization into account. An alternative, widely adapted in practice, is to adjust for covariates by running a regression (or covariance) analysis, with the outcome on the left-hand side and the treatment dummy and the covariates as explanatory variables, including possible interactions between covariates and treatment dummies. Freedman (2008) shows that the adjusted estimate of the ATE is biased in finite samples, with the bias depending on the correlation between the squared treatment effect and the covariates. Accepting some bias in exchange for greater precision will often make sense, though it certainly undermines any gold standard argument that relies on unbiasedness without consideration of precision.

1.4 *Should we randomize?*

The tension between randomization and precision that goes back to Fisher, Gosset, and Savage has been reopened in recent papers by Kasy (2016), Banerjee et al. (BCS) (2016) and Banerjee et al. (BCMS) (2017).

The trade-off between bias and precision can be formalized in several ways, for example by specifying a loss or utility function that depends on how a user is affected by deviations of the estimate of the ATE from the truth and then choosing an estimator or an experimental design that minimizes expected loss or maximizes expected utility. As Savage (1962, 34) noted, for a Bayesian, this involves allocating treatments and controls in "the specific layout that promised to tell him the most," but *without randomization*. Of course, this requires serious and perhaps difficult thought about the mechanisms underlying the ATE, which randomization avoids. Savage also notes that several people with different priors may be involved in an investigation and that individual priors may be unreliable because of "vagueness and temptation to self-deception," defects that randomization may alleviate, or at least evade. BCMS (2017) provide a proof of a Bayesian no-randomization theorem, and BCS (2016) provide an illustration of a school administrator who has long believed that school outcomes are determined, not by school quality, but by parental background, and who can learn the most by placing deprived children in (supposed) high-quality schools and privileged children in (supposed) low-quality schools, which is the kind of study setting to which case study methodology is well attuned. As BCS note, this allocation would not persuade those with different priors, and they propose randomization as a means of satisfying skeptical observers. As this example shows, it is not always necessary to encode prior information into a set of formal prior probabilities, though thought about what we are trying to learn is always required.

Several points are important. First, the anti-randomization theorem is *not* a justification of *any* non-randomized design, for example, one that allows selection on unobservables, but only of the optimal design that is most informative. According to Chalmers (2001) and Bothwell and Podolsky (2016), the development of random-

ization in medicine originated with Bradford-Hill, who used randomization in the first RCT in medicine—the streptomycin trial—because it prevented doctors selecting patients on the basis of perceived need (or against perceived need, leaning over backward as it were), an argument recently echoed by Worrall (2007). Randomization serves this purpose, but so do other non-discretionary schemes; what is required is that hidden information should not be allowed to affect the allocation as would happen, for example, if subjects could choose their own assignments.

Second, the ideal rules by which units are allocated to treatment or control depend on the covariates and on the investigators' priors about how they affect the outcomes. This opens up all sorts of methods of inference that are long familiar but that are excluded by pure randomization. For example, what philosophers call the hypothetico-deductive method works by using theory to make a prediction that can be taken to the data for potential falsification (as in the school example above). This is the way that physicists learn, as do other researchers when they use theory to derive predictions that can be tested against the data, perhaps in an RCT, but more frequently not. As Lakatos 1970 (among others) has stressed, some of the most fruitful research advances are generated by the puzzles that result when the data fail to match such theoretical predictions. In economics, good examples include the equity premium puzzle, various purchasing power parity puzzles, the Feldstein-Horioka puzzle, the consumption smoothness puzzle, the puzzle of why in India, where malnourishment is widespread, rapid income growth has been accompanied by s fall in calories consumed, and many others.

Third, randomization, by ignoring prior information from theory and from covariates, is wasteful and even unethical when it unnecessarily exposes people, or unnecessarily many people, to possible harm in a risky experiment. Worrall (2008) documents the (extreme) case of ECMO (Extracorporeal Membrane Oxygenation), a new treatment for newborns with persistent pulmonary hypertension that was developed in the 1970s by intelligent and directed trial and error within a well-understood theory of the disease and a good understanding of how the oxygenator should work. In early experimentation by the inventors, mortality was reduced from 80 to 20 percent. The investigators felt compelled to conduct an RCT, albeit with an

adaptive 'play-the-winner' design in which each success in an arm increased the probability of the next baby being assigned to that arm. One baby received conventional therapy and died, 11 received ECMO and lived. Even so, a standard randomized controlled trial was thought necessary. With a stopping rule of four deaths, four more babies (out of ten) died in the control group and none of the nine who received ECMO.

Fourth, the non-random methods use prior information, which is why they do better than randomization. This is both an advantage and a disadvantage, depending on one's perspective. If prior information is not widely accepted, or is seen as non-credible by those we are seeking to persuade, we will generate more credible estimates if we do not use those priors. Indeed, this is why BCS (2017) recommend randomized designs, including in medicine and in development economics. They develop a theory of an investigator who is facing an adversarial audience who will challenge any prior information and can even potentially veto results based on it (think of administrative agencies such as the FDA or journal referees). The experimenter trades off his or her own desire for precision (and preventing possible harm to subjects), which would require prior information, against the wishes of the audience, who wants nothing to do with those priors. Even then, the approval of the audience is only *ex ante*; once the fully randomized experiment has been done, nothing stops critics arguing that, in fact, the randomization did not offer a fair test because important other causes were not balanced. Among doctors who use RCTs, and especially meta-analysis, such arguments are (appropriately) common (see Kramer (2016)). We return to this topic in Section 2.1.

Today, when the public has come to question expert prior knowledge, RCTs will flourish. In cases where there is good reason to doubt the good faith of experimenters, randomization will indeed be an appropriate response. But we believe such a simplistic approach is destructive for scientific endeavor (which is not the purpose of the FDA) and should be resisted as a general prescription in scientific research. Previous knowledge needs to be built on and incorporated into new knowledge, not discarded. The systematic refusal to use prior knowledge and the associated preference for RCTs are recipes for preventing cumulative scientific pro-

gress. In the end, it is also self-defeating. To quote Rodrik (D. Rodrik, personal communication, April 6, 2016) "the promise of RCTs as theory-free learning machines is a false one."

1.5 *Statistical inference in RCTs*

The estimated ATE in a simple RCT is the difference in the means between the treatment and control groups. When covariates are allowed for, as in most RCTs in economics, the ATE is usually estimated from the coefficient on the treatment dummy in a regression that looks like (1), but with the heterogeneity in β ignored. Modern work calculates standard errors allowing for the possibility that residual variances may be different in the treatment and control groups, usually by clustering the standard errors, which is equivalent to the familiar two sample standard error in the case with no covariates. Statistical inference is done with *t*-values in the usual way. Unfortunately, these procedures do not always give the right standard errors and, to reiterate, the value of randomization is that it permits inference about estimates of ATEs, not that it guarantees the quality of these estimates, so credible standard errors are essential in any argument for RCTs.

Looking back at (1), the underlying objects of interest are the individual treatment effects β_i for each of the individuals in the trial sample. Neither they, nor their distribution $G(\beta)$ is identified from an RCT; because RCTs make so few assumptions which, in many cases, is their strength, they can identify only the mean of the distribution. In many observational studies, researchers are prepared to make more assumptions on functional forms or on distributions, and for that price we are able to identify other quantities of interest. Without these assumptions, inferences must be based on the difference in the two means, a statistic that is sometimes ill-behaved, as we discuss below. This ill-behavior has nothing to do with RCTs, per se, but within RCTs, and their minimal assumptions, we cannot easily switch from the mean to some other quantity of interest.

Fisher proposed that statistical inference should be done using what has become known as 'randomization inference', a procedure that is as non-parametric as the RCT-based estimate of an ATE itself. To test the null hypothesis that $\beta_i = 0$ for

all *i,* note that, under the null that the treatment has no effect on *any* individual, an estimated nonzero ATE can only be a consequence of the particular random allocation that generated it (assuming no difference in the distributions of covariates post-randomization). By tabulating all possible combinations of treatments and controls in our trial sample, and the ATE associated with each, we can calculate the exact distribution of the estimated ATE under the null. This allows us to calculate the probability of calculating an estimate as large as our actual estimate when the treatment has no effect. This randomization test requires a finite sample, but it will work for any sample size (see Imbens and Wooldridge (2009) for an excellent account of the procedure).

Randomization inference can be used to test the null hypotheses that *all* of the treatment effects are zero, as in the above example, but it cannot be used to test the hypothesis that the *average* treatment effect is zero, which will often be of interest. In agricultural trials, and in medicine, the stronger (sharp) hypothesis that the treatment has no effect whatever is often of interest. In many public health applications, we are content with improving average health, and in economic applications that involve money, such as welfare experiments or cost-benefit analyses, we are interested in whether the net effect of the treatment is positive or negative, and in these cases, randomization inference cannot be used. None of which argues against its wider use in social sciences when appropriate.

In cases where randomization inference cannot be used, we must construct tests for the differences in two means. Standard procedures will often work well, but there are two potential pitfalls. One, the 'Fisher-Behrens problem', comes from the fact that, when the two samples have different variances—which we typically want to permit—the t–statistic as usually calculated does not have the t-distribution. The second problem, which is much harder to address, occurs when the distribution of treatment effects is not symmetric (Bahadur and Savage (1956)). Neither pitfall is specific to RCTs, but RCTs force us to work with means in estimating treatment effects and, with only a few exceptions in the literature, social scientists who use RCTs appear to be unaware of the difficulties.

In the simple case of comparing two means in an RCT, inference is usually based on the two–sample *t*–statistic which is computed by dividing the ATE by the estimated standard error whose square is given by

$$\hat{\sigma}^2 = \frac{(n_1 - 1)^{-1} \sum_{i\epsilon 1} (Y_i - \bar{Y}_1)^2}{n_1} + \frac{(n_0 - 1)^{-1} \sum_{i\epsilon 0} (Y_i - \bar{Y}_0)^2}{n_0} \#(3)$$

where 0 refers to controls and 1 to treatments, so that there are n_1 treatments and n_0 controls, and $\bar{Y}_1$ and $\bar{Y}_0$ are the two means. As has long been known, the "*t*–statistic' based on (3) is not distributed as Student's *t* if the two variances (treatment and control) are not identical but has the Behrens–Fisher distribution. In extreme cases, when one of the variances is zero, the *t*–statistic has *effective* degrees of freedom half of that of the nominal degrees of freedom, so that the test-statistic has thicker tails than allowed for, and there will be too many rejections when the null is true.

Young (2017) argues that this problem is worse when the trial results are analyzed by regressing outcomes not only on the treatment dummy but also on additional covariates and when using clustered or robust standard errors. When the design matrix is such that the maximal influence is large, which is likely if the distribution of the covariates is skewed so that for some observations outcomes have large influence on their own predicted values, there is a reduction in the effective degrees of freedom for the *t*–value(s) of the average treatment effect(s) leading to spurious findings of significance. Young looks at 2,027 regressions reported in 53 RCT papers in the *American Economic Association* journals and recalculates the significance of the estimates using randomization inference applied to the authors' original data. In 30 to 40 percent of the estimated treatment effects in individual equations with coefficients that are reported as significant, he cannot reject the null of no effect for any observation; the fraction of spuriously significant results increases further when he simultaneously tests for all results in each paper. These spurious findings come in part from issues of multiple-hypothesis testing, both within regressions with several treatments and across regressions. Within regressions, treatments are largely orthogonal, but authors tend to emphasize significant *t*–values even when the corresponding *F*-tests are insignificant. Across equations,

results are often strongly correlated, so that, at worst, different regressions are reporting variants of the same result, thus spuriously adding to the 'kill count' of significant effects. At the same time, the pervasiveness of observations with high influence generates spurious significance on its own.

These issues are now being taken more seriously, at least in economics. In addition to Young (2017), Imbens and Kolesár (2016) provide practical advice for dealing with the Fisher-Behrens problem, and the best current practice tries to be careful about multiple hypothesis testing. Yet it remains the case that many of the results reported in the literature are spuriously significant.

Spurious significance also arises when the distribution of treatment effects contains outliers or, more generally, is not symmetric. Standard *t*–tests break down in distributions with enough skewness (see Lehmann and Romano (2005, 466–8)). How difficult is it to maintain symmetry? And how badly is inference affected when the distribution of treatment effects is not symmetric? One important example is expenditures on healthcare. Most people have zero expenditure in any given period, but among those who do incur expenditures, a few individuals spend huge amounts that account for a large share of the total. Indeed, in the famous Rand health experiment (see Manning, et al. (1987, 1988)), there is a single very large outlier. The authors realize that the comparison of means across treatment arms is fragile, and, although they do not see their problem exactly as described here, they obtain their preferred estimates using an approach that is explicitly designed to model the skewness of expenditures. Another example comes from economics, where many trials have outcomes valued in money. Does an anti-poverty innovation—for example microfinance—increase the incomes of the participants? Income itself is not symmetrically distributed, and this might also be true of the treatment effects if there are a few people who are talented but credit-constrained entrepreneurs and who have treatment effects that are large and positive, while the vast majority of borrowers fritter away their loans, or at best make positive but modest profits. A recent summary of the literature is consistent with this (see Banerjee, Karlan, and Zinman (2015)).

In some cases, it will be appropriate to deal with outliers by trimming, transforming, or eliminating observations that have large effects on the estimates. But if the experiment is a project evaluation designed to estimate the net benefits of a policy, the elimination of genuine outliers, as in the Rand Health Experiment, will vitiate the analysis. It is precisely the outliers that make or break the program. Transformations, such as taking logarithms, may help to produce symmetry, but they change the nature of the question being asked; a cost benefit analysis or healthcare reform costing must be done in dollars, not log dollars.

We consider an example that illustrates what can happen in a realistic but simplified case; the full results are reported in the Appendix. We imagine a population of individuals, each with a treatment effect β_i. The parent population mean of the treatment effects is zero, but there is a long tail of positive values; we use a left-shifted lognormal distribution. This could be a healthcare expenditure trial or a microfinance trial, where there is a long positive tail of rare individuals who incur very high costs or who can do amazing things with credit while most people cost nothing in the period studied or cannot use the credit effectively. A trial sample of $2n$ individuals is randomly drawn from the parent population and is randomly split between *n* treatments and *n* controls. Within each trial sample, whose true ATE will generally differ from zero because of the sampling, we run many RCTs and tabulate the values of the ATE for each.

Using standard t–tests, the (true in the parent distribution) hypothesis that the ATE is zero is rejected between 14 ($n = 25$) and 6 percent ($n = 500$) of the time. These rejections come from two separate issues, both of which are relevant in practice: (a) that the ATE in the trial sample differs from the ATE in the parent population of interest, and (b) that the *t*–values are not distributed as *t* in the presence of outliers. The problem cases are when the trial sample happens to contain one or more outliers, something that is always a risk given the long positive tail of the parent distribution. When this happens, everything depends on whether the outlier is among the treatments or the controls; in effect, the outliers become the sample, reducing the effective number of degrees of freedom. In extreme cases, one of which is

illustrated in Figure A.1, the distribution of estimated ATEs is bimodal, depending on the group to which the outlier is assigned. When the outlier is in the treatment group, the dispersion across outcomes is large, as is the estimated standard error, and so those outcomes rarely reject the null using the standard table of *t*–values. The over-rejections come from cases when the outlier is in the control group, the outcomes are not so dispersed, and the *t*–values can be large, negative, and significant. While these cases of bimodal distributions may not be common and depend on the existence of large outliers, they illustrate the process that generates the over-rejections and spurious significance. Note that there is no remedy through randomization inference here, given that our interest is in the hypothesis that the *average* treatment effect is zero.

Our reading of the literature on RCTs in social and public health policy areas suggests that they are not exempt from these concerns. Many trials are run on (sometimes very) small samples, they have treatment effects where asymmetry is hard to rule out—especially when the outcomes are in money—and they often give results that are puzzling, or at least not easily interpreted theoretically. In the context of development studies, neither Banerjee and Duflo (2012) nor Karlan and Appel (2011), who cite many RCTs, raise concerns about misleading inference, implicitly treating all results as reliable. Some of these results contradict standard theory. No doubt there are behaviors in the world that are inconsistent with conventional economics, and some can be explained by standard biases in behavioral economics, but it would also be good to be suspicious of the significance tests before accepting that an unexpected finding is well-supported and that theory must be revised. Replication of results in different settings may be helpful, if they are the right kind of places (see our discussion in Section 2). Yet it hardly solves the problem given that the asymmetry may be in the same direction in different settings, that it seems likely to be so in just those settings that are sufficiently like the original trial setting to be of use for inference about the population of interest, and that the 'significant' *t*–values will show departures from the null in the same direction. This, then, replicates the spurious findings.

1.6 *Familiar threats to unbiasedness*

It is of great importance to note that randomization, by itself, is not sufficient to guarantee unbiasedness if post-randomization differences are permitted to affect the two groups. This requires 'policing' of the experiment, for example by requiring that subjects, experimenters, and analysts are blinded and that differences in treatments or outcomes do not reveal their status to subjects. Familiar concerns about selection bias and the placebo, Pygmalion, Hawthorne, John Henry, and 'teacher/therapist' effects are widespread across studies of medical and social interventions. The difficulty of controlling for placebo effects can be especially acute in testing medical interventions (see Howick (2011), Chapter 7 for a critical review), as is the difficulty in controlling both for placebo effects and the effects of therapist variables in testing psychological therapies. For instance, Pitman, et al. (2017) suggest how difficult it will be to identify just what a psychological therapy consists of; Kramer and Stiles (2015) treat the 'responsiveness' problem of categorizing therapist responses to emerging context; and there has been a lively debate about whether cognitive mechanisms of change are responsible for the effectiveness of cognitive therapy for depression based on data that shows the changes in symptoms occur mainly before the cognitive techniques are brought into play (Ilardi and Craighead (1999), Vittengl et al. (2014)).

Many social and economic trials, medical trials, and public health trials are not blinded nor sufficiently controlled for other sources of bias, and indeed many cannot be, and a sufficient defense is rarely offered that unbiasedness is not undermined. Generally, it is recommended to extend blinding beyond participants and investigators to include those who measure outcomes and those who analyze the data, all of whom may be affected by both conscious and unconscious bias. The need for blinding in those who assess outcomes is particularly important in cases where outcomes are not determined by strictly prescribed procedures whose application is transparent and checkable but requires elements of judgment.

Beyond the need to control for 'psychological' or 'placebo' effects, blinding of trial participants is important in cases where there is no compulsion, so that people who are randomized into the treatment group are free to choose to refuse treat-

ment. In many cases it is reasonable to suppose that people choose to participate if it is in their interest to do so. In consequence, those who estimate (consciously or unconsciously) that their gain is not high enough to offset the perceived drawbacks of compliance with the treatment protocol may avoid it. The selective acceptance of treatment limits the analyst's ability to learn about people who decline treatment but who would have to accept it if the policy were implemented. In these cases, both the intention-to-treat estimator and the 'as treated' estimator that compares the treated and the untreated are affected by the kind of selection effects that randomization is designed to eliminate.

So, blinding matters for unbiasedness and is very often missing (see also Hernán et al. (2013)). This is not to say that one should assume without argument that non-blinding at any point will introduce bias. That is a matter to be assessed case-by-case. But the contrary cannot be automatically assumed. This brings to the fore the trade-off between using an RCT-based estimate that may well be biased, and in ways we do not have good ideas how to deal with, versus one from an observational study where blinding may have been easier, or some of these sources of bias may be missing or where we may have a better understanding of how to correct for them. For instance, blinding is sometimes automatic in observational studies, e.g. from administrative records. (See for example Horwitz et al. 2017 for a discussion of the complications of analyzing the result in the large Women's Health Trial when it was noted that due to the presence of side effects of the treatment "blinding was broken for nearly half of the HRT users but only a small percentage of the placebo users" [1248].)

Lack of blinding is not the only source of post-randomization bias. Subsequent treatment decisions can differ, and treatments and controls may be handled in different places, or by differently trained practitioners, or at different times of day, and these differences can bring with them systematic differences in the other causes to which the two groups are exposed. These can, and should, be guarded against. But doing so requires an understanding of what these causally relevant factors might be.

1.7 *A summary*

What do the arguments of this section mean about the importance of randomization and the interpretation that should be given to an estimated ATE from a randomized trial?

First, we should be sure that an unbiased estimate of an ATE for the trial population is likely to be useful enough to warrant the costs of running the trial.

Second, since randomization does not ensure orthogonality, to conclude that an estimate is unbiased, warrant is required that there are no significant post-randomization correlates with the treatment.

Third, the inference problems reviewed here cannot just be presumed away. When there is substantial heterogeneity, the ATE in the trial sample can be quite different from the ATE in the population of interest, even if the trial is randomly selected from that population; in practice, the relationship between the trial sample and the population is often obscure (see Longford and Nelder (1999)).

Fourth, beyond that, in many case the statistical inference will be fine, but serious attention should be given to the possibility that there are outliers in treatment effects, something that knowledge of the problem can suggest and where inspection of the marginal distributions of treatments and controls may be informative. For example, if both are symmetric, it seems unlikely (though certainly not impossible) that the treatment effects are highly skewed. Measures to deal with Fisher-Behrens should be used and randomization inference considered when appropriate to the hypothesis of interest.

All of this can be regarded as recommendations for improvement to current practice, not a challenge to it. More fundamentally, we strongly contest the often-expressed idea that the ATE calculated from an RCT is automatically reliable, that randomization automatically controls for unobservables, or worst of all, that the calculated ATE is true. If, by chance, it is close to the truth, the truth we are referring to is the truth *in the trial sample only*. To make any inference beyond that requires arguments of the kind we consider in the next section. We have also argued that, depending on what we are trying to measure and what we want to use that measure

for, there is no presumption that an RCT is the best means of estimating it. That too requires an argument, not a presumption.

Section 2: Using the results of randomized controlled trials

2.1 *Introduction*

Suppose we have estimated an ATE from a well-conducted RCT on a trial sample, and our standard error gives us reason to believe that the effect did not come about by chance. We thus have good warrant that the treatment causes the effect in our trial sample, up to the limits of statistical inference. What are such findings good for? The literature discussing RCTs has paid more attention to obtaining results than to considering what can justifiably be done with them. There is insufficient theoretical and empirical work to guide us how and for what purposes to use the findings. What there is tends to focus on the conditions under which the same results hold outside of the original settings or how they might be adapted for use elsewhere, with almost no attention to how they might be used for formulating, testing, understanding, or probing hypotheses beyond the immediate relation between the treatment and the outcome investigated in the study. Yet it cannot be that knowing *how to use* results is less important than knowing *how to demonstrate* them. Any chain of evidence is only as strong as it weakest link, so that a rigorously established effect whose applicability is justified by a loose declaration of simile warrants little. If trials are to be useful, we need paths to their use that are as carefully constructed as are the trials themselves.

The argument for the 'primacy of internal validity' made by Shadish, Cook, and Campbell (2002) may be reasonable as a warning that bad RCTs are unlikely to generalize, although as Cook (2014) notes "inferences about internal validity are inevitability probabilistic." Moreover, the primacy statement is sometimes incorrectly taken to imply that results of an internally valid trial will automatically, or often, apply 'as is' elsewhere, or that this should be the default assumption failing arguments to the contrary, as if a parameter, once well established, can be expected to be invariant across settings. The invariance assumption is often made in medicine, for example, where it is sometimes plausible that a particular procedure or drug works

the same way everywhere, though its effects cannot be the same at all stages of the disease. More generally, Horton (2000) gives a strong dissent and Rothwell (2005) provides arguments on both sides of the question. We should also note the recent movement to ensure that testing of drugs includes women and minorities because members of those groups suppose that the results of trials on mostly healthy young white males do not apply to them, as well as the increasing call for pragmatic trials, as in Williams et al. (2015): "[P]ragmatic trials … ask 'we now know it can work, but how well does it work in real world clinical practice?'"

Our approach to the use of RCT results is based on the observation that whether, and in what ways, an RCT result is evidence depends on exactly *what the hypothesis is* for which the result is supposed to be evidence, and that what kinds of hypotheses these will be depends on *the purposes to be served.* This should in turn affect the design of the trial itself. This is recognized in the medical literature in the distinction between explanatory and pragmatic trials and the proposals to adapt trial design to the question asked, as for example in Patsopoulos (2011, 218): "The explanatory trial is the best design to explore *if and how an intervention works"* whereas "The research question under investigation is *whether an intervention actually works in real life.*" It is also reflected in, for example, Rothman et l. (2013, 1013), whom we echo in arguing that simple extrapolation is not the sole purpose to which RCT results can be put: "The mistake is to think that statistical inference is the same as scientific inference." We shall distinguish a number of different purposes and discuss how, and when, RCTs can serve them: (a) simple extrapolation and simple generalization, (b) drawing lessons about the population enrolled in the trial, (c) extrapolation with adjustment, (d) estimating what happens if we scale up, (e) predicting the results of treatment on the individual, and (f) building and testing theory.

This list is hardly exhaustive. We noted in Section 1.4 one further use that we do not pursue here: The widespread and largely uncritical belief that RCTs give the right answer permits them to be used as dispute-reconciliation mechanisms to resolve political conflicts. For example, at the Federal level in the US, prospective policies are vetted by the non-partisan Congressional Budget Office (CBO), which makes its own estimates of budgetary implications. Ideologues whose programs are scored

poorly by the CBO have an incentive to support an RCT, not to convince themselves, but to convince opponents. Once again, RCTs are valuable when your opponents do not share your prior.

2.2 *Simple extrapolation and simple generalization*

Suppose a trial has (probabilistically) established a result in a specific setting. If `the same' result holds elsewhere, it is said to have *external validity*. External validity may refer just to the replication of the causal connection or go further and require replication of the magnitude of the ATE. Either way, the result holds—everywhere, or widely, or in some specific elsewhere—or it does not.

This binary concept of external validity is often unhelpful because it asks the results of an RCT to satisfy a condition that is neither necessary nor sufficient for trials to be useful, and so both overstates and understates their value. It directs us toward *simple extrapolation*—whether the same result holds elsewhere—or *simple generalization*—it holds universally or at least widely—and away from more complex but equally useful applications of the results. The failure of external validity interpreted as simple generalization or extrapolation says little about the value of the results of the trial.

There are several uses of RCTs that do not require applying their results beyond the original context; we discuss these in Section 2.4. Beyond that, there are often good reasons to expect that the results from a well-conducted, informative, and potentially useful RCT will *not* apply elsewhere in any simple way. Without further understanding and analysis, even successful replication tells us little either for or against simple generalization nor does much to support the conclusion that the next will work in the same way. Nor do failures of replication make the original result useless. We often learn much from coming to understand why replication failed and can use that knowledge in looking for how the factors that caused the original result might operate differently in different settings. Third, and particularly important for scientific progress, the RCT result can be incorporated into a network of evidence and hypotheses that test or explore claims that look very different from the results reported from the RCT. We shall give examples below of valuable uses for RCTs that

are not externally valid in the (usual) sense that their results do not hold elsewhere, whether in a specific target setting or in the more sweeping sense of holding everywhere, or everywhere in some specified domain.

The RAND health experiment (Manning et al. (1987, 88)) provides an instructive story if only because its results have permeated the academic and policy discussions about healthcare ever since. It was originally designed to test whether more generous insurance causes people to use more medical care and, if so, by how much. The incentive effects are hardly in doubt today; the immortality of the study comes rather from the fact that its multi-arm (response surface) design allowed the calculation of an elasticity for the study population, that medical expenditures decreased by –0.1 to –0.2 percent for every percentage increase in the copayment. According to Aron-Dine et al. (2013), it is this dimensionless and thus apparently exportable number that has been used ever since to discuss the design of healthcare policy; the elasticity has come to be treated as a universal constant. Ironically, they argue that the estimate cannot be replicated in recent studies, and that it is unclear that it is firmly based on the original evidence. The simple direct exportability of the result was perhaps illusory.

The drive to export and generalize RCTs results is at the core of the influential 'what works' movement across the medical and social sciences. At its most ambitious, this aims for universal reach. For example, in the development economics literature, Duflo and Kremer (2008, 93) argue that "credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations (NGOs) beyond national borders." Sometimes the results of a single RCT are advocated as having wide applicability, with especially strong endorsement when there is at least one replication.

Simple extrapolation is often used to move RCT results from one setting to another. Much of what is written in the 'what works' literature suggests that, unless there is evidence to the contrary, the direction and size of treatment effects can be transported from one place to another without serious adjustment. The Abdul Latif Jameel Poverty Action Lab (J-PAL) conducts RCTs around the world and summarizes

findings in an attempt to reduce poverty by the use of "scientific evidence to inform policy." Some of their reports convert results into a common cost-effectiveness measure. For example, *Improving Student Participation--Which programs most effectively get children into school?* classifies results into six categories: school time travel, subsidies and transfers, health, perceived returns, education quality, and gender specific barriers; results are reported in the common unit, "additional years of education for US$100 spent." "Health", which top-rated by far, includes two studies, "deworming" in Kenya (11.91) and "iron & vitamin A" in India (2.61); "perceived returns" to education has one study in the Dominican Republic (0.23); "subsidies and transfers" includes the most studies—six, with results ranging from 0.17 for "secondary scholarships" in Ghana to 0.01, for "CCT" (Conditional Cash Transfers) in Mexico and 0.09 and 0.07 for "CCT" in Malawi.

What can we conclude from such comparisons? A philanthropic donor interested in education, who assumes that marginal and average effects are the same, might learn that the best place to devote a marginal dollar is in Kenya, where it would be used for deworming. This is certainly useful, but it is not as useful as statements that deworming programs are everywhere more cost-effective than programs involving vitamin A or scholarships, or if not everywhere, at least over some domain, and it is these second kinds of comparison that would genuinely fulfill the promise of 'finding out what works.' But such comparisons only make sense if the results from one place can be relied on to apply in another, if the Kenyan results also hold in the Dominican Republic, Mexico, Ghana, or in some specific list of places.

What does J-PAL conclude? Here are two of their reported "Practical Implications": "Conditional and unconditional cash transfers can increase school enrolment and attendance, but are expensive to implement...Eliminating small costs can have substantial impacts on school participation." 'Can' here is admittedly an ambiguous word. It is certainly true in a logical sense that if a program has achieved a given result, then it can do so. But we suspect that the more natural sense for readers to take away is that the program 'may well' do so most other places, in the absence of special problems, or that that is at least the default assumption.

Trials, as is widely noted, often take place in artificial environments which raises well recognized problems for extrapolation. For instance, with respect to economic development, Drèze (J. Drèze, personal communications, November 8, 2017) notes, based on extensive experience in India, that "when a foreign agency comes in with its heavy boots and deep pockets to administer a 'treatment,' whether through a local NGO or government or whatever, there tends to be a lot going on other than the treatment." There is also the suspicion that a treatment that works does so because of the presence of the 'treators,' often from abroad, and may not do so with the people who will work it in practice.

J-PAL's manual for cost-effectiveness (Dhaliwal et al. (2012)) explains in (entirely appropriate) detail how to handle variation in costs across sites, noting variable factors such as population density, prices, exchange rates, discount rates, inflation, and bulk discounts. But it gives short shrift to cross-site variation in the size of ATEs, which also play a key part in the calculations of cost effectiveness. The manual briefly notes that diminishing returns (or the last-mile problem) might be important in theory but argues that the baseline levels of outcomes are likely to be similar in the pilot and replication areas, so that the ATE can be safely assumed to apply as is. All of this lacks a justification for extrapolating results, some understanding of when results can be extrapolated, when they cannot, or better still, how they should be modified to make them applicable in a new setting. Without well substantiated assumptions to support the projection of results, this is just induction by simple enumeration—swan 1 is white, swan 2 is white, . . . , so all swans are white; and, as Francis Bacon (1859, 1.105) taught, "…the induction that proceeds by simple enumerations is childish."

Bertrand Russell's chicken (Russell (1912)) provides an excellent example of the limitations to simple extrapolation from repeated successful replication. The bird infers, on repeated evidence, that when the farmer comes in the morning, he feeds her. The inference serves her well until Christmas morning, when he wrings her neck and serves her for dinner. Though this chicken did not base her inference on an RCT, had we constructed one for her, we would have obtained the same result that she did. Her problem was not her methodology, but rather that she did not un-

derstand the social and economic structure that gave rise to the causal relations that she observed. (We shall return to the importance of the underlying structure for understanding what causal pathways are likely and what are unlikely below.)

The problems with simple extrapolation and simple generalization extend beyond RCTs, to both 'fully controlled' laboratory experiments and to most non-experimental findings. Our argument here is that evidence from RCTs is *not* automatically simply generalizable, and that its superior internal validity, if and when it exists, does not provide it with any unique invariance across context. That simple extrapolation and simple generalization are far from automatic also tells us why (even ideal) RCTs of similar interventions give different answers in different settings and the results of large RCTs may differ from the results of meta-analyses on the same treatment (as in LeLorier et al. (1997)). Such differences do not necessarily reflect methodological failings and will hold across perfectly executed RCTs just as they do across observational studies.

Our arguments are not meant to suggest that extrapolation or even generalization is never reasonable. For instance, conditional cash transfers have worked for a variety of different outcomes in different places; they are often cited as a leading example of how an evaluation with strong internal validity leads to a rapid spread of the policy. Think through the causal chain that is required for CCTs to be successful: People must like money, they must like (or do not object too much) to their children being educated and vaccinated, there must exist schools and clinics that are close enough and well enough staffed to do their job, and the government or agency that is running the scheme must care about the wellbeing of families and their children. That such conditions hold in a wide range of (although certainly not all) countries makes it unsurprising that CCTs 'work' in many replications, though they certainly will not work in places where the schools and clinics do not exist, e.g. Levy (2006), nor in places where people strongly oppose education or vaccination. So, there are structural reasons why CCT results export where they do. Our objection is to the assumption that it is 'natural' that well-established results export; to the contrary, good reasons are needed to justify that they do.

To summarize. Establishing *causality* does nothing in and of itself to guarantee that the causal relation will hold in some new case, let alone in general. Nor does the ability of an ideal RCT to eliminate bias from selection or from omitted variables mean that the resulting ATE from the trial sample will apply anywhere else. The issue is worth mentioning only because of the enormous weight that is currently attached to policing the rigor with which causal claims are established by contrast with the rigor devoted to all those further claims—often unstated even—that go into warranting extrapolating or generalizing the relations.

2.3 *Support factors and the ATE*

The operation of a cause generally requires the presence of *support factors* (also known as 'interactive variables' or 'moderators'), factors without which a cause that produces the targeted effect in one place, even though it may be present and have the capacity to operate elsewhere, will remain latent and inoperative. What Mackie (1974) called INUS causality (Insufficient but Non-redundant parts of a condition that is itself Unnecessary but Sufficient for a contribution to the outcome) is the kind of causality reflected in equation (1). (See Rothman (1976, 2012) for the same idea in epidemiology, which uses the term 'causal pie' to refer to a set of causes that are jointly but not separately sufficient for a contribution to an effect.) A standard example is a house burning down *because* the television was left on, although televisions do not operate in this way without support factors, such as wiring faults, the presence of tinder, and so on.

The value of the ATE depends on the distribution of the values of the 'support factors' necessary for T to contribute to Y. This becomes clear if we rewrite (1) in the form

$$Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij} = \theta(w_i) T_i + \sum_{j=1}^{J} \gamma_j x_{ij} \quad \#(4)$$

where the function $\theta(.)$ controls how a k-vector w_i of k 'support factors' affect individual i's treatment effect β_i. The support factors may include some of the x's. Since the ATE is the average of the $\beta_i s$, two populations will have the same ATE if and only

if they have the same average for the net effect of the support factors necessary for the treatment to work, i.e. for the quantity in front of T_i. These are however just the kind of factors that are likely to be differently distributed in different populations.,

Given that support factors will operate with different strengths and effectiveness in different places, it is not surprising that the size of the ATE differs from place to place; for example, Vivalt's AidGrade website lists 29 estimates from a range of countries of the standardized (divided by local standard deviation of the outcome) effects of CCTs on school attendance; all but four show the expected positive effect, and the range runs from –8 to +38 percentage points (Vivalt (2016)). Even in this leading case, where we might reasonably conclude that CCTs 'work' in getting children into school, it would be hard to calculate credible cost-effectiveness numbers or to come to a general conclusion about whether CCTs are more or less cost effective than other possible policies. Both costs and effect sizes can be expected to differ in new settings, just as they have in observed ones, making these predictions difficult.

AidGrade uses standardized measures of effect size divided by standard deviation of outcome at baseline, as does the major multi-country study by Banerjee et al. (2015). But we might prefer measures that have an economic interpretation, such as J-PAL's 'additional months of schooling per US$100 spent' (for example if a donor is trying to decide where to spend, as we noted). Nutrition might be measured by height, or by the log of height. Even if the ATE by one measure carries across, it will only do so using another measure if the relationship between the two measures is the same in both situations. This is exactly the sort of thing that a formal analysis of what reasons justify simple extrapolation and how to adjust predictions when simple extrapolation is not justified forces us to think about. (Note also that the ATE in the original RCT can differ depending on whether the outcome is measured in levels or in logs; it is easy to construct examples where the two ATEs have different signs.)

The worry is not just that the distribution of values for the support factors in a new setting will differ from the distribution in the trial but that what those support factors are will differ, or indeed whether there are any at all in the new setting that can get the treatment to work there. Causal processes often require highly spe-

cialized economic, cultural, or social structures to enable them to work. Different structures will enable different processes with different causes and different support factors. Consider the Rube Goldberg machine that is rigged up so that flying a kite sharpens a pencil (Cartwright and Hardie (2012, 77)). The underlying structure affords a very specific form of (4) that will not describe causal processes elsewhere. The Rube Goldberg machine is an exaggerated example, but it makes transparent how unreliable simple extrapolation is likely to be when little knowledge of causal structure is available.

For more typical examples, consider systems design, where we aim to construct systems that will generate causal relations that we like and that will rule out causal relations that we do not like. Healthcare systems are designed to prevent nurses and doctors making errors; cars are designed so that drivers cannot start them in reverse; work schedules for pilots are designed so they do not fly too many consecutive hours without rest because alertness and performance are compromised. In philosophy, a system of interacting parts that underpins causal processes and makes some possible and some impossible, some likely and some unlikely is labelled a *mechanism.* (Note that this is only one of many meanings in philosophy and elsewhere for the term 'mechanism'; in particular it is not 'mechanism' in the sense of the causal pathway from treatment to outcomes, which is another common use, for example in Suzuki et al. (2011)). Mechanisms are particularly important in understanding the explanation of causal processes in biology and the philosophical literature is rife with biological examples, as in the account in the seminal Machamer et al. (2000) of how Shepherd (1988) uses biochemical mechanisms at chemical synapses to explain the process of transmitting electrical signals from one neuron to another. (See also Bechtel (2006), Craver (2007).) 'Mechanism' in this sense is nor restricted to physical parts and their interactions and constraints but includes social, cultural, and economic arrangements, institutions, norms, habits, and individual psychology. (See, for example, Seckinelgin (2016) on the importance of context in determining the effectiveness of HIV-AIDs therapies.)

As in the Rube Goldberg machine and in the design of cars and work schedules, the physical, social, and economic structure and equilibrium may differ in ways

that support, permit, or block different kinds of causal relations and thus render a trial in one setting useless in another. For example, a trial that relies on providing incentives for personal promotion is of no use in a state in which a political system locks people into their social and economic positions. Cash transfers that are conditional on parents taking their children to clinics cannot improve child health in the absence of functioning clinics. Policies targeted at men may not work for women. We use a lever to toast our bread, but levers only operate to toast bread in a toaster; we cannot brown toast by pressing an accelerator, even if the principle of the lever is the same in both a toaster and a car. If we misunderstand the setting, if we do not understand *why* the treatment in our RCT works, we run the same risks as Russell's chicken. (See Little (2007) and Howick et al. (2013) for many of the difficulties in using claims about mechanistic structure to support extrapolation, and Parkkinen et al. (2018) defending the importance of mechanistic reasoning both for internal validity and for extrapolation.)

2.4 *When RCTs speak for themselves: no extrapolation or generalization required*

For some things we want to learn, an RCT is enough by itself. An RCT may provide a counterexample to a general theoretical proposition, either to the proposition itself (a simple refutation test) or to some consequence of it (a complex refutation test). An RCT may also confirm a prediction of a theory, and although this does not confirm the theory, it is evidence in its favor, especially if the prediction seems inherently unlikely in advance. This is all familiar territory, and there is nothing unique about an RCT; it is simply one among many possible testing procedures. Even when there is no theory, or very weak theory, an RCT, by demonstrating causality in *some* population can be thought of as *proof of concept*, that the treatment is capable of working *somewhere* (as in the remark from Curtis Meinert, prominent expert on clinical trial methodology: "There is no point in worrying whether a treatment works the same or differently in men and women until it has been shown to work in someone" (quoted in Epstein (2007, 108))). This is one of the arguments for the importance of internal validity.

Nor is extrapolation called for when an RCT is used for evaluation, for example to satisfy donors that the project they funded achieved its aims in the population in which it was conducted. Even so, for such evaluations, say by the World Bank, to be useful to the world at large (to be global public goods) requires arguments and guidelines that justify using the results in some way elsewhere; the global public good is not an automatic by-product of the Bank fulfilling its fiduciary responsibility. We need something, some regularity or invariance, and that something can rarely be recovered by simply generalizing across trials.

A third non-problematic and important use of an RCT is when the parameter of interest is the ATE in a well-defined population from which the trial sample is itself a random sample. In this case the sample average treatment effect (SATE) is an unbiased estimator of the population average treatment effect (PATE) that, by assumption, is our target (see Imbens (2004) for these terms). We refer to this as the 'public health' case; like many public health interventions, the target is the average, 'population health,' not the health of individuals. One major (and widely recognized) danger of this use of RCTs is that exporting results from (even a random) sample to the population will not go through in any simple way if the outcomes of individuals or groups of individuals change the behavior of others—which is common in social examples and in public health whenever there is a possibility of contagion.

2.5 *Reweighting and stratifying*

Many advocates of RCTs understand that 'what works' needs to be qualified to `what works under which circumstances' and try to say something about what those circumstances might be, for example, by replicating RCTs in different places and thinking intelligently about the differences in outcomes when they find them. Sometimes this is done in a systematic way, for example by having multiple treatments within the same trial so that it is possible to estimate a 'response surface' that links outcomes to various combinations of treatments (see Greenberg and Schroder (2004) or Shadish et al. (2002)). For example, the RAND health experiment had multiple treatments, allowing investigation of how much health insurance increased expendi-

tures under different circumstances. Some of the negative income tax experiments (NITs) in the 1960s and 1970s were designed to estimate response surfaces, with the number of treatments and controls in each arm optimized to maximize precision of estimated response functions subject to an overall cost limit (see Conlisk (1973)). Experiments on time-of-day pricing for electricity had a similar structure (see Aigner (1985)).

The experiments by MDRC have also been analyzed across cities in an effort to link city features to the results of the RCTs within them (see Bloom et al. (2005)). Unlike the RAND and NIT examples, these are *ex post* analyses of completed trials; the same is true of Vivalt (2015), who finds, for the collection of trials she studied, that development-related RCTs run by government agencies typically find smaller (standardized) effect sizes than RCTs run by academics or by NGOs. Bold et al. (2013), who ran parallel RCTs on an intervention implemented either by an NGO or by the government of Kenya, found similar results there. Note that these analyses have a different purpose from meta-analyses that assume that different trials estimate the same parameter up to noise and average in order to increase precision.

Statistical approaches are also widely used to adjust the results from a trial population to predict those in a target population; these are designed to deal with the fact that treatment effects vary systematically with variations in the support factors. One procedure to deal with this is *post-experimental stratification*, which parallels post-survey stratification in sample surveys. The trial is broken up into subgroups that have the same combination of known, observable w's (age, race, gender, co-morbidities for example), then the ATEs within each of the subgroups are calculated, and then they are reassembled according to the configuration of w's in the new context. This can be used to estimate the ATE in a new context, or to correct estimates to the parent population when the trial sample is not a random sample of the parent. Other methods can be used when there are too many w's for stratification, for example by estimating the probability of each observation in the population included in the trial sample as a function of the w's, then weighting each observation by the inverse of these propensity scores. A good reference for these methods is Stuart et al. (2011), or in economics, Angrist (2004) and Hotz et al. (2005).)

These methods are often not applicable, however. First, reweighting works only when the observable factors used for reweighting include all (and only) genuine interactive causes (support/moderator factors). Second, as with any form of reweighting, the variables used to construct the weights must be present in both the original and new context. For example, if we are to carry a result forward in time, we may not be able to extrapolate from a period of low inflation to a period of high inflation; medical treatments that work in cold climates may not work in the tropics. As Hotz et al. (2005) note, it will typically be necessary to rule out such 'macro' effects, whether over time, or over locations. Third, reweighting also depends on the assumption that the same governing equation (4) covers both the trial and the target population.

Pearl and Bareinboim (2011, 2014) and Bareinboim and Pearl (2013, 2014) provide strategies for inferring information about new populations from trial results that are more general than reweighting. They suppose we have available both causal information and probabilistic information for population *A* (e.g. the experimental one), while for population *B* (the target) we have only (some) probabilistic information, and also that we know that certain probabilistic and causal facts are shared between the two and certain ones are not. They offer theorems describing what causal conclusions about population *B* are thereby fixed. Their work underlines the fact that exactly what conclusions about one population can be supported by information about another depends on exactly what causal and probabilistic facts they have in common. But as Muller (2015) notes, this, like the problem with simple reweighting, takes us back to the situation that RCTs are designed to avoid, where we need to start from a complete and correct specification of the causal structure. RCTs can avoid this in estimation—which is one of their strengths, supporting their credibility—but the benefit vanishes as soon as we try to carry their results to a new context.

This discussion leads to a number of points. First it underlines our previous arguments that we cannot get to general claims by simple generalization; there is no warrant for the convenient assumption that the ATE estimated in a specific RCT is

an invariant parameter, nor that the kinds of interventions and outcomes we measure in typical RCTs participate in general causal relations.

Second, thoughtful pre-experimental stratification in RCTs is likely to be valuable, or failing that, subgroup analysis, because it can provide information that may be useful for generalization or extrapolation. For example, Kremer and Holla (2009) note that, in their trials, school attendance is surprisingly sensitive to small subsidies, which they suggest is because there are a large number of students and parents who are on the (financial) margin between attending and not attending school; if this is indeed the mechanism for their results, a good variable for stratification would be distance from the relevant cutoff. We also need to know that this same mechanism works in any new target setting, as discussed at the end of Section 2.3.

Third, we need to be explicit about causal structure, even if that means more model building and more—or different—assumptions than advocates of RCTs are often comfortable with. We need something, some regularity or invariance, and that something can rarely be recovered by simply generalizing across trials. To be clear, modeling causal structure does not commit us to the elaborate and often incredible assumptions that characterize some structural modeling in economics, but there is no escape from thinking about the way things work; the why as well as the what.

Fourth, to use these techniques for reweighting and stratifying, we will need to know more than the results of the RCT itself, for example about differences in social, economic, and cultural structures and about the joint distributions of causal variables, knowledge that will often only be available through observational studies. We will also need external information, both theoretical and empirical, to settle on an informative characterization of the population enrolled in the RCT because how that population is described is commonly taken to be some indication of which other populations would yield similar results.

Many medical and psychological journals are explicit about this. For instance, the rules for submission recommended by the International Committee of Medical Journal Editors, ICMJE (2015, 14) insist that article abstracts "Clearly describe the selection of observational or experimental participants (healthy individuals or patients, including controls), including eligibility and exclusion criteria and a descrip-

tion of the source population." An RCT is conducted on a specific trial sample, somehow drawn from a population of specific individuals. The results obtained are features of that sample, of those *very* individuals at that *very* time, not any other population with any different individuals that might, for example, satisfy one of the infinite set of descriptions that the trial sample satisfies. If following the ICMJE advice is to produce warrantable extrapolation—simple or adjusted—from a trial population to some other, the descriptors for the trial population must be correctly chosen. As we have argued, they must pick out populations where the same form of equation (4) holds and that have approximately the same mean (or one that we know how to adjust) for the net effect of the support factors in the two populations.

This same issue is confronted already in study design. Apart from special cases, like post hoc evaluation for payment-for-results, we are not especially concerned to learn about the very individuals enrolled in the trial. Most experiments are, and should be, conducted with an eye to what the results can help us learn about other populations. This cannot be done without substantial assumptions about what might and what might not be relevant to the production of the outcome studied. So both intelligent study design and responsible reporting of study results involve substantial background assumptions.

Of course, this is true for all studies. But RCTs require special conditions if they are to be conducted at all and especially if they are to be conducted successfully—for example, local agreements, compliant subjects, affordable administrators, multiple blinding, people competent to measure and record outcomes reliably, a setting where random allocation is morally and politically acceptable, etc.—whereas observational data are often more readily and widely available. In the case of RCTs, there is danger that these kinds of considerations have too much effect. This is especially worrisome where the features that the trial sample should have are not justified, made explicit, or subjected to serious critical review.

The need for observational knowledge is one of many reasons why it is counter-productive to insist that RCTs are the gold standard or that some categories of evidence should be prioritized over others; these strategies leave us helpless in using RCTs beyond their original context. The results of RCTs must be integrated with

other knowledge, including the practical wisdom of policymakers, if they are to be useable outside the context in which they were constructed.

Contrary to much practice in medicine as well as in economics, conflicts between RCTs and observational results need to be explained, for example by reference to the different characteristics of the different populations studied in each, a process that will sometimes yield important evidence, including on the range of applicability of the RCT results themselves. While the validity of the RCT will sometimes provide an understanding of why the observational study found a different answer, there is no basis (or excuse) for the common practice of dismissing the observational study simply because it was not an RCT and therefore must be invalid. It is a basic tenet of scientific advance that, as collective knowledge advances, new findings must be able to explain and be integrated with previous results, even results that are now thought to be invalid; methodological prejudice is not an explanation.

2.6 *Using RCTs to build and test theory*

RCT results, as with any well-established scientific claims, can be used in the familiar hypothetico-deductive way to test theory.

For example, one of the largest and most technically impressive of the development RCTs is by Banerjee et al. (2015), which tests a 'graduation' program designed to permanently lift extremely poor people from poverty by providing them with a gift of a productive asset (from guinea-pigs, (regular-) pigs, sheep, goats, or chickens depending on locale), training and support, and life-skills coaching, as well as support for consumption, saving, and health services. The idea is that this package of aid can help people break out of poverty traps in a way that would not be possible with one intervention at a time. Comparable versions of the program were tested in Ethiopia, Ghana, Honduras, India, Pakistan, and Peru and, excepting Honduras (where the chickens died) find largely positive and persistent effects—with similar (standardized) effect sizes—for a range of outcomes (economic, mental and physical health, and female empowerment). One site apart, essentially everyone accepted their assignment. Replication of positive ATEs over such a wide range of

places certainly provides proof of concept for such a scheme. Yet Bauchet et al. (2015) fail to replicate the result in South India, where the control group got access to much the same benefits. (Heckman, et al. (2000) call this `substitution' bias). Even so, the results are important because, although there is a longstanding interest in poverty traps, many economists have been skeptical of their existence or that they could be sprung by such aid-based policies. In this sense, the study is an important contribution to the *theory* of economic development; it tests a theoretical proposition and will (or should) change minds about it.

Economists have been combining theory and randomized controlled trials in a variety of other ways since the early experiments. The trials help build and test theory and theory in turn can answer questions about new settings and populations that we cannot answer by simple extrapolation or generalization of the trial results. We will outline a few economics examples to give a sense of how the interweaving of theory and results can work.

Orcutt and Orcutt (1968) laid out the inspiration for the income tax trials using a simple static theory of labor supply. According to this, people choose how to divide their time between work and leisure in an environment in which they receive a minimum G if they do not work, and where they receive an additional amount $(1-t)\ w$ for each hour they work, where w is the wage rate, and t is a tax rate. The trials assigned different combinations of G and t to different trial groups, so that the results traced out the labor supply function, allowing estimation of the parameters of preferences, which could then be used in a wide range of policy calculations, for example to raise revenue at minimum utility loss to workers.

Following these early trials, there has been a continuing tradition of using trial results, together with the baseline data collected for the trial, to fit structural models that are to be used more generally. (Early examples include Moffitt (1979) on labor supply and Wise (1985) on housing; a more recent example is Heckman et al. (2013) for the Perry pre-school program. Development economics examples include Attanasio et al. (2012), Attanasio et al. (2015), Todd and Wolpin (2006), Wolpin (2013), and Duflo et al. (2012).) These structural models sometimes require

formidable auxiliary assumptions on functional forms or the distributions of unobservables, but they have compensating advantages, including the ability to integrate theory and evidence, to make out-of-sample predictions, and to analyze welfare, and the use of RCT evidence allows the relaxation of at least some of the assumptions that are needed for identification. In this way, the structural models borrow credibility from the RCTs and in return help set the RCT results within a coherent framework. Without some such interpretation, the welfare implications of RCT results can be problematic; knowing how people in general (let alone just people in the trial population) respond to some policy is rarely enough to tell whether or not they are made better off, Harrison (2014a, b). Traditional welfare economics draws a link from preferences to behavior, a link that is respected in structural work but often lost in the 'what works' literature, and without which we have no basis for inferring welfare from behavior. What works is not equivalent to what should be.

Even simple theory can do much to interpret, to extend, and to use RCT results. In both the RAND Health Experiment and negative income tax experiments, an immediate issue concerned the difference between short and long-run responses; indeed, differences between immediate and ultimate effects occur in a wide range of RCTs. Both health and tax RCTs aimed to discover what would happen if consumers/workers were *permanently* faced with higher or lower prices/wages, but the trials could only run for a limited period. A *temporarily* high tax rate on earnings is effectively a 'fire sale' on leisure, so that the experiment provided an opportunity to take a vacation and make up the earnings later, an incentive that would be absent in a permanent scheme. How do we get from the short-run responses that come from the trial to the long-run responses that we want to know? Metcalf (1973) and Ashenfelter (1978) provided answers for the income tax experiments, as did Arrow (1975) for the Rand Health Experiment.

Arrow's analysis illustrates how to use both structure and observational data in combination with results from one setting to predict results in another. He models the health experiment as a two-period model in which the price of medical care is lowered in the first period only, and shows how to derive what we want, which is the response in the first period if prices were lowered by the same proportion in

both periods. The magnitude that we want is *S*, the compensated price derivative of medical care in period 1 in the face of identical increases in p_1 and p_2 in both periods 1 and 2. This is equal to $s_{11} + s_{12}$, the sum of the derivatives of period 1's demand with respect to the two prices. The trial gives only s_{11}. But if we have post-trial data on medical services for both treatments and controls, we can infer s_{21}, the effect of the experimental price manipulation on post-experimental care. Choice theory, in the form of Slutsky symmetry says that $s_{12} = s_{21}$ and so allows Arrow to infer s_{12} and thus *S.* He contrasts this with Metcalf's alternative solution, which makes different assumptions—that two period preferences are intertemporally additive, in which case the long-run elasticity can be obtained from knowledge of the income elasticity of post-experimental medical care, which would have to come from an observational analysis.

These two alternative approaches show how we can choose, based on our willingness to make assumptions and on the data that we have, a suitable combination of (elementary and transparent) theoretical assumptions and observational data in order to adapt and use trial results. Such analysis can also help design the original trial by clarifying what we need to know in order to use the results of a temporary treatment to estimate the permanent effects that we need. Ashenfelter provides a third solution, noting that the two-*period* model is formally identical to a two-*person* model, so that we can use information on two-person labor supply to tell us about the dynamics. In the Rand case, internal evidence suggests that short-run and long-run responses were not in fact very different, but Arrow's analysis provides an illustration of how theory can form a bridge from what we get to what we want.

Theory can often allow us to reclassify new or unknown situations as analogous to situations where we already have background knowledge. In economics, one frequently useful way of doing this is when the new policy can be recast as equivalent to a change in the prices and incomes faced by respondents. The consequences of a new policy may be easier to predict if we can reduce it to equivalent changes in income and prices, whose effects are often well understood and well-studied. Todd and Wolpin (2008) and Wolpin (2013) make this point and provide examples. In the labor supply case, an increase in the tax rate has the same effect as a decrease in the

wage rate, so that we can rely on previous literature to predict what will happen when tax rates are changed. In the case of Mexico's PROGRESA conditional cash transfer program, Todd and Wolpin note that the subsidies paid to parents if their children go to school can be thought of as a combination of reduction in children's wages and an increase in parents' income, which allows them to predict the results of the conditional cash experiment with limited additional assumptions. If this works, as it partially does in their analysis, the trial helps consolidate previous knowledge and contributes to an evolving body of theory and empirical, including trial, evidence.

The program of thinking about policy changes as equivalent to price and income changes has a long history in economics; much of rational choice theory can be so interpreted (see Deaton and Muellbauer (1980) for many examples). When this conversion is credible, and when a trial on some apparently unrelated topic can be modeled as equivalent to a change in prices and incomes, and when we can assume that people in different settings respond similarly to changes in prices and incomes, we have a readymade framework for incorporating the trial results into previous knowledge, as well as for extending the trial results and using them elsewhere. Of course, all depends on the validity and credibility of the theory; people may not in fact treat a tax increase as a decrease in the price of leisure, and behavioral economics is full of examples where apparently equivalent stimuli generate non-equivalent outcomes. The embrace of behavioral economics by many of the current generation of researchers may account for their limited willingness to use conventional choice theory in this way. Unfortunately, behavioral economics does not yet offer a replacement for the general framework of choice theory that is so useful in this regard.

Theory can also help with the problems we raised in the summary of Section 1., that people who are randomized into the treatment group may refuse treatment. When theory is good enough to indicate how to represent the gain and losses that trial participants are likely to base compliance on, then analysis can sometimes help us adjust the trial estimates back to what we would like to know.

2.6 *Scaling up: using the average for populations*

Many RCTs are small-scale and local, for example in a few schools, clinics, or farms in a particular geographic, cultural, socio-economic setting. If successful according to a cost-effectiveness criterion, for example, it is a candidate for scaling-up, applying the same intervention for a much larger area, often a whole country, or sometimes even beyond, as when some treatment is considered for all relevant World Bank projects. Predicting the same results at scale as in the trial is a case of simple extrapolation. We discuss it separately, however, because it can raise special problems. The fact that the intervention might work differently at scale has long been noted in the economics literature, e.g. Garfinkel and Manski (1992), Heckman (1992), and Moffitt (1992), and is recognized in the recent review by Banerjee and Duflo (2009).

In medicine, where biological interactions between people are less common than are social interactions in social science, they can still be important. Infectious diseases are a well-known example, where immunization programs affect the dynamics of disease transmission through herd immunity (see Fine and Clarkson (1986) and Manski (2013, 52)). The social and economic setting also affects how drugs are actually used and the same issues can arise; the distinction between efficacy and effectiveness in clinical trials is in part recognition of the fact. We want here to emphasize the pervasiveness of such effects as well as to note again that this should not be taken as an argument against using RCTs but only against the idea that effects at scale are likely to be the same as in the trial.

An example of what are often called 'general equilibrium effects' comes from agriculture. Suppose an RCT demonstrates that in the study population a new way of using fertilizer had a substantial positive effect on, say, cocoa yields, so that farmers who used the new methods saw increases in production and in incomes compared to those in the control group. If the procedure is scaled up to the whole country, or to all cocoa farmers worldwide, the price will drop, and if the demand for cocoa is price inelastic—as is usually thought to be the case, at least in the short run—cocoa farmers' incomes will fall. Indeed, the conventional wisdom for many crops is that farmers do best when the harvest is small, not large. In this case, the scaled-up effect is *opposite in sign* to the trial effect. The problem is not with the trial results, which

can be usefully incorporated into a more comprehensive market model that incorporates the responses estimated by the trial. The problem is only if we assume that the aggregate looks like the individual. That other ingredients of the aggregate model must come from observational studies should not be a criticism, even for those who favor RCTs; it is simply the price of doing serious analysis.

There are many possible interventions that alter supply or demand whose effect, in aggregate, will change a price or a wage that is held constant in the original RCT. Indeed, any trial that changes the quantities that people demand or supply—including labor supply—must, as a matter of logic, affect other people because the new demand has to be met, or the new supply accommodated. In the language of the Rubin causal model, this is a failure of SUTVA, the stable unit treatment value assumption. Of course, each unit may be too small to have any perceptible effect by itself, so SUTVA holds to a high degree of approximation in the trial, but once we aggregate to the population, the effects will often be large enough to modify or reverse the result from the trial. Examples include that education will change the supplies of skilled versus unskilled labor, with implications for relative wage rates. Conditional cash transfers increase the demand for (and perhaps supply of) schools and clinics, which will change prices or waiting lines, or both. There are interactions between people that will operate only at scale. Giving one child a voucher to go to private school might improve her future, but doing so for everyone can decrease the quality of education for those children who are left in the public schools (see the contrasting studies of Angrist et al. (2002) and Hsieh and Urquiola (2006)). Educational or training programs may benefit those who are treated but harm those left behind; Crépon et al. (2014) recognize the issue and show how to adapt an RCT to deal with it.

Much of economics is concerned with analyzing equilibria, most obviously in the equilibrium of supply and demand. Multiple causal mechanisms are reconciled by the adjustment of some variable, such as a price. RCTs will often be useful in analyzing one or other mechanism, in which the equilibrating variable is held constant, and the results of those RCTs can be used to analyze and predict the equilibrium effects of policies. But the results of implementing policies will often look very differ-

ent from the trial results, as in the cocoa example above. If, as is often argued, economics is about the analysis of equilibrium, simple extrapolation of the results of an RCT will rarely be useful. Note that we are making no claim about the success of economic models, either in analysis or prediction. But the analysis of equilibrium is a matter of logical consistency without which we are left with contradictory propositions.

2.7 *Drilling down: using the average for individuals*

Just as there are issues with scaling-up, it is not obvious how to use the results from RCTs at the level of individual units, even individual units that were included in the trial. A well-conducted RCT delivers an ATE for the trial population but, in general, that average does *not* apply to everyone. It is not true, for example, as argued in the American Medical Association's *Users' guide to the medical literature* that "if the patient would have been enrolled in the study had she been there—that is she meets all of the inclusion criteria and doesn't violate any of the exclusion criteria—there is little question that the results are applicable" (see Guyatt et al. (1994, 60)). Even more misleading are the often-heard statements that an RCT with an *average* treatment effect insignificantly different from zero has shown that the treatment works for *no one.*

These issues are familiar to physicians practicing evidence-based medicine whose guidelines require "integrating individual clinical expertise with the best available external clinical evidence from systematic research" Sackett et al. (1996, 71)). Exactly what this means is unclear; physicians know much more about their patients than is allowed for in the ATE from the RCT (though, once again, stratification in the trial is likely to be helpful) and they often have intuitive expertise from long practice that can help them identify features in a particular patient that may influence the effectiveness of a given treatment for that patient (see Horwitz (1996)). But there is an odd balance struck here. These judgments are deemed admissible in discussion with the individual patient, but they don't add up to evidence to be made publicly available, with the usual cautions about credibility, by the standards adopted by most EBM sites. It is also true that physicians can have preju-

dices and `knowledge' that might be anything but. Clearly, there are situations where forcing practitioners to follow the average will do better, even for individual patients, and others where the opposite is true (see Kahneman and Klein (2009)). Horwitz et al. (2017) propose that medical practice should move from evidence-based medicine to what they call medicine-based evidence in which all individual case histories are assembled and matched to provide a basis for deviation from the means of RCTs.

Whether or not averages are useful to individuals raises the same issue throughout social science research. Imagine two schools, St Joseph's and St. Mary's, both of which were included in an RCT of a classroom innovation. The innovation is successful on average, but should the schools adopt it? Should St Mary's be influenced by a previous attempt in St Joseph's that was judged a failure? Many would dismiss this experience as anecdotal and ask how St Joseph's could have known that it was a failure without benefit of 'rigorous' evidence. Yet if St Mary's is like St Joseph's, with a similar mix of pupils, a similar curriculum, and similar academic standing, might not St Joseph's experience be *more* relevant to what might happen at St Mary's than is the positive *average* from the RCT? And might it not be a good idea for the teachers and governors of St Mary's to go to St Joseph's and find out what happened and why? They may be able to observe the mechanism of the failure, if such it was, and figure out whether the same problems would apply for them, or whether they might be able to adapt the innovation to make it work for them, perhaps even more successfully than the positive average in the trial.

Once again, these questions are unlikely to be easily answered in practice; but, as with exportability, there is no serious alternative to trying. Assuming that the average works for you will often be wrong, and it will at least sometimes be possible to do better; for instance, by judicious use of theory, reasoning by analogy, process tracing, identification of mechanisms, sub-group analysis, or recognizing various symptoms that a causal pathway is possible, as in Bradford-Hill (1965) (see also Cartwright (2015), Reiss (2017), and Humphreys and Jacobs (2017). As in the medical case, the advice to individual schools often lacks specificity. For example, the U.S. Institute of Education Sciences has provided a "user-friendly" guide to practices

supported by rigorous evidence (U.S. Department of Education (2003)). The advice, which is similar to recommendations throughout evidence-based social and health policy literature, is that the intervention be demonstrated effective through well-designed RCTs in more than one site and that "the trials should demonstrate the intervention's effectiveness in school settings similar to yours" (2003, 17). No operational definition of "similar" is provided.

Conclusions

It is useful to respond to two challenges that are often put to us, one from medicine and one from social science. The medical challenge is, "If you are being prescribed a new drug, wouldn't you want it to have been through an RCT?" The second (related) challenge is, "OK, you have highlighted some of the problems with RCTs, but other methods have all of those problems, plus problems of their own." We believe that we have answered both of these in the paper but that it is helpful to recapitulate.

The medical challenge is about *you*, a specific person, so that one answer would be that *you* may be different from the average, and *you* are entitled to and ought to ask about theory and evidence about whether it will work for *you*. This would be in the form of a conversation between *you* and *your* physician, who knows a lot about *you*. *You* would want to know how this class of drug is supposed to work and whether that mechanism is likely to work for *you*. Is there any evidence from other patients, especially patients like *you*, with *your* condition and in *your* circumstances, or are there suggestions from theory? What scientific work has been done to identify what support factors matter for success with this kind of drug? If the only information available is from the pharmaceutical company whose priors and financial interests might have somehow influenced the results, an RCT might seem like a good idea. But even then, and although knowledge of the mean effect among some group is certainly of value, *you* might give little weight to an RCT whose participants are selected in the way they were selected in the trial, or where there is little information about whether the outcomes are relevant to *you*. Recall that many new drugs are prescribed 'off-label', for a purpose for which they were not tested, and beyond that, that many new drugs are administered in the absence of an RCT because *you*

are actually being enrolled in one. For patients whose last chance is to participate in a trial of some new drug, this is exactly the sort of conversation *you* should have with your physician (followed by one asking her to reveal whether you are in the active arm, so that you can switch if not), and such conversations need to take place for *all* prescriptions that are new to you. In these conversations, the results of an RCT may have marginal value. If *your* physician tells *you* that she endorses evidence-based medicine, and that the drug will work for *you because* an RCT has shown that 'it works', it is time to find a physician who knows that *you* and the *average* are not the same.

The second challenge claims that other methods are always dominated by an RCT. That, as one of our referees challenged us, echoing Churchill, "that RCTs are horrible, except when compared to the alternatives." We believe that this challenge is not well-formulated. Dominated for answering what question, for what purposes? The chief advantage of the RCT is that it can, if well-conducted, give an unbiased estimate of an ATE in a study (trial) sample and thus provide evidence that the treatment caused the outcome in some individuals in that sample. Note that 'well-conducted"' rules out all of the things that almost always occur in practice, including attrition, intentional lack of blinding or unintentional unblinding, and other post-randomization confounding and selection biases (see Hernán et al. (2013)). If an unbiased estimate of the ATE is what you want and there's little background knowledge available and the price is right, then an RCT may be the best choice. As to other questions, the RCT result can be part—but usually only a small part—of the defense of (*a)* a general claim, (*b*) a claim that the treatment will cause that outcome for some other individuals, (*c*) a claim about what the ATE will be in some other population, or even (*d*) a claim about something very different that the RCT results tests. But they do little for these enterprises on their own. What is the best overall package of research work for tackling these questions—most cost-effective and most likely to produce correct results—depends on what we know and what different kinds of research will cost.

There are examples where an RCT does better than an observational study, and these seem to be the cases that come to mind for defenders of RCTs. For exam-

ple, regressions of whether people who get Medicaid do better or worse than people with private insurance are vitiated by gross differences in the other characteristics of the two populations. But it is a long step from that to saying that an RCT can solve the problem, let alone that it is the *only* way to solve the problem. It will not only be expensive per subject, but it can only enroll a selected and almost certainly unrepresentative study sample, it can be run only temporarily, and the recruitment to the experiment will necessarily be different from recruitment in a scheme that is permanent and open to the full qualified population. The subjects in the trial are likely to find out whether or not they are in the treatment arm, either because the treatment itself prevents blinding, or because side-effects or differences in protocol reveal their status; subjects may differentially leave the trial given this information. None of this removes the blemishes of the observational study, but there are many methods of mitigating its difficulties, so that, in the end, an observational study with credible corrections and a more relevant and much larger study sample—today often the complete population of interest through administrative records, where blinding and selection issues are absent—may provide a better estimate.

The medical community seems slow and reluctant to embrace other reliable methods of causal inference. The Academy of Medical Sciences (2017, 4) in its review of sources of evidence on the efficacy and effectiveness of medicine agrees with us that "The type of evidence, and the methods needed to analyse that evidence, will depend on the research question being asked." Still, it does not mention methods widely used in social and economic sciences such as instrumental variables, econometric modelling, deduction from theory, causal Bayesian nets, process tracing, or qualitative comparative analysis. Each of these has its strengths and weaknesses, each allows causal inference though not all allow an estimate of effect size, and each—as with every method—requires casual background knowledge as input in order to draw causal conclusions. But in the face of widespread unbinding and the increasing cost of RCTs, it is wasteful not to make use of these. Everything has to be judged on a case -by-case basis. There is no valid argument for a lexicographic preference for RCTs.

There is also an important line of enquiry that goes, not only beyond RCTs, but beyond the 'method of differences' that is common to RCTs, regressions, or any form of controlled or uncontrolled comparison. The hypothetico-deductive method confronts theory-based deductions with the data—either observational or experimental. As noted above, economists routinely use theory to tease out a new implication that can be taken to the data, and there are also good examples in medicine. One is Bleyer and Welch (2012)'s demonstration of the limited effectiveness of mammography screening; the data do not show the compensating changes in early and late stage breast-cancer incidence that would accompany the large-scale introduction of successful screening. This is a topic where RCTs have been indecisive and controversial, if only because they are 20–30 years old and therefore outdated relative to the current rapidly-changing environment (see Marmot et al. (2013)). Such uses of the hypothetico-deductive method are different from what seems to be usually meant by an 'observational study,' in which groups are compared with questionable controls for confounders, and where randomization, in spite of its inadequacies, is arguably better.

RCTs are the ultimate in non-parametric estimation of average treatment effects in trial samples because they make so few assumptions about heterogeneity, causal structure, choice of variables, and functional form. RCTs are often convenient ways to introduce experimenter-controlled variance—if you want to see what happens, then kick it and see, twist the lion's tail—but note that many experiments, including many of the most important (and Nobel Prize winning) experiments in economics, do not and did not use randomization (see Harrison (2013), Svorencik (2015)). But the credibility of the results, even internally, can be undermined by unbalanced covariates and by excessive heterogeneity in responses, especially when the distribution of effects is asymmetric, where inference on means can be hazardous. Ironically, the price of the credibility in RCTs is that we can only recover the mean of the distribution of treatment effects, and that only for the trial sample. Yet, in the presence of outliers in treatment effects or in covariates, reliable inference on means is difficult. And randomization in and of itself does nothing unless the details

are right; purposive selection into the experimental population, like purposive selection into and out of assignment, undermines inference in just the same way as does selection in observational studies. Lack of blinding, whether of participants, investigators, data collectors, or analysts, undermines inference, akin to a failure of exclusion restrictions in instrumental variable analysis.

The lack of structure can be seriously disabling when we try to use RCT results outside of a few contexts, such as program evaluation, hypothesis testing, or establishing proof of concept. Beyond that, the results cannot be used to help make predictions beyond the trial sample without more structure, without more prior information, and without having some idea of what makes treatment effects vary from place to place or time to time. There is no option but to commit to some causal structure if we are to know how to use RCT evidence out of the original context. Simple generalization and simple extrapolation do not cut the mustard. This is true of any study, experimental or observational. But observational studies are familiar with, and routinely work with, the sort of assumptions that RCTs claim to (but do not) avoid, so that if the aim is to use empirical evidence, any credibility advantage that RCTs have in estimation is no longer operative. And because RCTs tell us so little about *why* results happen, they have a disadvantage over studies that use a wider range of prior information and data to help nail down mechanisms.

Yet once that commitment has been made, RCT evidence can be extremely useful, pinning down part of a structure, helping to build stronger understanding and knowledge, and helping to assess welfare consequences. As our examples show, this can often be done without committing to the full complexity of what are often thought of as structural models. Yet without the structure that allows us to place RCT results in context, or to understand the mechanisms behind those results, not only can we not transport whether `it works' elsewhere, but we cannot do one of the standard tasks of economics, which is to say whether the intervention is actually welfare improving. Without knowing *why* things happen and *why* people do things, we run the risk of worthless casual (`fairy story') causal theorizing and have given up on one of the central tasks of economics and other social sciences.

We must back away from the refusal to theorize, from the exultation in our ability to handle unlimited heterogeneity, and actually SAY something. Perhaps paradoxically, unless we are prepared to make assumptions, and to say what we know, making statements that will be incredible to some, the credibility of the RCT does us very little good.

Citations

Abdul Latif Jameel Poverty Action Lab, MIT, (2017). Retrieved August 21, 2017 from: https://www.povertyactionlab.org/about-j-pal

Academy of Medical Sciences (2017). *Sources of evidence for assessing the safety, efficacy, and effectiveness of medicines*. Retrieved from https://acmedsci.ac.uk/file-download/86466482

Aigner, D. J. (1985). The residential electricity time-of-use pricing experiments. What have we learned?. In D. A. Wise & J. A. Hausman (Eds.), *Social experimentation* pp. 11-54). Chicago, Il: Chicago University Press for National Bureau of Economic Research.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *Economic Journal*, *114,* C52–C83.

Angrist, J. D., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: evidence from a randomized natural experiment. *American Economic Review*, *92(5)*, 1535–58.

Aron-Dine, A., Einav, L., & Finkelstein, A. (2013). The RAND health insurance experiment, three decades later. *Journal of Economic Perspectives*, *27(1)*, 197–222.

Arrow, K. J. (1975). Two notes on inferring long run behavior from social experiments. *Document No. P-5546*. Santa Monica, CA: Rand Corporation.

Ashenfelter, O. (1978). The labor supply response of wage earners. In J. L. Palmer & J. A. Pechman (Eds.), *Welfare in rural areas: the North Carolina–Iowa Income Maintenance Experiment* pp. 109-38. Washington, DC: The Brookings Institution.

Attanasio, O., Meghir, C., & Santiago, A. (2012). Education choices in Mexico: using a structural model and a randomized experiment to evaluate PROGRESA. *Review of Economic Studies*, *79(1)*, 37–66.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C., & Rubio-Codina, M. (2015). *Estimating the production function for human capital: results from a randomized controlled trial in Colombia* (Working Paper W15/06). London: Institute for Fiscal Studies.

Bacon, F. (1859). Novum Organum. In Ellis R. L., & Spedding, J. (Eds.), *The Philosophical Works of Francis Bacon*. London, England: Longmans.

Bahadur, R. R., & Savage, L. J. (1956). The non-existence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics*, *25,* 1115–22.

Banerjee, A., Chassang, S., Montero, S., & Snowberg, E. (2017). *A theory of experimenters* (Working Paper 23867). Cambridge, MA: National Bureau of Economic Research.

Banerjee, A., Chassang, S., & Snowberg, E. (2016). *Decision theoretic approaches to experiment design and external validity* (Working Paper 22167). Cambridge, MA: National Bureau of Economic Research.

Banerjee, A. & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, *1,* 151–78.

Banerjee, A. & Duflo, E. (2012). *Poor economics: a radical rethinking of the way to fight global poverty*. New York, NY: Public Affairs.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., & Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348(6236),* 1260799.

Banerjee, A., Karlan, D., & Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, *7(1),* 1–21.

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, *1(1),* 107–34.

Bareinboim, E., & Pearl, J. (2014). Transportability from multiple environments with limited experiments: Completeness results. In Welling, M., Ghahramani, Z., Cortes, C., & Lawrence, N. (Eds.). *Advances in neural information processing systems, 27*, 280–8.

Bauchet, J., Morduch, J., & Ravi, S. (2015). Failure vs displacement: Why an innovative anti-poverty program showed no net impact in South India. *Journal of Development Economics*, *116,* 1–16.

Bechtel, W., (2006). *Discovering cell mechanisms: The creation of modern cell biology*. Cambridge, England: Cambridge University Press.

Begg, C. B. (1990). Significance tests of covariance imbalance in clinical trials. *Controlled Clinical Trials*, *11(4),* 223–5.

Bhattacharya, D., & Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, *167(1),* 168–96.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, *96(4),* 988–1012.

Bleyer, A., & Welch, H. G. (2012). Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine*, *367,* 1998–2005.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2005). Modeling cross-site experimental differences to find out why program effectiveness varies. In Bloom, H. S. (Ed.). *Learning more from social experiments: Evolving analytical approaches*. New York, NY: Russell Sage.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling up what works: Experimental evidence on external validity in Kenyan education* (Working Paper 321). Washington, DC: Center for Global Development.

Bothwell, L. E., & Podolsky, S. H. (2016). The emergence of the randomized, controlled trial. *New England Journal of Medicine*, *375(6),* 501–4.

Cartwright, N., (1994). *Nature's capacities and their measurement*. Oxford, England: Clarendon Press.

Cartwright, N. (2015). *Single case causes: What is evidence and why* (CHESS Working paper 2015-02). Retrieved from: https://www.dur.ac.uk/resources/chess/CHESSWP_2015_02.pdf

Cartwright, N., & Hardie, J. (2012). *Evidence based policy: A practical guide to doing it better*. Oxford, England: Oxford University Press.

Chalmers, I. (2001). Comparing like with like: Some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. *International Journal of Epidemiology*, *30,* 1156–64.

Concato, J. (2013). Study design and `evidence' in patient-oriented research. *American Journal of Respiratory and Critical Care Medicine*, *187(11),* 1167–72.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled, trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, *342(25),* 1887–92.

Conlisk, J. (1973). Choice of response functional form in designing subsidy experiments. *Econometrica*, *41(4),* 643–56.

CONSORT 2010, 15. Baseline data. Retrieved November 9, 2017 from: http://www.consort-statement.org/checklists/view/32--consort-2010/510-baseline-data

Cook, T. D. (2014). Generating causal knowledge in the policy sciences: External validity as a task of both multi-attribute representation and multi-attribute extrapolation. *Journal of Policy Analysis and Management*, *33(2),* 527–36.

Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford, England: Clarendon Press.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2014). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics*, *128(2)*, 531–80.

Davey-Smith, G., & Ibrahim, S. (2002). Data dredging, bias, or confounding. *British Medical Journal*, *325,* 1437–8.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95(450),* 407–24.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48(2),* 424–55.

Deaton, A., & Muellbauer, J. (1980). *Economics and consumer behavior*. New York. NY: Cambridge University Press.

Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (December 3, 2012). Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education. Abdul Latif Jameel Poverty Action Lab, MIT. Retrieved from: http://www.povertyactionlab.org/publication/cost-effectiveness

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, *102(4),* 1241–78.

Duflo, E., & Kremer, M. (2008). Use of randomization in the evaluation of development effectiveness. In Easterly, W. (Ed.). *Reinventing foreign aid* (pp. 93–120). Washington, DC: Brookings.

Dynarski, S. (2015, January 18). Helping the poor in education: The power of a simple nudge. *New York Times*, p BU6. Retrieved from: https://www.nytimes.com/2015/01/18/upshot/helping-the-poor-in-higher-education-the-power-of-a-simple-nudge.html

Epstein, S. (2007). *Inclusion*: *The politics of difference in medical research.* Chicago, Il: Chicago University Press.

Feinstein, A. R., & Horwitz, R. I. (1997). Problems in the `evidence' of `evidence-based medicine. *American Journal of Medicine*, *103,* 529–35.

Fine, P. E. M., & Clarkson, J. A. (1986). Individual versus public priorities in the determination of optimal vaccination policies. *American Journal of Epidemiology*, *124(6),* 1012–20.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33,* 503–13.

Freedman, D. A. (2006). Statistical models for causation: What inferential leverage do they provide?. *Evaluation Review*, *30(6),* 691–713.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40,* 180–93.

Frieden, T. R. (2017). Evidence for health decision making—beyond randomized, controlled trials. *New England Journal of Medicine*, *377,* 465–75.

Garfinkel, I., & Manski, C. F. (1992). Introduction. In Garfinkel, I., & Manski, C. F. (Eds.). *Evaluating welfare and training programs* (pp. 1–22). Cambridge, MA: Harvard University Press.

Gerber, A. S., & Green, D. P. (2012). *Field Experiments*. New York, NY: Norton.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact evaluation in practice* (2nd ed). Washington, DC: Inter-American Development Bank and World Bank.

Greenberg, D., Shroder, M., & Onstott, M. (1999). The social experiment market. *Journal of Economic Perspectives*, *13(3),* 157–72.

Greenland, S. (1990). Randomization, statistics, and causal inference *Epidemiology*, *1(6)*, 421–9.

Greenland, S., & Mansournia, M. A. (2015). Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European Journal of Epidemiology*, *30,* 1101-1110.

Gueron, J. M., & Rolston, H. (2013). *Fighting for reliable evidence*. New York, NY: Russell Sage.

Guyatt, G., Sackett, D. L., & Cook, D. J. (1994). Users' guides to the medical literature II: How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients?. for the Evidence-Based Medicine Working Group, 1994. *Journal of the American Medical Association*, *271(1),* 59–63.

Harrison, G. W. (2013). Field experiments and methodological intolerance. *Journal of Economic Methodology*, *20(2),* 103–17.

Harrison, G. W. (2014a). Impact evaluation and welfare evaluation. *European Journal of Development Research*, *26,* 39–45.

Harrison, G. W. (2014b). Cautionary notes on the use of field experiments to address policy issues. *Oxford Review of Economic Policy*, *30(4),* 753–63.

Heckman, J. J. (1992). Randomization and social policy evaluation. In Manski, C. F., & Garfinkel, I. (Eds.). *Evaluating welfare and training programs* (pp. 547–70). Cambridge, MA: Harvard University Press.

Heckman, J. J., Hohman,N., & Smith, J., with the assistance of Khoo, M. (2000). Substitution and drop out bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics*, *115(2),* 651–94.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economics and Statistics*, *64(4),* 605–54.

Heckman, J. J., Lalonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor markets. In Ashenfelter, O., & Card, D. (Eds.). *Handbook of labor economics* (Vol 3A, pp. 1866–2097). Amsterdam, Netherlands: North-Holland.

Heckman, J. J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, *103(6),* 2052–86.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, Part 1: Causal models, structural models, and econometric policy evaluation. In

Heckman, J. J., & Leamer, E. E. (Eds.). *Handbook of Econometrics* (Vol 6B, pp. 4779–874). Amsterdam, Netherlands: North-Holland.

Hernán, M. A. (2004). A definition of a casual effect for epidemiological research. *Journal of Epidemiology and Community Health*, *58(4),* 265–71.

Hernán, M. A., Hernández-Diaz S., & Robins, J. M. (2013). Randomized trials analyzed as observational studies. *Annals of Internal Medicine*, *159(8),* 560–2.

Hill, A. B. (1965). The environment and disease: Association or causation?. *Proceedings of the Royal Society of Medicine, 58(5),* 295-300.

Horton, R. (2000). Common sense and figures: The rhetoric of validity in medicine. Bradford Hill memorial lecture 1999. *Statistics in Medicine*, *19,* 3149–64.

Horwitz, R. I. (1996). The dark side of evidence based medicine. *Cleveland Clinic Journal of Medicine*, *63(6),* 320–3

Horwitz, R. I., Hayes-Conroy, A., Caricchio, R., & Singer, B. H. (2017). From evidence-based medicine to medicine-based evidence. *American Journal of Medicine*, *130(11),* 1246–50.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experience at other locations. *Journal of Econometrics*, *125,* 241–70.

Howick, J. (2011). *The Philosophy of Evidence-Based Medicine.* Chichester, England. Wiley-Blackwell.

Howick, J., Glasziou, J. P., & Aronson, J. K. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical Medicine and Bioethics*, *34,* 275–91.

Hsieh, C., & Urquiola, M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. *Journal of Public Economics*, *90,* 1477–1503.

Humphreys, M., & Jacobs, A. (2017). *Qualitative inference from causal models*. Draft manuscript (version 0.2). Retrieved November 27, 2017 from: http://www.columbia.edu/~mh2245/qualdag.pdf

Hurwicz, L. (1966). On the structural form of interdependent systems. *Studies in logic and the foundations of mathematics*, *44,* 232–9.

Ilardi, S. S., & Craighead, W. E. (1999). Rapid early response, cognitive modification, and nonspecific factors in cognitive behavior therapy for depression: A reply to Tang and DeRubeis. *Clinical Psychology: Science and Practice*, *6,* 295–99.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86(1),* 4–29.

Imbens, G. W., & Kolesár, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, *98(4),* 701–12.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47(1),* 5–86.

International Committee of Medical Journal Editors, (2015). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. Retrieved August 20, 2016 from: http://www.icmje.org/icmje-recommendations.pdf

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64(6),* 515–26.

Karlan, D., & Appel, J. (2011). *More than good intentions: how a new economics is helping to solve global poverty*. New York, NY: Dutton.

Kasy, M. (2016). Why experimenters might not want to randomize, and what they could do instead. *Political Analysis*, *24(3),* 324-338. doi: 10.1093/pan/mpw012

Kramer, P. (2016). *Ordinarily well: The case for antidepressants*. New York, NY: Farrar, Straus, and Giroux.

Kramer, U., & Stiles, W. B. (2015). The responsiveness problem in psychotherapy: A review of proposed solutions. *Clinical Psychology: Science and Practice*, *22(3),* 277-95.

Kremer, M., & Holla, A. (2009). Improving education in the developing world: What have we learned from randomized evaluations?. *Annual Review of Economics*, *1,* 513–42.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos I., & Musgrave, A. (Eds.). *Criticism and the growth of knowledge: Proceedings of the international colloquium in the philosophy of science, London.* Cambridge, England: Cambridge University Press. doi: 10.1017/CBO9781139171434.00, 91–106.

Lalonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76(4),* 604–20.

Lehman, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed), New York, NY: Springer.

LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, *337,* 536-42.

Levy, S. (2006). *Progress against poverty: sustaining Mexico's Progresa-Oportunidades program*. Washington, DC: Brookings.

Little, D. (2007). *Across the boundaries: Extrapolation in biology and social science.* Oxford, England: Oxford University Press.

Longford, N. T., & Nelder, J. A. (1999). Statistics versus statistical science in the regulatory process. *Statistical Medicine*, *18,* 2311–20.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, *67,* 1–25.

Mackie, J. L., (1974). *The cement of the universe: a study of causation*. Oxford, England: Oxford University Press.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E., & Leibowitz, A. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review*, *77(3),* 251–77.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E., Benjamin, B., Leibowitz, A., Marquis, M. A., & Zwanziger, J. (1988). *Health insurance and the demand for medical care: Evidence from a randomized experiment*. Santa Monica, CA: RAND.

Manski, C. F. (2004). Treatment rules for heterogeneous populations. *Econometrica*, *72(4),* 1221–46.

Manski, C. F. (2013). *Public policy in an uncertain world: Analysis and decisions*. Cambridge, MA: Harvard University Press.

Manski, C. F., & Tetenov, A. (2016). Sufficient trial size to inform clinical practice. *PNAS*, *113(38),* 10518–23.

Marmot, M. G., Altman, D. G., Cameron, D. A., Dewar, J. A., Thomson, S. G., Wilcox M., & the Independent UK panel on breast cancer screening (2013). The benefits and harms of breast cancer screening: An independent review. *British Journal of Cancer*, *108(11),* 2205–40.

Metcalf, C. E., (1973). Making inferences from controlled income maintenance experiments. *American Economic Review*, *63(3),* 478–83.

Moffitt, R. (1979). The labor supply response in the Gary experiment. *Journal of Human Resources*, *14(4),* 477–87.

Moffitt, R. (1992). Evaluation methods for program entry effects. In Manski, C., & Garfinkel, I. (Eds.). *Evaluating welfare and training programs* (pp. 231–52). Cambridge, MA: Harvard University Press.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, *40(2),* 1263–82.

Muller, S. M. (2015). Causal interaction and external validity: Obstacles to the policy relevance of randomized evaluations. *World Bank Economic Review*, *29,* S217–S225.

Orcutt, G. H., & Orcutt, A. G. (1968). Incentive and disincentive experimentation for income maintenance policy purposes. *American Economic Review*, *58(4),* 754–72.

Parkkinen, V-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., Norrell, C., Russo, F., Shaw, B., & Williamson, J. (2008). *Evaluating evidence of mechanisms in medicine: Principles and procedures*. New York, NY: Springer.

Patsopoulos, N. A., (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience*, *13(2),* 217-24.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence* (pp.247-254). Menlo Park, CA: AAAI Press.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29(4),* 579-95.

Pitman, S. R., Hilsenroth, M. J., Goldman, R. E., Levy, S. R., Siegel, D. F., & Miller, R. (2017). Therapeutic technique of APA master therapists: Areas of difference and integration across theoretical orientations. *Professional Psychology: Research and Practice*, *48(3),* 156–66.

Rawlins, M. (2008). *De testimonio:* On the evidence for decisions about the use of therapeutic interventions. *The Lancet*, *372,* 2152–61.

Reichenbach, H. (1954). *Nomological statements and admissible operations*. Amsterdam, Netherlands: North-Holland.

Reichenbach, H. (1976). *Laws, modalities and counterfactuals*, with a foreword by W.C. Salmon. Berkeley and Los Angeles: CA, University of California Press.

Reiss, J. (2017). *Against external validity* (CHESS Working Paper 2017-03). Retrieved from: https://www.dur.ac.uk/resources/chess/CHESSK4UWP_2017_03_Reiss.pdf

Rothman, K. J. (1976). Causes. *American Journal of Epidemiology*, *104 (6),* 587–92.

Rothman, K. J. (2012). *Epidemiology. An introduction* (2nd ed). New York, NY: Oxford University Press.

Rothman, K. J., Gallacher, J. E. J., & Hatch, E. E. (2013). Why representativeness should be avoided. *International Journal of Epidemiology 42(4):* 1012–1014.

Rothwell, P. M. (2005). External validity of randomized controlled trials: 'To whom do the results of the trial apply'. *Lancet*, *365,* 82–93.

Rubin D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association*, *100(469),* 322-331.

Russell, B. (2008) [1912]. *The problems of philosophy*. Rockville, MD: Arc Manor.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal*, *312,* 71–2.

Savage, L. J., (1962). Subjective probability and statistical practice. In Barnard, G. A., & Cox, G. A. (Eds.). *The Foundations of Statistical Inference* (pp. 9–35). London, England: Methuen.

Scriven, M. (1974). Evaluation perspectives and procedures. In Popham, W. J. (Ed.). *Evaluation in education—current applications*. Berkeley, CA: McCutchan Publishing Corporation, 68–84.

Seckinelgin, H. (2017). *The politics of global AIDS: institutionalization of solidarity, exclusion of context*. Social Aspects of HIV, 3. Switzerland: Springer International Publishing.

Senn, S. (2013). Seven myths of randomization in clinical trials. *Statistics in Medicine, 32,* 1439–50.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shepherd, G. M. (1988). *Neurobiology* (2nd ed.). New York, NY: Oxford University Press.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society A*, *174(2),* 369–86.

Student (Gosset, W. S.) (1938). Comparison between balanced and random arrangements of field plots. *Biometrika*, *29(3/4),* 363–78.

Suzuki, E., Yamamoto, E., & Tsuda, T. (2011). Identification of operating mediation and mechanism in the sufficient-component cause framework. *European Journal of Epidemiology*, *26,* 347-57.

Svorencik, A. (2015). *The experimental turn in economics: a history of experimental economics*. Utrecht School of Economics, Dissertation Series #29. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2560026

Todd, P. E., & Wolpin, K. J. (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review*, *96(5),* 1384–1417.

Todd, P. E., & Wolpin, K. J. (2008). Ex ante evaluation of social programs. *Annales d'Economie et de la Statistique*, *91/92,* 263–91.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*, Washington, DC: Institute of Education Sciences.

Van der Weele, T. J. (2012). Confounding and effect modification: Distribution and measure. *Epidemiologic Methods, 1,* 55-82.

Vandenbroucke, J. P. (2004). When are observational studies as credible as randomized controlled trials?. *The Lancet*, *363,* 1728–31.

Vandenbroucke, J. P. (2009). The HRT controversy: Observational studies and RCTs fall in line. *The Lancet*, *373,* 1233–5.

Vittengl, J. R**.,** Clark, L. A., Thase, M. E., & Jarrett, R. B. (2014). Are improvements in cognitive content and depressive symptoms correlates or mediators during Acute- Phase Cognitive Therapy for Recurrent Major Depressive Disorder?. *Inter-*

national Journal of Cognitive Therapy, *7(3),* 255–71. doi:10.1521/ijct.2014.7.3.251.

Vivalt, E. (2016). How much can we generalize from impact evaluations? NYU, unpublished. Retrieved from: http://evavivalt.com/wp-content/uploads/2014/12/Vivalt_JMP_latest.pdf (retrieved, Nov. 28, 2017)

Williams, H. C., Burden-Teh, E., & Nunn, A. J. (2015). What is a pragmatic clinical trial?. *Journal of Investigative Dermatology*, *135(6),* 1-3.

Wise, D. A. (1985). A behavioral model versus experimentation: The effects of housing subsidies on rent. In Brucker, P., & Pauly, R. (Eds.). *Methods of Operations Research* 50 (pp. 441–89). Verlag, Germany: Anton Hain.

Wolpin, K. I. (2013). *The limits of inference without theory*. Cambridge, MA: MIT Press.

Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass*, *2/6,* 981–1022.

Worrall, J. (2008). Evidence and ethics in medicine. *Perspectives in Biology and Medicine*, *51(3),* 418–31.

Young, A. (2017). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results (Working Paper). London School of Economics. Retrieved from: http://personal.lse.ac.uk/YoungA/ChannellingFisher.pdf

Ziliak, S. T. (2014). Balanced versus randomized field experiments in economics: Why W. S. Gosset aka 'Student' matters. *Review of Behavioral Economics*, *1,* 167–208.

Appendix: Monte Carlo experiment for an RCT with outliers

In this illustrative example, there is a parent population each member of which has his or her own treatment effect; these are continuously distributed with a shifted lognormal distribution with zero mean so that the population ATE is zero. The individual treatment effects $\square$ are distributed so that $\square + e^{0.5} \sim \square(0,1)$, for standardized lognormal distribution $\square$. In the absence of treatment, everyone in the sample records zero, so the sample average treatment effect in any one trial is simply the mean outcome among the *n* treatments. For values of *n* equal to 25, 50, 100, 200, and 500 we draw from the parent population 100 trial samples each of size 2*n*; with five values of *n*, this gives us 500 trial samples in all; because

of sampling the true ATE's in each trial sample will not be zero. For each of these 500 samples, we randomize into *n* controls and *n* treatments, estimate the ATE and its estimated *t*–value (using the standard two-sample *t*–value, or equivalently, by running a regression with robust *t*–values), and then repeat 1,000 times, so we have 1,000 ATE estimates and *t*–values for each of the 500 trial samples. These allow us to assess the distribution of ATE estimates and their nominal *t*–values for each trial.

The results are shown in Table A1. Each row corresponds to a sample size. In each row, we show the results of 100,000 individual trials, composed of 1,000 replications on each of the 100 trial (experimental) samples. The columns are averaged over all 100,000 trials.

Table A1: RCTs with skewed treatment effects

Sample size	Mean of ATE estimates	Mean of nominal *t*–values	Fraction null rejected (percent)
25	0.0268	–0.4274	13.54
50	0.0266	–0.2952	11.20
100	–0.0018	–0.2600	8.71
200	0.0184	–0.1748	7.09
500	–0.0024	–0.1362	6.06

Note: 1,000 randomizations on each of 100 draws of the trial sample randomly drawn from a lognormal distribution of treatment effects shifted to have a zero mean.

The last column shows the fractions of times the null that is true in the population is rejected in the trial samples and is our key result. When there are only 50 treatments and 50 controls (row 2), the (true) null is rejected 11.2 percent of the time, instead of the 5 percent that we would like and expect if we were unaware of the problem. When there are 500 units in each arm, the rejection rate is 6.06 percent, much closer to the nominal 5 percent.

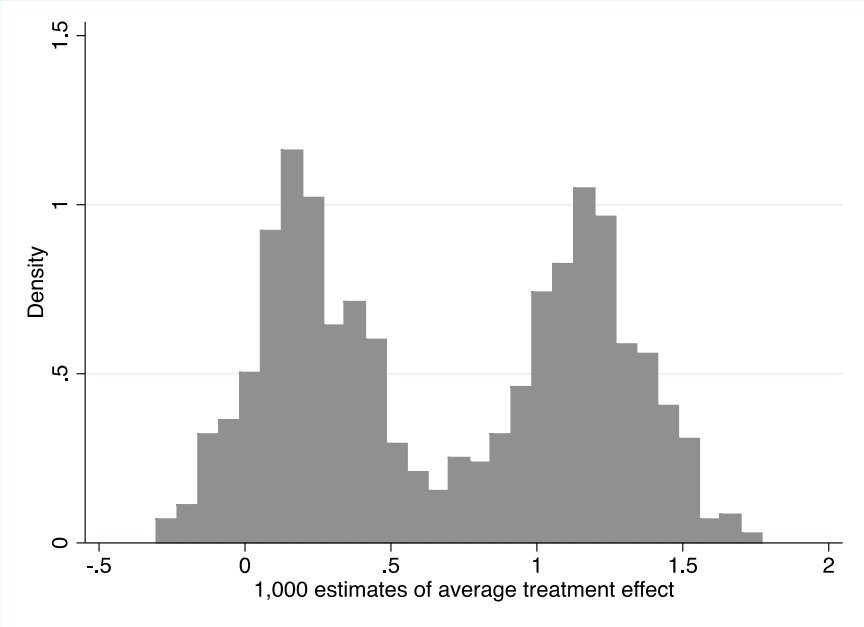

Figure A1: Estimates of an ATE with an outlier in the trial sample

Figure A1 illustrates the estimated ATEs from an extreme trial sample from the simulations in the second row with 100 observations in total; the histogram shows the 1,000 estimates of the ATE for that trial sample. This trial sample has a single large outlying treatment effect of 48.3; the mean (s.d.) of the other 99 observations is –0.51 (2.1); when the outlier is in the treatment group, we get the observations around right-hand mode, when it is in the control group, we get the left-hand mode.