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

This version, October 2017

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. 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

RCTs would be more useful if there were more realistic expectations of them and if their pitfalls were better recognized. For example, and contrary to many claims in the applied literature, randomization does *not* equalize everything but the treatment across treatments and controls, 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) confounders. Estimates apply to the trial sample only, sometimes a convenience sample, and usually selected; justification is required to extend them to other groups, including any population to which the trial sample belongs. Demanding "external validity" is unhelpful because it expects too much of an RCT while undervaluing its contribution. Statistical inference on ATEs involves hazards that are not always recognized. 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 and 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 currently widely visible in economics today and have been used in the subject at least since the 1960s (see Greenberg and Shroder (2004) for a compendium). It is often claimed that such trials can discover "what works" in economics, as well as in political science, education, and social policy. 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 econometric 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.

There are now "What Works" centers using and recommending RCTs in a range of areas of social concern across Europe and the Anglophone world. These centers see RCTs as their preferred tool and indeed often prefer RCT evidence lexicographically. As one of many examples, the US Department of Education's standard for "strong evidence of effectiveness" requires a "well-designed and implemented" RCT; no observational study can earn such a label. This "gold standard" claim about RCTs is less common in economics, but Imbens (2010, 407) writes that "randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top." The Abdul Latif Jameel Poverty Action Lab (J-PAL), whose stated mission is "to reduce poverty by ensuring that policy is informed by scientific evidence", advertises that its affiliated professors "conduct randomized evaluations to test and improve the effectiveness of programs and policies aimed at reducing poverty", J-PAL (2017). The lead page of its website (echoed in the 'Evaluation' section) notes "843 ongoing and completed randomized evaluations in 80 countries" with no mention of any studies that are not randomized.

In medicine, the gold standard view has long been widespread, e.g. for drug trials by the FDA; a notable exception is the recent paper by Frieden (2017), ex-di-

rector 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.

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 knowledge is already available. 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, let alone an 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.

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 terms of bias and efficiency, or expected loss. None of this material is new, but we know of no similar treatment, and we wish to dispute many of the claims that are frequently made in the applied literature. Some routine misunderstandings are: (a) randomization ensures a fair trial by ensuring that, at least with high probability, treatment and control groups differ only in the treatment; (b) RCTs provide not only unbiased estimates of ATEs but also precise estimates; (c) statistical inference in RCTs, which requires only the simple comparison of means, is straightforward, so that standard significance tests are reliable.

Nothing we say in the paper should be taken as a general argument against RCTs; we are simply trying to challenge unjustifiable claims, and expose misunderstandings. We are not against RCTs, only magical thinking about them. The misunderstandings are important because we believe that they contribute to the common perception that RCTs always provide the strongest evidence for causality and for effectiveness.

In the second part of the paper, we discuss how to use the evidence from RCTs. The non-parametric and theory-free nature of RCTs, which is arguably an advantage in estimation, 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. Much of the literature, perhaps inspired by Campbell and Stanley's (1963) famous "primacy of internal validity", appears to believe that internal validity is not only necessary but almost sufficient to guarantee the usefulness of the estimates in different contexts. But 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 the commitment has been made to seeing RCTs 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 not advanced through magical thinking.

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. These issues about how to incorporate prior information into randomized trials are central to Section 1.

In economics, the strengths and weaknesses of RCTs are well explored in the volumes by Hausman and Wise (1985) and by Garfinkel and Manski (1992); in the latter, the introduction by Garfinkel and Manski is a balanced summary of what randomized trials can and cannot do. The paper in that volume by Heckman (1992) raises many of the issues that he and his coauthors have explored in subsequent papers, see in particular Heckman and Smith (1995), and Heckman, Lalonde and Smith (1999) who focus on labor market experiments. Manski (2013) contains a good summary of both strengths and weaknesses.

There is also a more contested recent literature. On the one hand, there are procedures that take as fundamental the unrestricted individual treatment effects of individuals and seek non-parametric approaches to estimating their average. On the other hand, these procedures are contrasted with an approach that uses elements of economic theory to define parameters of interest and to identify magnitudes that are likely to be invariant to policy manipulation or across contexts, where invariance is defined in the sense of Hurwicz (1966). The introduction in Imbens and Wooldridge (2009) provide an eloquent defense of the treatment-effect formulation. It emphasizes the credibility that comes from a theory-free specification with almost unlimited heterogeneity in treatment effects. The introduction in Heckman and Vytlacil (2007) makes an equally eloquent case against, noting that the crucial ingredients of treatments in RCTs are often not clearly specified—so that we often do not know what the treatment really is—and that the treatment effects are hard to link to invariant parameters that would be useful elsewhere. Aspects of the same debate feature in Imbens (2010), Athey and Imbens (2017), Angrist and Pischke (2017), Heckman (2005, 2008, 2010) and Heckman and Urzua (2010).

Deaton (2010) complains about the use of instrumental variables, including randomization, as a substitute for thinking about and constructing models of economic development. He argues against the idea that using RCTs to evaluate projects to discover "what works" can ever yield a systematic body of scientific knowledge that can be used to reduce or eliminate poverty. That paper is an argument against the usefulness of the heterogeneous treatment approach. It argues that refusing to

model heterogeneity, though avoiding assumptions, precludes the sort of cumulative research program that might yield useful policy. The paper's claim that RCTs have no special claim to generate credible and useful knowledge was challenged by Imbens (2010); some of his arguments are answered below. Cartwright (2007) and Cartwright and Munro (2010) challenge any "gold standard" view of RCTs. Cartwright (2011, 2012, 2016) and Cartwright and Hardie (2012) focus on the question of how to use the results of RCTs and what we can learn when an experiment shows that some policy works somewhere. Section 2 pursues these issues in general and through case studies.

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

In this section, we explore how to estimate average treatment effects (ATEs) and the role of randomization. We note first that estimating ATEs is only one of many uses for the data generated by an RCT. 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 trialists. Given random allocation to treatments and controls, the data from the trial allow the identification of two 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 estimated ATE is the difference in means of the two distributions and is the focus of much of the literature in social science and medicine. Yet policy makers and researchers may well be interested in other features of the two distributions. For example, if Y is income, they may be interested in whether a treatment reduced income inequality, or in what it did to the 10^{th} or 90^{th} percentiles of the income distribution, even though different people occupy those percentiles in the treatment and control distributions (see Bitler et al (2006) for an 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) contains an application.) These uses are important, but we focus on ATEs here and do not consider these other uses of RCTs any further in this paper.

1.1 *What does randomization do*?

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

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

where, Y_i is the outcome for unit *i*, 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.* The *x*'s are the 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. Because the heterogeneity of the individual treatment effects, β_i, is unrestricted, we allow the possibility that the treatment interacts with the *x*'s or other variables, so that the effects of *T* can depend on 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.

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

$$\overline{Y}^1 - \overline{Y}^0 = \overline{\beta}^1 + \sum_{j=1}^{J} \gamma_j (\overline{x}_{ij}^1 - \overline{x}_{ij}^0) = \overline{\beta}^1 + (\overline{S}^1 - \overline{S}^0) \quad (2)$$

The first term on the far-right-hand side of (2), which is the ATE in the trial sample, is what we want, 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 average 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? What, exactly, is the role of randomization? 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 statistical, medical, and econometric work. For each subject, a match 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? 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*, subject to the caveat that no correlations of the *x*'s with *Y* are introduced post-randomization, for example by subjects not accepting their assignment. The expectation here 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 $\bar{\beta}^1$ delivers an unbiased estimate of the ATE among the treated in the trial sample, and it does so whether or not the causes are observed. Unbiasedness does not require us to know anything about the other causes though it does require that they not change after randomization so as to make them correlated with the treatment, which is an important caveat to which we shall return. If the RCT is repeated many times on the same trial sample, then, assuming our caveat holds in the trials, the last term in (2) will be zero when averaged over an infinite number of (entirely hypothetical) trials, and the average of the estimated ATEs will be the true ATE in the trial sample. Of course, none of this is true in any one trial where the difference in means will be equal to the average treatment effect among those treated *plus* the term that reflects the 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; by chance the randomization in our single trial 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.

The unbiasedness result can easily be compromised. In particular, the treatment must not be correlated with any other cause. Random assignment is designed to aid with this, but it is not sufficient if, for example, there is lack of blinding so that individuals are aware of their assignment, or if those administering the treatment are so aware, and if that awareness triggers another cause. Similarly, researchers sometimes return to individuals who were randomized years before, so that there has been time for the subjects or others to learn their assignment or for other causes to be influenced by the assignment. This again opens up the possibility of unbalanced effects of causes other than the treatment we are interested in. We have already noted that unbiasedness refers to the trial sample, which may or may not be representative of the population of interest.

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 innovation: not that randomization balanced other causes between treatments and controls but that, conditional on our caveat above, 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. Given the absence of treatment-related post-randomization changes in other causes, randomization yields an unbiased estimate of the ATE in the trial sample as well as a sound method for measuring error of estimation in that sample; therein lies its virtue, not that it yields precise estimates through balance.

1.2 *Misunderstandings: claiming too much*

Everything so far should be perfectly familiar, but exactly what randomization does is frequently lost in the practical 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 do. Randomization is an alternative when we do not know enough, but is generally inferior to good control. 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. These misunderstandings are not so much among the trialists who will often give a correct account when pressed. They come from imprecise statements by trialists that are taken literally by the lay audience that the trialists are keen to reach.

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, Martinez, Premand, Rawlings, and Vermeersch (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.

A similar quote comes from John List, one of the most imaginative and successful scholars who use RCTs:

> "complications that are difficult to understand and control represent key reasons *to conduct* experiments, not a point of skepticism. This is because randomization acts as an instrumental variable, balancing unobservables across control and treatment groups." Al-Ubaydli and List (2013) (italics in the original.)

And from Dean Karlan, founder and President of Yale's *Innovations for Poverty Action*, which runs development RCTs around the world:

> "As in medical trials, we isolate the impact of an intervention by randomly assigning subjects to treatments and control groups. This makes it so that all those other factors which could influence the outcome are present in treatment and control, and thus any difference in outcome can be confidently attributed to the intervention." Karlan, Goldberg and Copestake (2009)

And from the medical literature, 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, p. 18)

Claims are even made that RCTs reveal knowledge without possibility of error. Judy Gueron, the long-time president of MDRC, 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 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 study results apply to different populations (Heckman, Lalonde, and Smith (1999)). In epidemiology, Davey-Smith and Ibrahim (2002) state that "observational studies propose, RCTs dispose". A 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 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), in their standard text for RCTs in political science, write that running an RCT is "a research strategy that does not require, let alone measure, all potential confounders." This is true 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 confounders.

1.3 *Sample size, balance, and precision*

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). Even with very large sample sizes, if there is a large number of causes, balance on *each* cause may be infeasible. Even with just three causes with three values each, there are 27 cells to balance, and in most social and medical cases there will be more. 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 of those billions, only one might be important, and if that one is unbalanced, the results of a single trial can be 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).

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, trialists in economics (and in some other disciplines) usually 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—the observable x's in (1) or interactive factors represented in β—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, unbiasedness is guaranteed by the randomization, whatever the tests show, and as the next paragraph demonstrates, the test is not informative about the balance that would lead to *precision*.

If we write μ^0 and μ^1 for the (vectors of) true means in the trial sample (i.e. the means over all possible randomizations) of the observed causes of Y in the control and treatment groups at the point of assignment, the null hypothesis is (presumably, as judged by the typical balance test) that the two vectors are identical, with the alternative being that they are not. But if the randomization has been correctly done the null hypothesis is true *by construction* (see e.g. Altman (1985) and Senn (1994)), which may help explain why it so rarely fails in practice. As Begg (1990) 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, aa false positive." This is, of course, consistent with Fisher's comments about the plots in the field, which notes that two samples of plots randomly drawn from the same field can look very unbalanced. Indeed, although we cannot "test" it in this way, we know that the null hypothesis is also true for the *unobservable* causes. Note the contrast with the statement quoted above claiming that RCTs guarantee balance on causes across treatment and control groups. Those statements refer to balance of causes at the point of assignment in *any single trial*, which is not guaranteed by randomization, whereas the balance tests are about the balance of causes at the point of assignment *in expec-*

tation over many trials, which *is* guaranteed by randomization. The confusion is perhaps understandable, but it is a confusion nevertheless. 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 covariates. But as Scriven (1974, 103) 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 of 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, Chassang, and Snowberg (BCS) (2016) and Banerjee, Chassang, Montero, and Snowberg (BCMS) (2016).

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) 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 (2016) 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 that case study methodology is well attuned to. As BCS note, this allocation would not persuade those with different priors, and they propose randomization as a means of satisfying skeptical observers.

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 the optimal design that is most informative. According to Chalmers (2001) and Bothwell and Podolsky (2016), the development of randomization 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.

Second, the ideal rules by which units are allocated to treatment or control depend on the covariates and on the investigators' priors about how the covariates affect the outcomes. This opens up all sorts of methods of inference that are long familiar to economists 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 economists when they use theory to derive predictions that can be tested against the data, perhaps in an RCT, but more frequently not. Some of the most fruitful research programs in economics have been generated by the puzzles that result when the data fail to

match such theoretical predictions, such as the equity premium puzzle, various purchasing power parity puzzles, the Feldstein-Horioka puzzle, the consumption smoothness puzzle, the puzzle of calorie decline in the face of malnourishment and income growth, and many others.

Third, randomization, by running roughshod over 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, 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. 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 (2016) 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)).

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, as in many pharmaceutical trials, randomization will indeed be an appropriate response. But we believe such arguments are 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 in the face of aggressive ignorance. The systematic refusal to use prior knowledge and the associated preference for RCTs are recipes for preventing cumulative scientific progress. In the end, it is also self-defeating. To quote Rodrik (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. These procedures do not always give the right answer.

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 shall 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 must be a consequence of the particular random allocation that generated it. 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 there are no effects of treatment. 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). Imbens (2010) argues that it is this randomization inference plus the unbiasedness of the ATE that provides the twin non-parametric pillars that support placing RCTs at the "very top" of the hierarchy of evidence.

Randomization inference can be used for null hypotheses that specify 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 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 usual t–statistic 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 very 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 without covariates, 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

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

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 (2016) 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 controls and when using clustered or robust standard errors. When the design matrix is such that the maximal influence is large, 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 2,003 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. In addition to Young (2016), 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 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, p. 466–8)). How difficult is it to maintain symmetry? And how badly is inference affected when the distribution of treatment effects is not symmetric? In economics, 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 be true of the treatment effects too 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)). Another 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, Newhouse 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 a structural approach that is explicitly designed to model the skewness of expenditures.

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 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. We have a microfinance trial in mind, where there is a long positive tail of rare individuals who can do amazing things with credit while most people cannot use it 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 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 development economics suggests that they are not exempt from these concerns. Many development 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 in terms of economic theory. 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. No doubt there are behaviors in the world that are inconsistent with standard 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.

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, care must be taken (e.g. by blinding) that there are no significant post-randomization correlates with the treatment. This is a well-known lesson but many social and economic trials are not blinded and insufficient defense is offered that unbiasedness is not undermined. Indeed, lack of blinding is not the only source of post-randomization bias. 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. 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.

Beyond that, in many cases, 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 an argument 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 in economics, as indeed in medicine and in social policy, has paid more attention to obtaining results than to considering what can be done with them. There is little theoretical or empirical work to guide us how and for what purposes to use the findings of RCTs, such as the conditions under which the same results hold outside of the original settings, how they might be adapted for use elsewhere, or 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, but it 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. An 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 see Horton (2000) for a strong dissent and Rothwell (2005) for examples 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.

2.2 *Using results, transportability, and external validity*

Suppose a trial has 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 transportability 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 a trial 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 more useful applications of the results. The failure of external validity interpreted as simple generalization or extrapolation says little about the value of the trial.

First, there are several uses of RCTs that do not require transportability beyond the original context; we discuss these in the next subsection. Second, 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 or to support for 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 extremely useful 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.

Bertrand Russell's chicken (Russell (1912)) provides an excellent example of the limitations to straightforward 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 understand the social and economic structure that gave rise to the causal relations that she observed.

So, establishing *causality* does nothing in and of itself to guarantee *generalizability*. 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 in economics to the discovery and labeling of causal relations, a weight that is hard to justify for effects that may have only local applicability, what might be labeled 'anecdotal causality'. The operation of a cause generally requires the presence of "support 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 often the kind of causality we see. 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. This is standard fare in epidemiology, which uses the term `causal pie' to refer to a set of causes that are jointly but not separately sufficient for an effect.

If we rewrite (1) in the form

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

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, and indeed we do generally find different ATEs in different economic (and other social policy) RCTs in different places even in the cases where (unusually) they all point in the same direction.

Causal processes often require highly specialized economic, cultural, or social structures to enable them to work. 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. Neither the same ATE nor the same qualitative causal relations can be expected to hold where the specific form for (4) is different. Indeed, we continually attempt to design 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.

As in the Rube Goldberg machine and in the design of cars and work schedules, the economic structure and equilibrium may differ in ways that support 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.

2.3 *When RCTs speak for themselves: no transportability 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*. This is one of the arguments for the importance of internal validity.

Nor is transportation 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 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. When the components of treatments change across studies, evaluations need not lead to cumulative knowledge. Or

as Heckman et al (1999, 1934) note, "the data produced from them [social experiments] are far from ideal for estimating the structural parameters of behavioral models. This makes it difficult to generalize findings across experiments or to use experiments to identify the policy-invariant structural parameters that are required for econometric policy evaluation."

Of course, when we ask exactly what those invariant structural parameters are, whether they exist, and how they should be modeled, we open up major fault lines in modern applied economics. For example, we do not intend to endorse intertemporal dynamic models of behavior as the only way of recovering the parameters that we need. We also recognize that the usefulness of simple price theory is not as universally accepted as it once was. But the point remains that we need something, some regularity or some 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 scaling up 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 will be common in economic examples but perhaps less so in health. There is also an issue of timing if time elapses between the trial and the implementation.

In economics, a `public-health-style' example is the imposition of a commodity tax, where the total tax revenue is of interest and policymakers do not care who pays the tax. Indeed, theory can often identify a specific, well-defined quantity whose measurement is key for a policy (see Deaton and Ng (1998) for an example of what Chetty (2009) calls a "sufficient" statistic). In this case, the behavior of a random sample of individuals might well provide a good guide to the tax revenue that

can be expected. Another case comes from work on poverty programs where the sponsors are most concerned about the budget; we discuss these cases at the end of this Section. Even here, it is easy to imagine behavioral effects coming into play that drive a wedge between the trial and its full-scale implementation, for example if compliance is higher when the scheme is widely publicized, or if government agencies implement the scheme differently from trialists.

2.4 *Transporting results laterally and globally*

The program of RCTs in economics, as in other areas of social science, has the broader goal of finding out `what works.' At its most ambitious, this aims for universal reach, and the development economics literature frequently argues 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", Duflo and Kremer (2008, 93). Sometimes the results of a single RCT are advocated as having wide applicability, with especially strong endorsement when there is at least one replication. For example, Kremer and Holla (2009, 3) use a Kenyan trial as the basis for a blanket statement without specifying context, "Provision of free school uniforms, for example, leads to 10%-15% reductions in teen pregnancy and dropout rates." Duflo and Kremer (2008, 104), writing about another trial, are more cautious, citing two evaluations and restricting themselves to India: "One can be relatively confident about recommending the scaling-up of this program, at least in India, on the basis of these estimates, since the program was continued for a period of time, was evaluated in two different contexts, and has shown its ability to be rolled out on a large scale." Even a number of replications do not provide a sound basis for inference. Without theory 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.

The problem of generalization extends 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 transportation is far from automatic also

tells us why (even ideal) RCTs of similar interventions give different answers in different settings. Such differences do not necessarily reflect methodological failings and will hold across perfectly executed RCTs just as they do across observational studies.

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, not of how much health insurance increased expenditures 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 (originally known as the Manpower Development Research Corporation) have also been analyzed across cities in an effort to link city features to the results of the RCTs within them (see Bloom, Hill, and Riccio (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.

Although there are issues with all of methods of investigating differences across trials, without some discipline it is too easy to come up with `just-so' or fairy stories that account for differences. We risk a procedure that, if a result is replicated in full or in part in at least two places, puts that treatment into the `it works' box and, if the result does not replicate, casually interprets the difference in a way that allows at least some of the findings to survive.

How can we do better than simple generalization and simple extrapolation? Many writers emphasize the role of *theory* in transporting and using the results of trials, and we shall discuss this in the next subsection. But statistical approaches are also widely used; these are designed to deal with the possibility that treatment effects vary systematically with other variables. Referring back to (4), it is clear that, supposing the same form of (4) obtains, if the distribution of the *w* values is the same in the new circumstances as in the old, the ATE in the original trial will hold in the new circumstances. In general, of course, this condition will not hold, nor do we have any obvious way of checking it unless we know what the support factors are in both places. One procedure to deal with interactions 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 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, Imbens, and Mortimer (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. 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. As Hotz et al (2005) note, it will typically be necessary to rule out such `macro' effects, whether over time, or over locations. Third, it also depends on assuming that the same governing equation (4) covers 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, 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. While it is true that general causal claims exist— that gravitational masses attract each other, or that people respond to incentives—these use relatively abstract concepts and operate at a much higher level than the claims that can be reasonably inferred from a typical RCT.

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 transportation. 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.

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. 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, we will typically 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 the results are likely to be exportable to. 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 description 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.

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. (For example, the ICMJE guidelines (2015, 14) go on to say: "Because the relevance of such variables as age, sex, or ethnicity is not always known at the time of study design, researchers should aim for inclusion of representative populations into all study types and at a minimum provide descriptive data for these and other relevant demographic variables," p14.) 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 populations 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.

2.5 *Using theory for generalization*

Economists have been combining theory and randomized controlled trials since the early experiments. 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, Pinto and Savelyev (2013) for the Perry pre-school program. Development economics examples include Attanasio, Meghir, and Santiago (2012), Attanasio et al (2015), Todd and Wolpin (2006), Wolpin (2013), and Duflo, Hanna, and Ryan (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.

Light touch 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 to transport and adapt results from one setting to 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.

Theory can often allow us to reclassify new or unknown situations as analogous to situations where we already have background knowledge. One frequently useful way of doing this is when the new policy can be recast as equivalent to a change in the budget constraint that respondents face. 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 relevantly 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 trialists 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 problem we raised of delineating the population to which the trial results immediately apply and for thinking about moving from this population to populations of interest. Ashenfelter's (1978) analysis is again a good illustration and predates much similar work in later literature. The income tax experiments offered participation in the trial to a random sample of the population of interest. Because there was no blinding and no compulsion, people who were randomized into the treatment group were free to *choose to refuse treatment*. As in many subsequent analyses, Ashenfelter supposes that people choose to

participate if it is in their interest to do so, depending on what has become known in the RCT and Instrumental Variables literature as their own idiosyncratic `gain.' The simple labor supply model gives an approximate condition: If the treatment increases the tax rate from t_0 to t_1 with an offsetting increase in *G*, then an individual assigned to the experimental group will decline to participate if

$$(t_1 - t_0)w_0h_0 + \frac{1}{2}s_{00}(t_1 - t_0) > G_1 - G_0 \qquad (5)$$

where subscript 1 refers to the treatment situation, 0 to the control, h_0 is hours worked, and s_{00} is the (negative) utility-constant response of hours worked to the tax rate. If there is no substitution, the second term on the left-hand side is zero, and people will accept treatment if the increase in *G* more than makes up for the increases in taxes payable, the `breakeven' condition. In consequence, those with higher earnings are less likely to accept treatment. Some better-off people with high substitution effects will also accept treatment if the opportunity to buy more cheap leisure is sufficient enticement.

The selective acceptance of treatment limits the analyst's ability to learn about the better-off or low-substitution people who decline treatment but who would have to accept it if the policy were implemented. Both the intention-to-treat estimator and the `as treated' estimator that compares the treated and the untreated are affected, not just by the labor supply effects that the trial is designed to induce, but by the kind of selection effects that randomization is designed to eliminate. Of course, the analysis that leads to (5) can perhaps help us say something about this and help us adjust the trial estimates back to what we would like to know. Yet this is no easy matter because selection depends, not only on observables, such as pre-experimental earnings and hours worked, but on (much harder to observe) labor supply responses that likely vary across individuals. Paraphrasing Ashenfelter, we cannot estimate the effects of a permanent compulsory negative income tax program from a transitory voluntary trial without strong assumptions or additional evidence.

Much of the modern literature, for example on training programs, wrestles with the issue of exactly who is represented by the RCT results, including not only

who participates in the first place but who leaves before the trial is completed (see again Heckman, Lalonde and Smith (1999)). As in the examples above, modeling attrition within a trial can yield estimates of behavioral responses that can be used to transport the findings to other settings (see Chan and Hamilton (2006), Chassang, Padró I Miguel, and Snowberg (2012) and Chassang et al (2015)). When people are allowed to reject their randomly assigned treatment according to their own (real or perceived) advantage, or to drop out of a trial on an estimate of the benefits and costs from doing so, we have come a long way away from the random allocation in the standard conception of a randomized controlled trial. Moreover, the absence of blinding is common in social and economic RCTs, and while there are trials, such as welfare trials, that effectively compel people to accept their assignments, and some where the treatment is generous enough to do so, there are trials where subjects have much freedom and, in those cases it is less than obvious to us what role, if any, randomization plays in warranting the results.

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. 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). 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. Of course, these considerations might not be decisive in deciding whether or not to promote the innovation, and there may still be long term gains if, for example, some farmers find something better to do than growing cocoa. But, 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. 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 (2002)). 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.

Scaling up can also disturb the political equilibrium. An exploitative government may not allow the mass transfer of money from abroad to a powerless segment of the population, though it may permit a small-scale RCT of cash transfers, perhaps even in the hope that a large-scale implementation will yield opportunities for predation. Provision of healthcare by foreign NGOs may be successful in trials, but have unintended negative consequences to scale because of general equilibrium

effects on the supply of healthcare personnel, or because it disturbs the nature of the contract between the people and a government that is using tax revenue to provide services. In India, the government spends large sums on food subsidies through a system (the PDS) that is both corrupt and inefficient, with much of the grain that is procured failing to find its way to the intended beneficiaries. Localized RCTs on whether or not families are better off with cash transfers are not informative about how politicians would change the amount of the transfer if faced with unanticipated inflation, and at least as important, whether the government could cut procurement from relatively wealthy and politically powerful farmers. Without a political and general equilibrium analysis, it is impossible to think about the effects of replacing food subsidies with cash transfers (see e.g. Basu (2010)).

Even in medicine, where biological interactions between people are less common than are social interactions in social science, interactions can be important. Infectious diseases are one 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.

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. 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 prejudices 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, Kahneman and Klein (2009).

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 transportability, 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. 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, US Department of Education (2003). The advice, which is similar to recommendations in development economics, 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.

2.8 *Examples and illustrations from economics*

Our arguments in this Section should not be controversial, yet we believe that they represent an approach that is different from most current practice. To document this and to fill out the arguments, we provide some examples. While these are occasionally critical, our purpose is constructive; indeed, we believe that misunderstandings about how to use RCTs have artificially limited their usefulness, as well as alienated some who would otherwise use them.

Conditional cash transfers (CCTs) are interventions that have been tested using RCTs (and other RCT-like methods) and are often cited as a leading example of how an evaluation with strong internal validity leads to a rapid spread of the policy, e.g. Angrist and Pischke (2010) among many others. 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 (2001), nor in places where people strongly oppose education or vaccination.

Similarly, given that the support factors will operate with different strengths and effectiveness in different places, it is also 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 (2015). 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.

The range of estimates illustrates that the simple view of external validity—that the ATE transports from one place to another—is not reasonable. 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 additional months of schooling per $100 spent (for example if a donor is trying to decide where to spend, see below). 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 transportability 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.)

Much of the economics literature, like the medical literature, works with the view of external validity that, unless there is evidence to the contrary, the direction and size of treatment effects can be transported from one place to another. The J-PAL website reports its findings under a general heading of policy relevance, subdivided by a selection of topics. Under each topic, there is a list of relevant RCTs from a range of different settings around the world. These are conveniently converted

into a common cost-effectiveness measure so that, for example, under "education", subhead "student participation", there are four studies from Africa: on informing parents about the returns to education in Madagascar, on deworming, on school uniforms, and on merit scholarships, all from Kenya. The units of measurement are additional years of student education per $100, and among these four studies, the average effects of spending $100 are 20.7 years, 13.9 years, 0.71 years and 0.27 years respectively. (Note that this is a different—and much superior—standardization from the effect size standardization discussed below.)

What can we conclude from such comparisons? For a philanthropic donor interested in education, and if marginal and average effects are the same, they might indicate that the best place to devote a marginal dollar is in Madagascar, where it would be used to inform parents about the value of education. This is certainly useful, but it is not as useful as statements that information or deworming programs are everywhere more cost-effective than programs involving school uniforms 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 we can transport the results from one place to another, if the Kenyan results also hold in Madagascar, Mali, or Namibia, or some other list of places. 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 play an equal 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 transported as is. All of this lacks a justification for transportability, some understanding of when results transport, when they do not, or better still, how they should be modified to make them transportable.

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, Morduch, and Ravi (2015) fail to replicate the result in South India, where the control group got access to much the same benefits, what Heckman, Hohman, and Smith (2000) call "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.

A number of difficulties remain. As the authors note, such trials cannot tell us which component of the treatment accounted for the results, or which might be dispensable—a much more expensive multifactorial trial would be required—though it seems likely in practice that the costliest component—the repeated visits for training and support—is likely to be the first to be cut by cash-strapped politicians or administrators. And as noted, it is not clear what should count as (simple) replication in international comparisons; it is hard to think of the uses of standardized effect sizes, except to document that effects exist everywhere and that they are similarly large relative to local variation in such things.

The effect size—the ATE standardized by being expressed in numbers of standard deviations of the original outcome—though conveniently dimensionless, has little to recommend it. As with much of RCT practice, it strips out any economic content—no rates of return, or benefits minus costs—and it removes any discipline on what is being compared. Apples and oranges become immediately comparable, as do treatments whose inclusion in a meta-analysis is limited only by the imagination of the analysts in claiming similarity. Training programs for physical fitness can be pooled with training programs for welding, or marketing, or even obedience training for pets. In psychology, where the concept originated, this results in endless disputes about what should and should not be pooled in a meta-analysis. Goldberger and Manski (1995, 769) note that "standardization accomplishes nothing except to give quantities in noncomparable units the superficial appearance of being in comparable units. This accomplishment is worse than useless—it yields misleading inferences." Beyond that, Simpson (2017) notes that restrictions on the trial sample—often good practice to reduce background noise and to help detect an effect—will reduce the baseline standard deviation and inflate the effect size. More generally, effect sizes are open to manipulation by exclusion rules. It makes no sense to claim replicability on the basis of effect sizes, let alone to use them to rank projects. Effect sizes are irrelevant for policymaking.

The graduation study can be taken as the closest to fulfilling the `finding out what works' aim of the RCT movement in development. Yet it is silent on perhaps the crucial aspect for policy, which is that the trial was run in partnership with NGOs, whereas what we would like to know is whether it could be replicated by governments, including those governments that are incapable of getting doctors, nurses, and teachers to show up to clinics or schools, Chaudhury et al (2005), Banerjee, Deaton and Duflo (2004), or of regulating the quality of medical care in either the public or private sectors, Filmer, Hammer and Pritchett (2000) or Das and Hammer (2005). In fact, we already know a great deal about `what works.' Vaccinations work, maternal and child healthcare services work, and classroom teaching works. Yet knowing this does not get those things done. Adding another program that works under ideal conditions is useful only where conditions are in fact ideal, in

which case it would likely be unnecessary. Finding out what works is not the magic key to economic development. Technical knowledge, though always worth having, requires suitable institutions and suitable incentives if it is to do any good.

A similar point is documented in the contrast between a successful trial that used cameras and threats of wage reductions to incentivize attendance of teachers in schools run by an NGO in Rajasthan in India, Duflo, Hanna, and Ryan (2012), and the subsequent failure of a follow-up program in the same state to tackle mass absenteeism of health workers, Banerjee, Duflo, and Glennerster (2008). In the schools, the cameras and timekeeping worked as intended, and teacher attendance increased. In the clinics, there was a short-run effect on nurse attendance, but it was quickly eliminated. (The ability of agents eventually to undermine policies that are initially effective is common enough and not easily handled within an RCT.) In both trials, there were incentives to improve attendance, and there were incentives to find a way to sabotage the monitoring and restore workers to their accustomed positions; the force of these incentives is a `high-level' cause, like gravity, or the principle of the lever, that works in much the same way everywhere. For the clinics, some sabotage was direct—the smashing of cameras—and some was subtler, when government supervisors provided official, though specious, reasons for missing work. We can only conjecture why the causality was switched in the move from NGO to government; we suspect that working for a highly-respected local NGO is a different contract from working for the government, where not showing up for work is widely (if informally) understood to be part of the deal. The incentive lever works when it is wired up right, as with the NGOs, but not when the wiring cuts it out, as with the government. Knowing `what works' in the sense of the treatment effect on the trial population is of limited value without understanding the political and institutional environment in which it is set. This underlines the need to understand the underlying social, economic, and cultural structures—including the incentives and agency problems that inhibit service delivery—that are required to support the causal pathways that we should like to see at work.

Trials in economic development often take place in artificial environments. Drèze (2016) notes, based on extensive experience in India, "when a foreign agency

comes in with its heavy boots and suitcases of dollars to administer a `treatment,' whether through a local NGO or government or whatever, there is 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 work do so with the people who will work it in practice.

There is also much to be learned from many years of economic trials in the United States, particularly from the work of MDRC, from the early income tax trials, as well as from the Rand Health Experiment. Following the income tax trials, MDRC has run many randomized trials since the 1970s, mostly for the Federal government but also for individual states and for Canada (see the thorough and informative account by Gueron and Rolston (2011) for the factual information underlying the following discussion). MDRC's program, like that of JPAL in development, is intended to find out `what works' in the state and federal welfare programs. These programs are conditional cash transfers in which poor recipients are given cash provided they satisfy certain conditions such as work requirements or training, which are often the subject of the trial. What are the benefits and costs of various alternatives, both to the recipients and to the local and federal taxpayers? All of these programs are deeply politicized, with sharply different views over both facts and desirability. Many engaged in these disputes feel certain of what should be done and what its consequences will be so that, by their lights, control groups are unethical because they deprive some people of what the advocates `know' will be certain benefits. Given this, it is perhaps surprising that RCTs have become the accepted norm for this kind of policy evaluation in the US.

The reasons owe much to political institutions, as well as to the common belief, explored in Section 1, that RCTs can reveal the truth. At the Federal level, prospective policies are vetted by the non-partisan Congressional Budget Office (CBO), which makes its own estimates of the budgetary implications of the program. 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. And control groups are

easier to put in place when there are insufficient funds to cover the whole population. There was also a widespread and largely uncritical belief that RCTs give the right answer, at least for the budgetary implications, which, rather than the wellbeing of the recipients, were often the primary concern; note that all of these trials are on poor people by rich people who are typically more concerned with cost than with the wellbeing of the poor, Greenberg, Schroder and Onstott (1999). MDRCs trials could therefore be effective dispute-reconciliation mechanisms both for those who saw the need for evidence and for those who did not (except instrumentally). The outcome here fits with our `public health' case; what the politicians need to know is not the outcomes for individuals, or even how the outcomes in one state might transport to another, but the average budgetary cost in a specific place, something that a good RCT conducted on a representative sample of the target population can deliver, at least in the absence of general equilibrium effects, timing effects, etc.

These RCTs by MDRC and other contractors have demonstrated both the feasibility of large-scale social trials including the possibility of randomization in these settings (where many participants were hostile to the idea), as well as their usefulness to policymakers. They also seem to have changed beliefs, for example in favor of the desirability of work requirements as a condition of welfare, even among many originally opposed. There are also limitations; the trials appear to have had at best a minor influence on scientific thinking about behavior in labor markets and, in that sense, they are more about `plumbing' than science, Duflo (2017). The results of similar programs have often been different across different sites, and there has to date been no firm understanding of why; indeed, the trials are not designed to reveal this, Moffitt (2004). Finally, and perhaps crucially for the potential contribution to economic science, there has been little success in understanding either the underlying structures or chains of causation, in spite of a determined effort from the beginning to open the black boxes.

The RAND health experiment, Manning et al (1975a, b), provides a different but equally 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, Einav, and Finkelstein (2013), it is this dimensionless and thus apparently transportable 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 it is even unclear that it is firmly based on the original evidence. This points, once again, to the central importance of transportability for the usefulness, both short and long term, of a trial. Here, the simple direct transportability of the result seems to have been largely illusory though, as we have argued, this does not mean that more complex constructions based on the results of the trial would not have done better.

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, 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 most likely work for *you because* an RCT has shown that 'it works', it is time to find a new physician.

The second challenge claims that other methods are always dominated by an RCT. This kind of 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. If that is what you want to know 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, or even (*c*) a claim about what the ATE will be in some other population. 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 example, 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. 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—may provide a better estimate. Everything has to be judged on a case -by-case basis. There is no rigorous 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 such as Bleyer and Welch (2012)'s demonstration of the limited impact on breast cancer incidence of mammography screening, a topic where other methods have generated great controversy and little consensus.

RCTs are the ultimate in non-parametric estimation of average treatment effects in the 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, 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, 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, trialists, 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 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 the standard stuff 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.

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, all the credibility of the RCT is for naught.

RCTs in economics on health, labor, and development have proven their worth in providing proofs of concept and at testing predictions that some policies must always work or can never work. But, as elsewhere in economics, we cannot find out *why* something works by simply demonstrating that it *does* work, no matter how often, which leaves us uninformed as to whether the policy *should* be implemented. Beyond that, small scale, demonstration RCTs are not capable of telling us what would happen if these policies were implemented to scale, of capturing unintended consequences that typically cannot be included in the protocols, or of modeling what will happen if schemes are implemented differently than in the trial, for example by governments, whose motives and operating principles are different from the NGOs or academics who typically run trials. While it is true that abstract knowledge is always likely to be beneficial, successful policy depends on institutions and on politics, matters on which RCTs have little to say. The results of RCTs can and should feed into public debate about what should be done, but we are on dangerous ground when they are used, on grounds of their supposed epistemic superiority, to insulate policy from democratic processes.

Citations

Aigner, Dennis J., 1985, "The residential electricity time-of-use pricing experiments. What have we learned?" in David A. Wise and Jerry A. Hausman, *Social experimentation*, Chicago, Il. Chicago University Press for National Bureau of Economic Research, 11–54.

Al-Ubaydil, Omar, and John A. List, 2013, "On the generalizability of experimental results in economics," in G. Frechette and A. Schotter, *Methods of modern experimental economics,* Oxford University Press.

Altman, Douglas G., 1985, "Comparability of randomized groups," *Journal of the Royal Statistical Society, Series D (The Statistician)*, 34(1), Statistics in health, 125–36.

Angrist, Joshua D., 2004, "Treatment effect heterogeneity in theory and practice," *Economic Journal*, 114, C52–C83.

Angrist, Joshua D., Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer, 2002, "Vouchers for private schooling in Colombia: evidence from a randomized natural experiment," *American Economic Review*, 92(5), 1535–58.

Angrist, Joshua D. and Jörn-Steffen Pischke, 2010, "The credibility revolution in empirical economics: how better research design is taking the con out of econometrics," *Journal of Economic Perspectives*, 24(2), 3–30.

Angrist, Joshua D. and Jörn-Steffen Pischke, 2017, "Undergraduate econometrics instruction: through our classes, darkly," *Journal of Economic Perspectives*, 31(2), 125–44.

Aron-Dine, Aviva, Liran Einav, and Amy Finkelstein, 2013, "The RAND health insurance experiment, three decades later," *Journal of Economic Perspectives*, 27(1), 197–222.

Arrow, Kenneth J., 1975, "Two notes on inferring long run behavior from social experiments," Document No. P-5546, Santa Monica, CA. Rand Corporation.

Ashenfelter, Orley, 1978, "The labor supply response of wage earners," in John L. Palmer and Joseph A. Pechman, eds., *Welfare in rural areas: the North Carolina–Iowa Income Maintenance Experiment*, Washington, DC. The Brookings Institution. 109–38.

Athey, Susan and Guido W. Imbens, 2017, "The state of applied econometrics: causality and policy evaluation," *Journal of Economic Perspectives*, 31(2), 3–32.

Attanasio, Orazio, Costas Meghir, and Ana Santiago, 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, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio Codina, 2015, "Estimating the production function for human capital: results from a randomized controlled trial in Colombia," London. Institute for Fiscal Studies, Working Paper no W15/06.

Bahadur, R. R., and Leonard J. Savage, 1956, "The non-existence of certain statistical procedures in nonparametric problems," *Annals of Mathematical Statistics*, 25: 1115–22.

Banerjee, Abhijit, Sylvain Chassang, Sergio Montero, and Erik Snowberg, 2016, "A theory of experimenters," processed, July 2016.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg, 2016, "Decision theoretic approaches to experiment design and external validity," Cambridge, MA. NBER Working Paper no 22167, April.
Banerjee, Abhijit, Angus Deaton, and Esther Duflo, 2004, "Healthcare delivery in rural Rajasthan," *Economic and Political Weekly*, 39(9), 944–9.
Banerjee, Abhijit and Esther Duflo, 2009, "The experimental approach to development economics," *Annual Review of Economics*, 1, 151–78.
Banerjee, Abhijit and Esther Duflo, 2012, *Poor economics: a radical rethinking of the way to fight global poverty*, Public Affairs.
Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry, 2015, "A multifaceted program causes lasting progress for the very poor: evidence from six countries," *Science*, 348 (6236), 1260799.
Banerjee, Abhijit, Esther Duflo, and Rachel Glennerster, 2008, "Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system," *Journal of the European Economic Association*, 6(2–3), 487–500.
Banerjee, Abhijit V. and Ruimin He, 2003, "The World Bank of the future," *American Economic Review*, 93(2), 39–44.
Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman, 2015, "Six randomized evaluations of microcredit: introduction and further steps," *American Economic Journal: Applied Economics*, 7(1), 1–21.
Bareinboim, Elias and Judea Pearl, 2013, "A general algorithm for deciding transportability of experimental results," *Journal of Causal Inference*, 1(1), 107–34.
Bareinboim, Elias and Judea Pearl, 2014, "Transportability from multiple environments with limited experiments: completeness results," in M. Welling, Z. Ghahramani, C. Cortes, and N. Lawrence, eds., *Advances of Neural Information Processing*, 27, (NIPS Proceedings), 280–8.
Bauchet, Jonathan, Jonathan Morduch, and Shamika Ravi, 2015, "Failure vs displacement: why an innovative anti-poverty program showed no net impact in South India," *Journal of Development Economics*, 116, 1–16.
Basu, Kaushik, 2010, "The economics of foodgrain management in India," Ministry of Finance, Delhi. http://finmin.nic.in/workingpaper/Foodgrain.pdf
Begg, Colin B., 1990, "Significance tests of covariance imbalance in clinical trials," *Controlled Clinical Trials*, 11(4), 223–5.
Bhattacharya, Debopam and Pascaline Dupas, 2012, "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 167 (1), 168–96.
Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes, 2006, "What mean impacts miss: distributional effects of welfare reform experiments," *American Economic Review*, 96(4), 988–1012.
Bleyer, Archie, and H. Gilbert Welch, 2012, "Effect of three decades of screening mammography on breast-cancer incidence," *New England Journal of Medicine*, 367, 1998–2005
Bloom, Howard S., Carolyn J. Hill, and James A. Riccio, 2005, "Modeling cross-site experimental differences to find out why program effectiveness varies," in Howard

S. Bloom, ed., *Learning more from social experiments: evolving analytical approaches*, New York, NY. Russell Sage.
Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, 2013, "Scaling up what works: experimental evidence on external validity in Kenyan education," Washington, DC. Center for Global Development, Working Paper 321.
Bothwell, Laura E. and Scott H. Podolsky, 2016, "The emergence of the randomized, controlled trial," *New England Journal of Medicine*, 375(6), 501–4. doi: 10.1056/NEJMp1604635
Campbell, D. T. and J. C. Stanley, 1963, *Experimental and quasi-experimental designs for research*. Chicago. Rand McNally.
Cartwright, Nancy, 1994, *Nature's capacities and their measurement*. Oxford. Clarendon Press.
Cartwright, Nancy, 2007, "Are RCTs the gold standard?" *Biosocieties*, 2, 11–20.
Cartwright, Nancy, 2011, "A philosopher's view of the long road from RCTs to effectiveness," *The Lancet*, 377, 1400–01.
Cartwright, Nancy, 2012, "Presidential address: will this policy work for you? Predicting effectiveness better: how philosophy helps," *Philosophy of Science*, 79, 973–89.
Cartwright, Nancy, 2016. "Where is the rigor when you need it?" in I. Marinovic, ed., *Foundations and Trends in Accounting: special issue on causal inference in capital markets research*, 10 (2–4): 106–24.
Cartwright, Nancy and Jeremy Hardie, 2012, *Evidence based policy: a practical guide to doing it better*, Oxford. Oxford University Press.
Cartwright, Nancy and Eileen Munro, 2010, "The limitations of RCTs in predicting effectiveness," *Journal of Experimental Child Psychology*, 16(2),
Chalmers, Iain, 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.
Chan, Tat Y. and Barton H. Hamilton, 2006, "Learning, private information, and the economic evaluation of randomized experiments," *Journal of Political Economy*, 114(6), 997–1040.
Chassang, Sylvain, Gerard Padró I Miguel, and Erik Snowberg, 2012, "Selective trials: a principal–agent approach to randomized controlled experiments," *American Economic Review*, 102(4), 1279–1309.
Chassang, Sylvain, Erik Snowberg, Ben Seymour, and Cayley Bowles, 2015, "Accounting for behavior in treatment effects: new applications for blind trials," *PLoS One*, 10(6), e0127227. doi: 10:1371/journal.pone.0127227.
Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers, 2005, "Missing in action: teacher and health worker absence in developing countries," *Journal of Economic Perspectives*, 19(4), 91–116.
Chyn, Eric, 2016, "Moved to opportunity: the long-run effect of public housing demolition on labor market outcomes of children," University of Michigan. http://www-personal.umich.edu/~ericchyn/Chyn_Moved_to_Opportunity.pdf

Chetty, Raj, 2009, "Sufficient statistics for welfare analysis: a bridge between structural and reduced-form methods," *Annual Review of Economics*, 1, 451–87.
Conlisk, John, 1973, "Choice of response functional form in designing subsidy experiments," *Econometrica*, 41(4), 643–56.
Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, 2014, "Do labor market policies have displacement effects? evidence from a clustered randomized experiment," *Quarterly Journal of Economics*, 128(2), 531–80.
Das, Jishnu and Jeffrey Hammer, 2005," Which doctor? Combining vignettes and item response to measure clinical competence," *Journal of Development Economics*, 78, 348–83.
Davey-Smith, George, and Shah Ibrahim, 2002, "Data dredging, bias, or confounding," *British Medical Journal*, 325, 1437–8.
Deaton, Angus, 2010, "Instruments, randomization, and learning about development," *Journal of Economic Literature*, 48(2), 424–55.
Deaton, Angus and Nancy Cartwright, 2016, "Understanding and misunderstanding randomized controlled trials," http://www.princeton.edu/~deaton/download.html?pdf=Deaton_Cartwright_RCTs_with_ABSTRACT_August_25.pdf
Deaton, Angus and John Muellbauer, 1980, *Economics and consumer behavior*, New York. Cambridge University Press.
Deaton, Angus and Serena Ng, 1998, "Parametric and nonparametric approaches to price and tax reform," *Journal of the American Statistical Association*, 93 (443), 900–9.
Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch, 2012, "Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education," J–PAL, MIT, December 3rd. http://www.povertyactionlab.org/publication/cost-effectiveness
Drèze, Jean, 2016, Personal email communication.
Duflo, Esther, 2017, "The economist as plumber," *American Economic Review*, 107(5), 1–26.
Duflo, Esther, Rema Hanna, and Stephen P. Ryan, 2012, "Incentives work: getting teachers to come to school," *American Economic Review*, 102(4), 1241–78.
Duflo, Esther and Michael Kremer, 2008, "Use of randomization in the evaluation of development effectiveness," in William Easterly, ed., *Reinventing foreign aid*. Washington, DC. Brookings, 93–120.
Dynarski, Susan, 2015, "Helping the poor in education: the power of a simple nudge," *New York Times*, Jan 17, 2015.
Fine, Paul E. M. and Jacqueline A. Clarkson, 1986, "Individual versus public priorities in the determination of optimal vaccination policies," *American Journal of Epidemiology*, 124(6), 1012–20.
Fisher, Ronald A., 1926, "The arrangement of field experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–13.
Filmer, Deon, Jeffrey Hammer, and Lant Pritchett, 2000, "Weak links in the chain: a diagnosis of health policy in poor countries," *World Bank Research Observer*, 15(2), 199–204.

Freedman, David A., 2008, "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 40, 180–93.
Frieden, Thomas R., 2017, "Evidence for health decision making—beyond randomized, controlled trials," *New England Journal of Medicine*, 377, 465–75.
Garfinkel, Irwin and Charles F. Manski, 1992, "Introduction," in Irwin Garfinkel and Charles F. Manski, eds., *Evaluating welfare and training programs*, Cambridge, MA. Harvard University Press. 1–22.
Gerber, Alan S. and Donald P. Green, 2012, *Field Experiments*, New York. Norton.
Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch, 2016, *Impact evaluation in practice*, 2nd Edition, Washington, DC. Inter-American Development Bank and World Bank.
Goldberger, Arthur S. and Charles F. Manski, 1995, "Review Article: The Bell Curve by Herrnstein and Murray," *Journal of Economic Literature*, 33(2), 762–76.
Greenberg, David and Mark Shroder, 2004, *The digest of social experiments* (3rd ed.), Washington, DC. Urban Institute Press.
Greenberg, David, Mark Shroder, and Matthew Onstott, 1999, "The social experiment market," *Journal of Economic Perspectives*, 13(3), 157–72.
Gueron, Judith M. and Howard Rolston, 2013, *Fighting for reliable evidence*, New York, Russell Sage.
Guyatt, Gordon, David L. Sackett, and Deborah J. Cook for the Evidence-Based Medicine Working Group, 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?" *Journal of the American Medical Association*, 271(1), 59–63.
Harrison, Glenn W., 2013, "Field experiments and methodological intolerance," *Journal of Economic Methodology*, 20(2), 103–17.
Harrison, Glenn W., 2014, "Impact evaluation and welfare evaluation," *European Journal of Development Research*, 26, 39–45.
Harrison, Glenn W., 2014, "Cautionary notes on the use of field experiments to address policy issues," *Oxford Review of Economic Policy*, 30 (4), 753–63.
Hausman, Jerry A. and David A. Wise, 1985, "Technical problems in social experimentation: cost versus ease of analysis," in Jerry A. Hausman and David A. Wise, eds., *Social Experimentation*, Chicago, IL. Chicago University Press. 187–220.
Heckman, James J., 1992, "Randomization and social policy evaluation," in Charles F. Manski and Irwin Garfinkel, eds., *Evaluating welfare and training programs*, Cambridge, MA. Harvard University Press. 547–70.
Heckman, James J., 1997, "Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations," *Journal of Human Resources*, 32(3), 441–62.
Heckman, James J., 2005, "The scientific model of causality," *Sociological Methodology*, 35 (1), 1–97.
Heckman, James J., 2008, "Econometric causality," *International Statistical Review*, 76 (1), 1–27.

Heckman, James J., 2010, "Building bridges between structural and program evaluation approaches to evaluating policy," *Journal of Economic Literature*, 48(2), 356–98.
Heckman, James J., Neil Hohman, and Jeffrey Smith, with the assistance of Michael Khoo, 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, James J., Robert J. Lalonde, and Jeffrey A. Smith, 1999, "The economics and econometrics of active labor markets," Chapter 31 in Ashenfelter, Orley and David Card, eds. *Handbook of labor economics*, Amsterdam. North-Holland, 3(A), 1866–2097.
Heckman, James J., Rodrigo Pinto, and Peter Savelyev, 2013, "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes," *American Economic Review*, 103(6), 2052–86.
Heckman, James J. and Jeffrey Smith, 1995, "Assessing the case for social experiments," *Journal of Economic Perspectives*, 9(2), 85–110.
Heckman, James J., Jeffrey Smith, and Nancy Clements, 1997, "Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts," *Review of Economic Studies*, 64(4), 487–535.
Heckman, James J. and Sergio Urzúa, 2010, "Comparing IV with structural models: what simple IV can and cannot identify," *Journal of Econometrics*, 156, 27–37.
Heckman, James J. and Edward Vytlacil, 2005, "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.
Heckman, James J. and Edward J. Vytlacil, 2007, "Econometric evaluation of social programs, Part 1: causal models, structural models, and econometric policy evaluation," Chapter 70 in James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, 6B, 4779–874.
Horton, Richard, 2000, "Common sense and figures: the rhetoric of validity in medicine: Bradford Hill memorial lecture 1999," *Statistics in medicine*, 19, 3149–64.
Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer, 2005, "Predicting the efficacy of future training programs using past experience at other locations," *Journal of Econometrics*, 125, 241–70.
Hsieh, Chang-tai and Miguel Urquiola, 2006, "The effects of generalized school choice on achievement and stratification: evidence from Chile's voucher program," *Journal of Public Economics*, 90, 1477–1503.
Hurwicz, Leonid, 1966, "On the structural form of interdependent systems," *Studies in Logic and the Foundations of Mathematics*, 44, 232–9.
Imbens, Guido W., 2004, "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics*, 86(1), 4–29.
Imbens, Guido W., 2010, "Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua," *Journal of Economic Literature*, 48 (2), 399–423.
Imbens, Guido W. and Joshua D. Angrist, 1994, "Identification and estimation of local average treatment effects," *Econometrica*, 62(2), 467–75.
Imbens, Guido W. and Michal Kolesár, 2016, "Robust standard errors in small samples: some practical advice," *Review of Economics and Statistics*, 98(4), 701–12.

Imbens, Guido W. and Jeffrey M. Wooldridge, 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*, http://www.icmje.org/icmje-recommendations.pdf (accessed, August 20, 2016.)
J_PAL, 2017, https://www.povertyactionlab.org/about-j-pal, (accessed, August 21, 2017).
Kahneman, Daniel and Gary Klein, 2009, "Conditions for intuitive expertise: a failure to disagree," *American Psychologist*, 64(6), 515–26.
Karlan, Dean and Jacob Appel, 2011, *More than good intentions: how a new economics is helping to solve global poverty*, New York. Dutton.
Karlan, Dean, Nathaneal Goldberg and James Copestake, 2009, "Randomized controlled trials are the best way to measure impact of microfinance programs and improve microfinance product designs," *Enterprise Development and Microfinance*, 20(3), 167–76.
Kasy, Maximilian, 2016, "Why experimenters might not want to randomize, and what they could do instead," *Political Analysis*, 1–15 doi: 10.1093/pan/mpw012
Kramer, Peter, 2016, *Ordinarily well: the case for antidepressants*, New York. Farrar, Straus, and Giroux.
Kremer, Michael and Alaka Holla, 2009, "Improving education in the developing world: what have we learned from randomized evaluations?" *Annual Review of Economics*, 1, 513–42.
Lalonde, Robert J., 1986, "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review*, 76 (4), 604–20.
Lehman, Erich. L. and Joseph P. Romano, 2005, *Testing statistical hypotheses* (third edition), New York. Springer.
Levy, Santiago, 2006, *Progress against poverty: sustaining Mexico's Progresa-Oportunidades program*, Washington, DC. Brookings.
Mackie, John L., 1974, *The cement of the universe: a study of causation*, Oxford. Oxford University Press.
Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett Keeler, and Arleen Leibowitz, 1988a, "Health insurance and the demand for medical care: evidence from a randomized experiment," *American Economic Review*, 77(3), 251–77.
Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett Keeler, Bernadette Benjamin, Arleen Leibowitz, M. Susan Marquis, and Jack Zwanziger, 1988b, *Health insurance and the demand for medical care: evidence from a randomized experiment*, Santa Monica, CA. RAND.
Manski, Charles F., 2004, "Treatment rules for heterogeneous populations," *Econometrica*, 72(4), 1221–46.
Manski, Charles F., 2013, *Public policy in an uncertain world: analysis and decisions*, Cambridge, MA. Harvard University Press.
Manski, Charles F. and Aleksey Tetenov, 2016, "Sufficient trial size to inform clinical practice," *PNAS*, 113(38), 10518–23.

Metcalf, Charles E., 1973, "Making inferences from controlled income maintenance experiments," *American Economic Review*, 63(3), 478–83.
Moffitt, Robert, 1979, "The labor supply response in the Gary experiment," *Journal of Human Resources*, 14(4), 477–87.
Moffitt, Robert, 1992, "Evaluation methods for program entry effects," Chapter 6 in Charles Manski and Irwin Garfinkel, *Evaluating welfare and training programs*, Cambridge, MA. Harvard University Press, 231–52.
Moffitt, Robert, 2004, "The role of randomized field trials in social science research: a perspective from evaluations of reforms of social welfare programs," *American Behavioral Scientist*, 47(5), 506–40
Morgan, Kari Lock and Donald B. Rubin, 2012, "Rerandomization to improve covariate balance in experiments," *Annals of Statistics*, 40(2), 1263–82.
Muller, Seán M., 2015, "Causal interaction and external validity: obstacles to the policy relevance of randomized evaluations," *World Bank Economic Review*, 29, S217–S225.
Orcutt, Guy H. and Alice G. Orcutt, 1968, "Incentive and disincentive experimentation for income maintenance policy purposes," *American Economic Review*, 58(4), 754–72.
Pearl, Judea and Elias Bareinboim, 2011, "Transportability of causal and statistical relations: a formal approach," *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, AAAI Press, 247–54,
Pearl, Judea and Elias Bareinboim, 2014, "External validity: from do-calculus to transportability across populations," *Statistical Science*, 29(4), 579-95.
Rodrik, Dani, 2006, personal email communication.
Rothwell, Peter M., 2005, "External validity of randomized controlled trials: 'to whom do the results of the trial apply'", *Lancet*, 365, 82–93.
Russell, Bertrand, 2008 [1912], *The problems of philosophy*, Rockville, MD. Arc Manor.
Sackett, David L., William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes and W. Scott Richardson, 1996, "Evidence based medicine: what it is and what it isn't," *British Medical Journal*, 312 (January 13), 71–2.
Savage, Leonard J., 1962, "Subjective probability and statistical practice," in G.A. Barnard and D. R. Cox, eds., *The Foundations of Statistical Inference*, London. Methuen. 9–35.
Scriven, Michael, 1974, "Evaluation perspectives and procedures," in W. James Popham, ed., *Evaluation in education—current applications*, Berkeley, CA. McCutchan Publishing Corporation.
Senn, Stephen, 1994, "Testing for baseline balance in clinical trials," *Statistics in Medicine*, 13, 1715–26.
Senn, Stephen, 2013, "Seven myths of randomization in clinical trials," *Statistics in Medicine* 32, 1439–50.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell, 2002, *Experimental and quasi-experimental designs for generalized causal inference*, Boston, MA. Houghton Mifflin.
Simpson, Adrian, 2017, **"**The misdirection of public policy: comparing and combining standardised effect sizes," *Journal of Educational Policy* 32(4), 450-66.

Stuart, Elizabeth A., Stephen R. Cole, and Catharine P. Bradshaw, and Philip J. Leaf, 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 (W. S. Gosset), 1938, "Comparison between balanced and random arrangements of field plots," *Biometrika*, 29 (3/4), 363–78.
Svorencik, Andrej, 2015, *The experimental turn in economics: a history of experimental economics*, Utrecht School of Economics, Dissertation Series #29, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2560026
Todd, Petra E. and Kenneth J. Wolpin, 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, Petra E. and Kenneth J. Wolpin, 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.
Vandenbroucke, Jan P., 2004, "When are observational studies as credible as randomized controlled trials?" *The Lancet*, 363:1728–31.
Vandenbroucke, Jan P.. 2009, "The HRT controversy: observational studies and RCTs fall in line," *The Lancet*, 373, 1233–5.
Vivalt, Eva, 2015, "How much can we generalize from impact evaluations?" NYU, unpublished. http://evavivalt.com/wp-content/uploads/2014/10/Vivalt-JMP-10.27.14.pdf
White, Halbert, 1980, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 50(1), 1–25.
Wise, David A., 1985, "A behavioral model versus experimentation: the effects of housing subsidies on rent," in P. Brucker and R. Pauly, eds. *Methods of Operations Research*, 50, Verlag Anon Hain. 441–89.
Wolpin, Kenneth I., 2013, *The limits of inference without theory*, Camridge, MA. MIT Press.
Worrall, John, 2007, "Evidence in medicine and evidence-based medicine," *Philosophy Compass*, 2/6, 981–1022.
Worrall, John, 2008, "Evidence and ethics in medicine," *Perspectives in Biology and Medicine*, 51(3), 418–31.
Yates, Frank, 1939, "The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments," *Biometrika*, 30 (3/4), 440–66.
Young, Alwyn, 2016, "Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results," London School of Economics, Working Paper, Feb.
Ziliak, Stephen T., 2014, "Balanced versus randomized field experiments in economics: why W. S. Gosset aka 'Student' matters," *Review of Behavioral Economics*, 1, 167–208.

Appendix: Monte Carlo experiment for an RCT with outliers

In this illustrative example, there is 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 β are distributed so that $\beta + e^{0.5} \sim \Lambda(0,1)$, for standardized lognormal distribution Λ. 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.

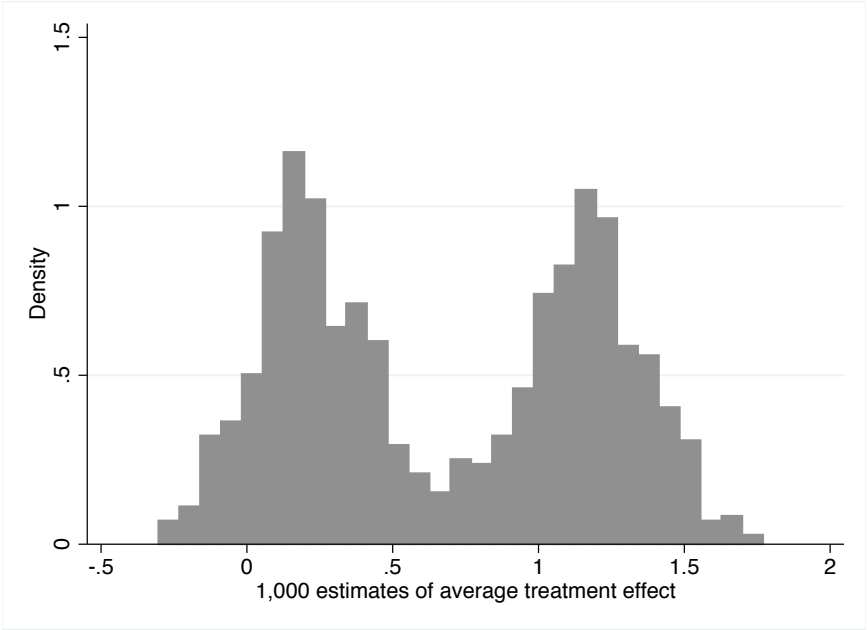

Figure A1: Estimates of an ATE with an outlier in the trial sample

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 right-hand side of the figure, when it is in the control group, we get the left-hand side.