Understanding and misunderstanding randomized controlled trials

Angus Deaton and Nancy Cartwright

Princeton University

Durham University and UC San Diego

This version, August 2016

We acknowledge helpful discussions with many people over the many 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, Bo Honoré, and Julian Reiss. Ulrich Mueller had a major influence on shaping Section 1 of the paper. We have benefited from generous comments on an earlier version by Tim Besley, Chris Blattman, Sylvain Chassang, Steven Durlauf, Jean Drèze, William Easterly, Jonathan Fuller, Lars Hansen, Jim Heckman, Jeff Hammer, Macartan Humphreys, Helen Milner, Suresh Naidu, Lant Pritchett, Dani Rodrik, Burt Singer, 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). Deaton acknowledges financial support through the National Bureau of Economic Research, Grants 5R01AG040629-02 and P01 AG05842-14 and through Princeton University's Roybal Center, Grant P30 AG024928.

ABSTRACT

RCTs are valuable tools whose use is spreading in economics and in other social sciences. They are seen as desirable aids in scientific discovery and for generating evidence for policy. Yet some of the enthusiasm for RCTs appears to be based on misunderstandings: that randomization provides a fair test by equalizing everything but the treatment and so allows a precise estimate of the treatment alone; that randomization is required to solve selection problems; that lack of blinding does little to compromise inference; and that statistical inference in RCTs is straightforward, because it requires only the comparison of two means. None of these statements is true. RCTs do indeed require minimal assumptions and can operate with little prior knowledge, an advantage when persuading distrustful audiences, but a crucial disadvantage for cumulative scientific progress, where randomization adds noise and undermines precision. The lack of connection between RCTs and other scientific knowledge makes it hard to use them outside of the exact context in which they are conducted. Yet, once they are seen as part of a cumulative program, they can play a role in building general knowledge and useful predictions, provided they are combined with other methods, including conceptual and theoretical development, to discover not "what works," but why things work. Unless we are prepared to make assumptions, and to stand on what we know, making statements that will be incredible to some, all the credibility of RCTs is for naught.

Introduction

Randomized trials are currently much used in economics and are widely considered to be a desirable method of empirical analysis and discovery. There is a long history of such trials in the subject. There were four large federally sponsored negative income tax trials in the 1960s and 1970s. In the mid-1970s, there was a famous, and still frequently cited, trial on health insurance, the Rand health experiment. There was then a period during which randomized controlled trials (RCTs) received less attention by academic economics; even so, randomized trials on welfare, social policy, labor markets, and education have continued since the mid-1970s, some with substantial involvement and discussion by academic economists, see Greenberg and Shroder (2004).

Recent randomized trials in economic development have attracted attention, and the idea that such trials can discover "what works" has been widely adopted 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 parameter estimates that are more credible than other empirical methods that do not involve the comparison of randomly selected treatment and control groups. RCTs are seen as largely exempt from many of the econometric problems that characterize observational studies. When RCTs are not feasible, researchers often mimic randomized designs by using observational data to construct two groups that, as far as possible, are identical and differ only in their exposure to treatment.

The preference for randomized trials has spread beyond trialists to the general public and the media, which typically reports favorably on them. They are seen as accurate, objective, and 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 huge range of areas of social concern across Europe and the Anglophone world, such as the US Department of Education's What Works Clearing House, The Campbell Collaboration (parallel to the Cochrane Collaboration in health), the Scottish Intercollegiate Guidelines Network (SIGN), the US Department of Health and Human Services Child Welfare Information Gateway, the US Social and Behavioral Sciences Team, and others. The British government has established eight new (well-financed) What Works Centers similar to the National Institute for Health and Care Excellence (NICE), with more planned. They extend NICE's evaluation of health treatment into aging, early intervention, education, crime, local economic growth, Scottish service delivery, poverty, and wellbeing. These centers see randomized controlled trials as their

preferred tool. There is a widespread desire for careful evaluation—to support what is sometimes called the "audit society"—and everyone assents to the idea that policy should be based on evidence of effectiveness, for which randomized trials appear to be ideally suited. Trials are easily, if not very precisely, explained along the lines that random selection generates two otherwise identical groups, one treated and one not; results are easy to compute—all we need is the comparison of two averages; and unlike other methods, it seems to require no specialized understanding of the subject matter. It seems a truly general tool that (nominally) works in the same way in agriculture, medicine, sociology, economics, politics, and education. It is supposed to require no prior knowledge, whether suspect or not, which is seen as a great advantage.

In this paper, we present two sets of arguments, one on conducting RCTS and on how to interpret the results, and one on how to use the results once we have them. Although we do not care for the terms—for reasons that will become apparent—the two sections correspond roughly to internal and external validity.

Randomized controlled trials are often useful, and have been important sources of empirical evidence for causal claims and evaluation of effectiveness in many fields. Yet many of the popular interpretations—not only among the general public, but also among trialists—are incomplete and sometimes misleading, and these misunderstandings can lead to unwarranted trust in the impregnability of results from RCTs, to a lack of understanding of their limitations, and to mistaken claims about how widely their results can be used. All these, in turn, can lead to flawed policy recommendations.

Among the misunderstandings are the following: (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 average treatment effects, but also precise estimates; (c) randomization is necessary to solve the selection problem; (d) lack of blinding, which is common in social science experiments, does not seriously compromise inference; (e) statistical inference in RCTs, which requires only the simple comparison of means, is straightforward, so that standard significance tests are reliable.

While many of the problems of RCTs are shared with observational studies, some are unique, for example the fact that randomizing itself can change outcomes independently of treatment. More generally, it is almost never the case that an RCT can be judged superior to a well-conducted observational study simply by virtue of being an RCT. The idea that all methods

have their flaws, but RCTs always have fewest, is one of the deepest and mort pernicious misunderstandings.

In the second part of the paper, we discuss the uses and limitations of results from RCTs for making policy. The non-parametric and theory-free nature of RCTs, which is arguably an advantage in estimation, is a serious disadvantage when we try to use the results outside of the context in which they were obtained. Much of the literature, in economic development and elsewhere, perhaps inspired by Campbell and Stanley's (1963) famous "primacy of internal validity," assumes that internal validity is enough to guarantee the usefulness of the estimates in different contexts. Without understanding RCTs within the context of the knowledge that we already possess about the world, much of it obtained by other methods, we do not know how to use trial results. But once the commitment has been made to seeing RCTs within this broader structure of knowledge and inference, and when they are designed to fit within it, they can play a useful role in building general knowledge and policy predictions; for example, an RCT can be a good way of estimating a key policy magnitude. The broader context within which RCTs need to be set includes not only models of economic structure, but also the previous experience that policymakers have accumulated about local settings and implementation. Most importantly for economic development, the use of RCT results should be sensitive to what people want, both individually and collectively. RCTs should not become yet another technical fix that is imposed on people by bureaucrats or foreigners; RCT results need to be incorporated into a democratic process of public reasoning, Sen (2011). Greenberg, Shroder, and Onstott (1999) document that, even before the recent wave of RCTs in development, most RCTs in economics have been carried out by rich people on poor people, and the fact should make us especially sensitive to avoid charges of paternalism.

Section 1: Interpreting the results of RCTs

1.1 *Prolog*

RCTs were first popularized by Fisher's agricultural trials in the 1930s and are today often described by the Rubin counterfactual causal model, which itself traces back to Neyman in 1923, see Freedman (2006) for a description of the history: Each unit *i* (a person, a pupil, a school, an agricultural plot) is assumed to have two possible outcomes, Y_{io} and Y_{i1}, the former occurring if there is no treatment at the time in question, the latter if the unit is treated. The difference between the two outcomes $Y_{i1} - Y_{i0}$ is the individual treatment effect, which we shall denote β_i. Treatment effects are typically different for different units. No unit can be both treated and

untreated at the same time, so only one or other of the outcomes occurs; the other is counterfactual so that individual treatment effects are in principle unobservable.

We note parenthetically that while we use the counterfactual framework here, we do not endorse it, nor argue against other approaches that do not use it, such as the Cowles commission econometric framework where the causal relations are coded as structural equations, see also Pearl (2009.) Imbens and Wooldridge (2009, Introduction) provide an eloquent defense of the Rubin formulation, emphasizing the credibility that comes from a theory-free specification with unlimited heterogeneity in treatment effects. Heckman and Vytlacil (2007, Introduction) make an equally eloquent case against, noting that the treatments in RCTs are often unclearly specified and that the treatment effects are hard to link to invariant parameters that would be useful elsewhere.

The basic theorem governing RCTs is a remarkable one. It states that the average treatment effect is the average outcome in the treatment group minus the average outcome in the control group. While we cannot observe the *individual* treatment effects, we can observe their mean. The estimate of the average treatment effect (ATE) is simply the difference between the means in the two groups, and it has a standard error that can be estimated and used to make significance statements according to the statistical theory that applies to the difference of two means, on which more below in Section 1.3. The difference in means is an *unbiased* estimator of the mean treatment effect.

The theorem is remarkable because it requires so few assumptions; no model is required, no assumptions about covariates are needed, the treatment effects can be heterogeneous, and nothing is assumed about the shapes of statistical distributions other than the statistical question of the existence of the mean of the counterfactual outcome values. In terms of one of our running themes, it requires no expert knowledge, or no acceptance of priors, expert or otherwise. The theorem also has its limitations; the proof uses the fact that the difference in two means is the mean of the individual differences, i.e. the treatment effects. This is not true for the median (the difference in two medians is not the median of the differences which is the median treatment effect). It also does not allow us to estimate any percentile of the distribution of treatment effects, or its variance. (Quantile estimates of treatment effects are not the quantiles of the distribution of treatment effects, but the differences in the quantiles of the two marginal distributions of treatments and controls; the two measures coincide if the experiment has no effect on ranks, an assumption that would be convenient but is hard to justify, at least in

general.) All of these statistics can be of interest for policy but RCTs are not informative about them, or at least not without further assumptions, for example on the distribution of treatment effects, see Heckman, Smith, and Clements (1997), and much of the attraction of RCTs is the absence of such assumptions.

The basic theorem tells us that the difference in means is an unbiased estimator of the average treatment effect but says nothing about the variance of this estimator. In general, a biased estimator that is typically closer to the truth will often be better than an unbiased estimator that is typically wide of the truth. There is nothing to say that a non-RCT estimator, in spite of bias, might not have a lower mean squared error (MSE), one measure of the distance of the estimate from the truth, or a lower value of a "loss function" that defines the loss to the experimenter of missing the target.

It is useful to think of the mean average treatment effect from an RCT in terms of sampling from a finite population, as when the Bureau of the Census estimates average income of the US population in 2013. For the RCT, the population is the population of units whose average treatment effect is of interest; note the importance of defining the population of interest because, given the heterogeneity of treatment effects, the average treatment effect will vary across different populations, just as average incomes differ across different subpopulations of the US. Finite population sampling theory tells us how to get accurate estimates of means from samples; in the RCT case, the sample is the study sample, both treatments and controls. In principle, the study sample could be a random sample of the parent population of interest, in which case it is representative of it, but that is seldom the case. Because the estimate is population specific, it is not (or need not be) thought of as the parameter of a super-population, or otherwise generalizable in any way. Average income in the US in 2013 may be of interest in its own right; but it will not be the same as average income in 2014, nor will it be the same as average income of whites, or of the populations of Wyoming or New York. Exactly the same is true of the estimate of an average treatment effect; it applies to the study sample in which the trial was done, at the time when it was done, and its use outside of those confines, though often possible, requires argument and justification. Without such an argument, we cannot claim that an ATE is "the" mean treatment effect any more than that average income in the US in 2013 is "the" average income of the US in any other year. Of course, knowing average income in 2013 can be useful for making other calculations, such as an estimate of income in 2014, or of a sub-

population that we know is richer or poorer; the fact that an estimate does not universally generalize does not make it useless. We shall return to these issues in Section 2.

1.2. *Precision, balance, and randomization*

1.2.1 *Precision and bias*

We should like our estimate of the average treatment effect to be as close to the truth as possible. One way to assess closeness is the mean square error (MSE), defined as

$$MSE = E(\widehat{\theta} - \theta)^2 \tag{1}$$

where θ is the true average treatment effect, and $\widehat{\theta}$ is its estimate from a particular trial. The expectation is taken over repeated randomizations of treatments and controls using the same study population. It is also standard to rewrite (1) as

$$MSE = E\left((\widehat{\theta} - E(\widehat{\theta})\right)^2 + \left(E(\widehat{\theta}) - \theta\right)^2 = \mathrm{var}(\widehat{\theta}) + bias(\widehat{\theta},\theta)^2 \tag{2}$$

so that mean square error is the sum of the variance of the estimator—which we typically know something about from the estimated standard error—and the square of the bias—which in the case of a(n ideal) randomized controlled trial is zero. The elementary, but crucial point is that, while it is certainly good that the bias is zero, that fact does nothing to make the distance from the truth as small as it might be, which is what we really care about. An unbiased estimator that is nearly always wide of the target is not as useful as one that is always near to it, even if, on average, it is off center. More generally, it will often be desirable to trade in some unbiasedness for greater precision. Experiments are often expensive, so we cannot always rely on large samples to bring the estimate close to the truth and resolve these issues for us. Much of this Section is concerned with how to design experiments to maximize precision.

Unbiasedness alone cannot therefore justify the often-expressed preference for RCTs over other estimators. The minimalist assumptions required for an RCT to be unbiased are an attraction although, as we shall see in this Section, this advantage usually comes at the cost of lowered precision and of difficulties in knowing how to use the result, as we shall see in Section 2. Yet there is an often expressed belief that RCTs are somehow guaranteed to be precise, simply because they are RCTs. Occasionally bias and precision are explicitly confused; the JPAL website, in its explanation of why it is good to randomize, says that RCTs "are generally considered the most rigorous and, all else equal, produce the most accurate (i.e. unbiased) results." Shadish, Cook, and Campbell (2002, p. 276), in what is (rightly) considered one of the bibles of causal inference in social science, state without qualification that "randomized experiments provide a

precise answer about whether a treatment worked" (p. 276) and "The randomized experiment is often the preferred method for obtaining a *precise* and statistically unbiased estimate of the effects of an intervention," (p. 277) our italics.

Contrast this with Cronbach et al (1980) who quotes Kendall's (1957) pastiche of Longfellow, "Hiawatha designs an experiment," where Hiawatha's insistence on unbiasedness leads to his never hitting the target and to his eventual banishment.

1.2.2 *Balance and precision in a linear all-cause model*

A useful way to think about precision and what an RCT does and does not do is to use a schematic linear causal model of the form:

$$Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij} \tag{3}$$

where, as before, 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 causes of the outcome, and we suppose that (3) captures all the causes of Y_i. *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, and we shall have occasion to make this explicit below. An obvious and important example is when the treatment if effective only in the presence of a particular value of one of the *x*'s.

We do not need *i* subscripts on the $\gamma's$ 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. Because the $\beta's$ can depend on the *x*'s, the effects of the *x*'s on the outcome can depend on T_i, or, equivalently, the effects of treatment can depend on covariates.

In an experiment, *with or without randomization*, we can represent the treatment group as having $T_i = 1,$ and the control group as having $T_i = 0.$ So when we subtract the average outcomes among the controls from the average outcomes among the treatments, we will get

$$\overline{Y}^1 - \overline{Y}^0 = \overline{\beta}^1 + \sum_{j=1}^{J} \gamma_j (\overline{x}_{ij}^1 - \overline{x}_{ij}^0) = \overline{\beta}^1 + (\overline{S}^1 - \overline{S}^0) \tag{4}$$

The first term on the far right hand side, which is the average treatment effect, is what we want, but the second term or error term, which is the sum of the net average balances of other causes

across the two groups, will generally be non-zero—because of selection or many other reasons—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 when the sum of their net differences $\overline{S}^1 - \overline{S}^0$ 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 effect among the treated, so that we have the ultimate precision and we know the answer exactly, at least in this linear case.

1.2.3 *Balancing acts: real and magical*

How do we get balance, or something close to it? What, exactly, is the role of randomization? In a laboratory experiment, where there is good background knowledge of the other causes, the experimenter has a good chance of controlling all of the other causes, aiming to ensure that the last term in (4) is close to zero. Failing such knowledge and control, an alternative is *matching*, frequently used in statistical, medical, and econometric work. For each treatment, a match is found that is as close as possible on all suspected causes, so that, once again, the last term in (4) can be kept small. Again, when we have a good idea of the causes, matching may also deliver a precise estimate. Of course, when there are important unknown or unobservable causes, neither laboratory control nor matching offers protection.

What does randomization do? Because the treatments and controls come from the same underlying distribution, randomization guarantees, by construction, that the last term on the right in (4) is zero *in expectation* at baseline (much can happen to disturb this beyond baseline). This is true whether or not the causes are observed. If the RCT is repeated many times on the same trial population, then the last term will be zero when averaged over an infinite number of (entirely hypothetical) trials. Of course, this does nothing to make it zero in any one trial where the difference in means will be equal to the average treatment effect among those treated *plus* a 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 the randomization that limits its size; by chance, there can be one (or more) important excluded cause(s) that is very unequally distributed between treatment and controls. This imbalance will vary over replications of the trial, and its average size will ideally be captured by the standard error of the estimated ATE, which gives us some idea of how likely we are to be away from the truth. Getting the standard error and associated significance statements right are therefore of great importance.

Exactly what randomization does is frequently lost in the practical literature, and there is often a confusion between perfect control, on the one hand—as in a laboratory experiment or perfect matching with no unobservable causes—and control in expectation—which is what RCTs do. We suspect that at least some of the popular and professional enthusiasm for RCTs, as well as the belief that they are precise by construction, comes from misunderstandings about balance. These misunderstandings are not so much among the trialists who, when pressed, will give a correct account, but come from imprecise statements by trialists that are taken as gospel by the lay audience that the trialists are keen to reach.

Such a misunderstanding is well captured by the following quote from the World Bank's online manual on impact evaluation:

> "We can be very confident that our estimated average impact, given as the difference between the outcome under treatment (the mean outcome of the randomly assigned treatment group) and our estimate of the counterfactual (the mean outcome of the randomly assigned comparison group) constitute the true impact of the program, since *by construction we have eliminated all observed and unobserved factors that might otherwise plausibly explain the difference in outcomes.*" Gertler et al (2011) (our italics.)

This statement confuses *actual* balance in any single trial with balance in expectation over many entirely hypothetical trials. If the statement above 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, at least in the absence of measurement error. We should not only be confident of our estimate; we would know the truth, as the quote says.

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 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).

Many writers are more cautious, and modify statements about treatment and control groups being identical with terms such as "statistically identical," "reasonably similar" or do not differ "systematically." And we have no doubt that all of the authors quoted above understand the need for these qualifications. But to the uninformed reader, the qualified statements are unlikely to be differentiated from the unqualified statements quoted above. Nor is it always clear what some of these terms mean. For example, if two people are selected at random from a population, and it so happens that one is female and one male, in what sense they are statistically identical? While it is true that they were randomly selected from the same parent distribution, which provides the basis for inference, the calculation of standard errors, and significance statements, it does nothing to help with balance or precision in any given trial.

1.2.4 *Sample size and statistical inference in unbalanced trials*

Is a single trial more likely to be balanced, and thus more precise, when the sample size is large? Indeed, 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 are a large number of causes, balance on *each* cause may be infeasible. Vandenbroucke (2004) notes that there are three million 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.

However, as (4) makes clear, we do not need balance on *all* causes, only on their net effect, the term $\overline{S}^1 - \overline{S}^0$ which does not require balance on each cause individually. Yet there is no guarantee that even the net effect will be small. For example, there may only be one omitted unobserved cause whose effect is large, one single base pair say, so that if that one cause is unbalanced across treatments and controls, that there is individual or even net balance on other less important causes is not going to help.

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.

A simple case illustrates. Suppose that there is one hidden cause in (3), a binary variable *x* that is unity with probability *p* and 0 otherwise. With *n* controls and *n* treatments, the difference in fractions with *x*=1 in the two groups has mean 0 and variance $1/np(1-p)$. With *n*=100 and *p*=0.5, the standard error around 0 is 0.2 so that, if this unobserved confounder has a large effect on the outcome, the imbalance could easily mask the effect of treatment, or be mistaken as evidence for the effectiveness of a truly ineffective treatment.

Lack of balance in the above example or in the net effect of either observables or non-observables in (4) 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. The randomization does not guarantee balance but it provides the basis for making probability statements about the various possible outcomes, which is also clear in the example in the previous paragraph. This was also Fisher's argument for randomization. Senn writes "the probability calculation applied to a clinical trial automatically *makes an allowance for the fact that the groups will almost certainly be unbalanced.*" (italics in the original.) If the design is such that, even with perfect randomization, successive replications tend to generate large imbalances, the resulting imprecision of the ATE will show up in its standard error. Of course, the usefulness of this requires that the calculated standard errors permit correct significance statements, which, as we shall see in the next subsection, is often far from straightforward. In the example above, an extreme, but entirely possible, case occurs when, by chance, the unobserved confounder is

perfectly correlated with the treatment; unless there are *actual* replications, the false certainty that such an experiment provides will be reinforced by false significance tests.

1.2.4 *Testing for balance*

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 (3), which are either causes in their own right or interact with the $\beta's$—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 if we are concerned that the random number generator might have failed (because we are drawing playing cards, rolling dice, or spinning bottle tops, though presumably not if the randomization is done by a random number generator, always supposing that there is such a thing as randomness, Singer and Pincus (1998)), or if we are worried that the randomization is undermined by non-blinded subjects or trialists systematically undermining the allocation. Otherwise, as the next paragraph shows, the test makes no sense and is not informative, which does not seem to stop it being routinely used.

If we write μ^0 and μ^1 for the (vectors of) population means (i.e. the means over all possible randomizations) of the observed *x*'s 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. Indeed, although we cannot "test" it, we know that the null hypothesis is also true for the *unobservable* components of *x*. Note the contrast with the statements 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 expectation over many trials*, which *is* guaranteed by randomization. The confusion is perhaps understandable, but it is confusion nevertheless. Of course, it makes sense to look for balance between observed covariates using some more appropriate distance measure for example the normalized difference in means, Imbens and Wooldridge (2009, equation 3).

1.2.5 *Methods for balancing*

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. But as Scriven (1974, 103) notes: "cause hunting, like lion hunting, is only likely to be successful if we have a considerable amount of relevant background knowledge," or even more strongly, "no causes in, no causes out," Cartwright (1994, Chapter 2). 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 average treatment effects in different strata which can be useful in adapting or transporting the results to other locations, see Section 2.

Stratification is not possible when 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. An alternative is to *re-randomize*, repeating the randomization until the distance between the observed covariates is less than some predetermined criteria. Morgan and Rubin (2012) suggest the Mahalanobis *D*–statistic, 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) has analyzed this method and argues "if adjustment made a substantial difference, we would suggest much caution when interpreting the results." But a substantial difference is exactly what we would like to see, at least some of the time, if the adjustment moves the estimate closer to the truth. Freedman 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. There is also no general guarantee that the regression adjustment will generate a more precise estimate, although it will do so if there are equal numbers of treatments and controls or if the treatment effects are constant over units (in which case there will also be no bias). Even with bias, the regression adjustment is attractive if it does in-

deed trade off bias for precision, though presumably not to RCT purists for whom unbiasedness is the *sine qua non.* Note again that the increased precision, when it exists, comes from using prior knowledge about the variables that are likely to be important for the outcome. That the background knowledge or theory is widely shared and understood will also provide some protection against data mining by searching through covariates in the search for (perhaps falsely) estimated precision.

1.2.6 *Should we randomize?*

The tension between randomization and precision goes back to the early debate between Fisher and Student (Gosset) who never accepted Fisher's arguments for randomization, see also Ziliak (2014). In his debate with Fisher about agricultural trials, Student argued that randomization ignored relevant prior information, for example about how likely confounders would be distributed across the test plots, so that randomization wasted resources and led to unnecessarily poor estimates. This general question of whether randomization is desirable has been reopened in recent papers by Kasy (2016), Banerjee, Chassang, and Snowberg (2016) and Banerjee, Chassang, Montero, and Snowberg (2016).

Refer back to the MSE introduced above, and consider designing an experiment that will make this as small as possible. Unfortunately, this is not generally possible; for example, the "estimator" of 3, say, for the ATE has the lowest possible mean-squared error if the true ATE is actually 3. Instead, we need to average the MSE over a distribution of possible ATEs. This leads to a decision theory approach to estimation whereby a Bayesian econometrician will estimate the ATE by choosing the allocation of treatment and controls so as to minimize the expected value of a loss function—the MSE being one example. Such an approach requires us to specify a prior on the ATE, or more generally, on the expectation of outcomes conditional on the covariates. These priors are formal versions of the issue that has already come up repeatedly, that to get good estimators, we need to know something about how the covariates affect the outcome. Kasy (2016) solves this problem for the case of expected MSE and shows that *randomization is undesirable*; it simply adds noise and makes the MSE larger. He uses a non-parametric prior that has proved useful in a number of other applications—we could presumably do even better if we were prepared to commit further, and he provides code to implement his method, which shows a 20 percent reduction in MSE compared with randomization (14 percent for stratified randomization) for the well-known Tennessee STAR class-size experiment.

Banerjee et al propose a more general loss function and prove the comparable theorem, that randomization leads to larger losses than the optimal non-random purposive assignment. These authors recommend randomization on other grounds, which we will discuss below, but agree that, for standard statistical efficiency or maximization of expected utility randomization should not be used in experimental design. Student was right.

Several points should be noted. First, the anti-randomization theorem is *not* a justification of *any* non-experimental design, for example one that compares outcomes of those who do or do not self-select into treatment. Selection effects are real enough, and if selection is based on unobservable causes, comparison of treated and controls will be biased. One acceptable non-random scheme is to use the observable covariates to divide the study sample into cells within which all observations have the same value and then divide each cell into treatments and controls. Within each cell, or for those units on which we have no information, we can choose any way we like, including randomly, though randomization has no advantage or disadvantage. Such allocations rule out self-selection (or doctor or program administrator selection) where the individual (doctor, or administrator) has information not visible to the person assigning treatments and controls. The key is that the person who makes the assignment (the analyst) uses all of the information that he or she possesses, and that once this has been taken into account, all units are interchangeable conditional on that information, so that assignment beyond that does not matter. Of course, the program administrators must enforce the analyst's assignment, so that private information that they or the units possess is not allowed to affect the assignment, conditional on the information used by the analyst. Given this, selection on unobservables is ruled out, and does not affect the results. Randomization is *not* required to eliminate selection bias.

Whether it is really possible for the analyst to assign arbitrarily is an open question, as is whether "randomization" from a random-number generator will do so. Even machine-generated sequences have causes, and even if the analyst has only a set of uninformative labels for the units, those too must come from somewhere, so that it is possible that those causes are linked to the unobserved causes in the experiment. We do not attempt to deal here with these deep issues on the meaning of randomization, but see Singer and Pincus (1998).

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 argu-

ment more recently echoed by Worrall (2007). Randomization serves this purpose, but so do other non-discretionary schemes; what is required is that the hidden information not affect the allocation. While it is true that doctors cannot be allowed to make the assignment, it is not true that randomization is the only scheme that can be enforced.

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 excluded by pure randomization. For example, the hypothetico-deductive method works by using theory to make a prediction that can be taken to the data; here the predictions would be of the form that a unit with characteristics *x* will respond in a particular way to treatment, falsification of which can be tested by an appropriate allocation of units to treatment. Banerjee, Chassang and Snowberg (2016) provide such examples.

Third, randomization, by running roughshod over prior information from theory and from the covariates, is wasteful and even unethical when it unnecessarily exposes people, or unnecessarily many people, to possible harm in a risky experiment, see Worrall (2002) for an egregious case of how an unthinking demand for randomization and the refusal to accept prior information put children's lives directly at risk.

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 Banerjee, Chassang and Snowberg (2016) recommend randomized designs, including in medicine and in development economics. They develop a theory of an investigator who is facing an adversarial audience that will challenge any prior information and can even potentially veto results that are based on it (think administrative agencies or journal referees). The experimenter trades off his or her own desire for precision (and preventing possible harm to subjects), which uses prior information, against the wishes of the audience, who want nothing of the priors. Even then, the approval of this 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. Among doctors who use RCTs, and especially meta-analysis, such arguments are (appropriately) common; see again Kramer (2016).

As we noted in the Introduction, much of the public has come to question expert prior knowledge, and Banerjee, Chassang, Montero and Snowberg (2016) have provided an elegant (positive) account of why RCTs will flourish in such an environment. In cases where there is good reason to doubt the good faith of experimenters, as in some pharmaceutical trials, randomization will indeed be the appropriate response. But we believe such arguments are deeply destructive for scientific endeavor and should be resisted as a general prescription for scientific research. Economists and other social scientists know a great deal, and there are many areas of theory and prior knowledge that are jointly endorsed by large numbers of knowledgeable researchers. Such information needs to be built on and incorporated into new knowledge, not discarded in the face of aggressive know-nothing 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."

1.3 *Statistical inference in RCTs*

If we are to interpret the results of an RCT as demonstrating the causal effect of the treatment in the trial population, we must be able to tell whether the difference between the control and treatment means could have come about by chance. Any conclusion about causality is hostage to our ability to calculate standard errors and accurate *p*–values. But this is not generally possible without assumptions that go beyond those needed to support the basic theorem of RCTs. In particular, it has long been known that the mean—and *a fortiori* the difference between two means—is a statistic that is sensitive to outliers. Indeed Bahadur and Savage (1956) demonstrate that, without restrictions on the parent distributions, standard *t*–tests are inherently unreliable.

The key problem here is *skewness*; standard *t*–tests break down in distributions with large skewness, see Lehmann and Romano (2005, p. 466–8). In consequence, RCTs will not work well when the distribution of the individual treatment effects is strongly asymmetric, at least if the standard two-sample *t*–statistics (or equivalently White's (1980) heteroskedastic robust regression *t*–values) are used. While we may be willing to assume that treatment effects are symmetric in some cases, the need for such an assumption—which requires prior knowledge about the specific process being studied—undermines the argument that RCTs are largely assumption free and do not depend on such knowledge. There is a deep irony here. In the search for robustness and the desire to do away with unnecessary assumptions, the RCT can deliver the mean of

the ATE, yet the mean—as opposed to the median, which cannot be estimated by an RCT—does not permit robust probability statements about the estimates of the ATE

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. 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, 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 designed to explicitly model the skewness of expenditures.

In some cases, it will be appropriate to deal with outliers by trimming, 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.

1.3.1 *Spurious statistical significance: an illustrative example*

We consider an example that illustrates what can happen in a realistic but simplified case. There is a *parent population*, or *population of interest*, defined as the collection of units for which we would like to estimate an average treatment effect. It might be all villages in India, or all recipients of food subsidies, or all users of health care in the US. From this population we have a sample that is available for randomization, the *trial or experimental sample*; in a randomized controlled trial, this will subsequently be randomly divided into treatments and controls. Ideally, the trial sample would be randomly selected from the parent sample, so that the sample average treatment effect would be an unbiased estimator of the population average treatment effect; indeed in some cases the complete population of interest is available for the trial. Clearly,

in these ideal cases, it is straightforward to use standard sampling theory to generalize the trial results from the sample to the population. However, for a number of practical and conceptual reasons, the trial sample is rarely either the whole population or a randomly selected subset, see Shadish et al (2002, pp. 341–8) for a good discussion of both practical and theoretical obstacles.

In our 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 average treatment effect is zero. The individual treatment effects β are distributed so that $\beta + e^{0.5} \sim \Lambda(0,1)$, for standardized lognormal distribution Λ. We have something like 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 (experimental) sample of $2n$ individuals is randomly drawn from the parent and is randomly split between *n* treatments and *n* controls. 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 100 trial/experimental samples each of size 2*n*; with five values of *n*, this gives us 500 trial/experimental samples in all. 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.

Table 1: 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 results are shown in Table 1. 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.

The last column shows the fractions of times the true null is rejected and is the 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.

Why does the standard application of the *t*–distribution give such strange results when all we are doing is estimating a mean? 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.

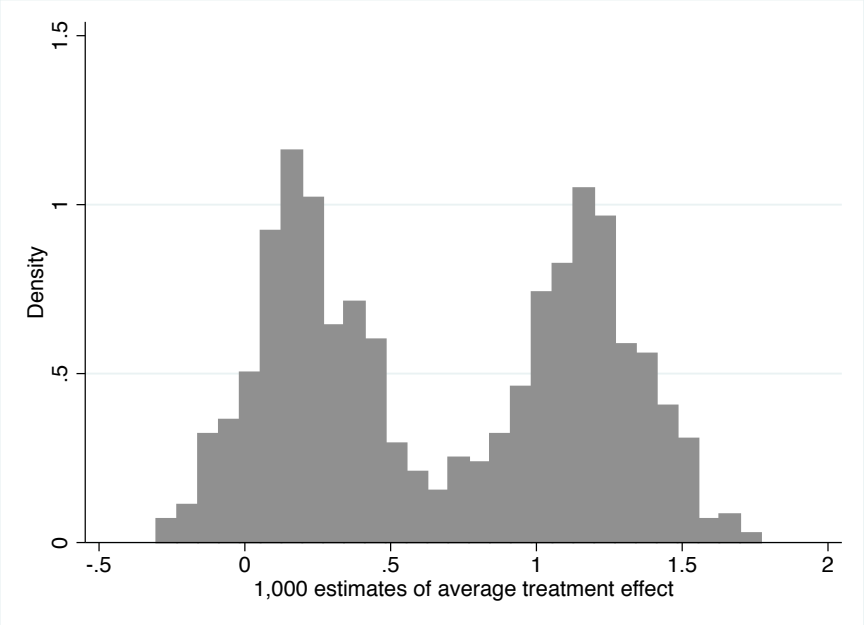

Figure 1: Estimates of an ATE with an outlier in the trial sample

Figure 1 illustrates the estimated average treatment effects from an extreme case from the simulations with 100 observations in total, the second row of Table 1; the histogram shows the 1,000 estimates of the ATE. The 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 not, we get the left-hand side. On the right-hand side, when the outlier is among 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 the left-hand side of the figure 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 large outliers, they illustrate the process that generates the over-rejections and spurious significance.

We could escape these problems if we could calculate the *median* treatment effect, but RCTs cannot (without further assumption) identify the median, only the mean, and it is the mean that is at risk because of the Bahadur-Savage theorem. Note too that there is only moderate comfort to be taken in large sample sizes. While the last row is certainly better than the others, there are still many trial samples that are going to give sample average effects that are significant, even when the number we want is zero. The proof of the Bahadur-Savage theorem works by noting that for *any* sample size, it is always possible to find an outlier that will give a misleading *t*–value. Nor is there an escape here by using the Fisher exact method for inference; the Fisher method tests the null hypothesis that *all* of the treatment effects are zero whereas what we are interested in here, at least if we want to do project evaluation or cost-benefit analysis, is that the *average* treatment effect is zero.

The problems illustrated above, that stem from the Bahadur-Savage theorem, are certainly not confined to RCTs, and occur more generally in econometric and statistical work. However, the analysis here illustrates that the simplicity of ideal RCTs, subtracting one mean from another, brings no exemption from troublesome problems of inference. Escape from these issues, as in the Rand Health Experiment, requires explicit modeling, or might be best handled by estimating quantiles of the treatment distribution, which again requires additional assumptions.

Our reading of the literature on RCTs in development 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, treating all results as solid. 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 theory should 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)—but it hardly solves the problem given that the asymmetry may be in the same direction in different settings (and 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 trial population), and that the "significant" *t*–values will show departures from the null in the same direction, thus replicating spurious findings.

1.2.11: *Significance tests: Fisher-Behrens, robust inference, and multiple hypotheses*

Skewness of treatment effects is not the only threat to accurate significance tests. The two–sample *t*–statistic is computed by dividing the ATE by the estimated standard error whose square is given by

$$\widehat{\sigma}^2 = \frac{(n_1 - 1)^{-1} \sum_{i \in 1} (Y_i - \widehat{\mu}_1)^2}{n_1} + \frac{(n_0 - 1)^{-1} \sum_{i \in 0} (Y_i - \widehat{\mu}_0)^2}{n_0} \tag{5}$$

where 0 refers to controls and 1 to treatments, so that there are n_1 treatments and n_0 controls, and $\hat{\mu}_1$ and $\hat{\mu}_0$ are the two means. As has been long known, this *t*–statistic is not distributed as Student's *t* if the two variances (treatment and control) are not identical; this is known as the Behrens–Fisher problem. 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.

In a remarkable recent paper, Young (2016) argues that this problem gets much worse when the trial results are analyzed by regressing outcomes not only on the treatment dummy, but also on additional controls, some of which might interact with the treatment dummy. Again the problem concerns outliers in combination with the use of 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 2003 regressions reported in 53 RCT papers in the *American Economic Association* journals and recalculates the significance of the estimates using Fisher's randomization inference applied to the authors' original data; see again Imbens and Wooldridge (2009) for a good modern account of Fisher's method. 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; 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 the well-known problem 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.

Our sense is that these issues are being taken more seriously in recent work, especially as concerns multiple hypothesis testing. Young himself is a strong proponent of RCTs in general and believes that randomization inference will yield correct inferences. Yet randomization inference can *only* test the null that *all* treatment effects are zero, that the experiment does nothing to anyone, whereas many investigators are interested in the weaker hypothesis that the *average* treatment effect is zero. This simply makes matters worse since the stronger hypothesis implies the weaker hypothesis and there are presumably undiscovered cases where the ATE is spuriously significant, even when the Fisher test rejects that all treatment effects are zero. Note that testing does not always match logic; it is possible to reject the null that the ATE is zero even when we can simultaneously accept the (joint) hypothesis that all treatment effects are zero; this is familiar from OLS regression, where an *F*–test can show joint insignificance, even when a *t*–test of some linear combination is significant.

It is clear that, as of now, all reported significance levels from RCT results in economics should be treated with considerable caution. Greater care about skewness and outliers would help, as would greater use of the Fisher method and of procedures that deal correctly with multiple hypothesis testing. Yet if the null hypothesis is that the *average* treatment effect is zero, as in most project evaluation, the Fisher test is not available, so that we currently do not have a reliable set of procedures. Robust or clustered standard errors are necessary to allow for the

possibility that treatment changes variances, and the inclusion of covariates is necessary to control for imbalance in finite samples.

1.3 *Blinding*

Blinding is rarely possible in economics or social science trials, and this is one of the major differences from most (although not all) RCTs in medicine, where blinding is standard, both for those receiving the treatment and those administering it. Indeed, the ability to blind has been one of the key arguments in favor of randomization, from Bradford-Hill in the 1950s, see Chalmers (2003), to welfare trials today, Gueron and Rolston (2013). Consider first the blinding of subjects. Subjects in social RCTs usually know whether they are receiving the treatment or not and so can react to their assignment in ways that can affect the outcome other than through the operation of the treatment; in econometric language, this is akin to a violation of exclusion restrictions, or a failure of exogeneity. In terms of (1), there is a pathway from the treatment assignment to another unobserved cause, which will result in a biased ATE. This is not to argue in favor of instrumental variables over RCTs, or vice versa, but simply to note that, without blinding, RCTs do not automatically solve the selection problem any more than IV estimation automatically solves the selection problem. In both cases, the exogeneity (exclusion restriction) argument needs to be explicitly made and justified. Yet the literature in economics gives great attention to the validity of exclusion restrictions in IV estimation, while tending to shrug off the essentially identical problems with lack of blinding in RCTs.

Note also that knowledge of their assignment may cause people to want to cross over from treatment to control, or vice versa, to drop out of the program, or to change their behavior in the trial depending on their assignment. In extreme cases, only those members of the trial sample who expect to benefit from the treatment will accept treatment. Consider, for example, a trial in which children are randomly allocated to two schools that teach in different languages, Russian or English, as happened during the breakup of the former Yugoslavia. The children (and their parents) know their allocation, and the more educated, wealthier, and less-ideologically committed parents whose children are assigned to the Russian-medium schools can (and did) remove their children to private English-medium schools. In a comparison of those who accepted their assignments, the effects of the language of instruction will be distorted in favor of the English schools by differences in family characteristics. This is a case where, even if the random number generator is fully functional, a later balance test will show systematic differences in ob-

servable background characteristics between the treatment and control groups; even if the balance test is passed, there may still be selection on unobservables for which we cannot test.

More generally, when people know their allocation, when they have a stake in the outcome, and when the treatment effect is different for different people, there are incentives and opportunities for selection in response to the randomization, and that selection can contaminate the estimated average treatment effect, see Heckman (1997) who makes the same point in the context of instrumental variables. Those who were randomized by a lottery into going to Vietnam will have different treatment effects depending on their labor market prospects, and those with better prospects are more likely to resist the draft. As we shall see in the next subsection, various statistical corrections are available for a few of the selection problems non-blinding presents, but all rely on the kind of assumptions that, while common in observational studies, RCTs are designed to avoid. Our own view is that assumptions and the use of prior knowledge are what we need to make progress in any kind of analysis, including RCTs whose promise of assumption-free learning is always likely to be illusory.

There may be a tendency in economics to focus on the selection bias effects of non-blinding because some solutions are available, but selection bias is not the only serious source of bias in social and medical trials. Concerns about the placebo, Pygmalion, Hawthorne, John Henry, and 'teacher/therapist' effects are widespread across studies of medical and social interventions. This literature argues that double blinding should be replaced by quadruple blinding; blinding should extend beyond participants and investigators and include those who measure outcomes and those who analyze the data, all of whom may be affected by both conscious and unconscious bias. The need for blinding in those who assess outcomes is particularly important in any cases where outcomes are not determined by strictly prescribed procedures whose application is transparent and checkable but requires elements of judgment; a good example is therapists who are asked to assess the extent of depression in clinical trials of anti-depressants, see Kramer (2016).

The lesson here is that blinding matters and is very often missing. There is no reason to suppose that a poorly blinded trial with random assignment trumps better blinded studies with alternative allocation mechanisms, or matched studies.

1.13 *What do RCTs do in practice?*

The execution of an RCT will often deviate from its design. People may not accept their assignment, controls may manage to get treatment, and vice versa, and people may accept their as-

signment, but drop out before the completion of the study. In some designs, the trial works by giving people incentives to participate, for example by mailing them a voucher that gives them subsidized access to a school or to a savings product. If the aim is to evaluate the voucher scheme itself, no new issue arises. However, if the aim is to find out what the education or savings program does, and the voucher is simply a device to induce variation, much depends on whether or not people decide to use the voucher which, like attrition and crossover, is subject to purposive decisions by the subjects inducing differences between treatments and controls.

Everything depends on the purpose of the trial. In the example above, we may want to evaluate the voucher program, or we may want to find out what the saving product does for people. We are sometimes interested in establishing causality, and sometimes in estimating an average treatment effect; in the economics literature, some writers define internal validity as getting the ATE right, while others, following the original definition of the term, define internal validity as getting causality right. Sometimes the trial limits itself to establishing causality (or to estimating an ATE) in only the trial sample, but some trials are more ambitious, and try to establish causality (or estimate an ATE) for a broader population of interest. When, as is common in economics trials, no limits are placed on the heterogeneity of treatment responses, different trial samples and different populations will generally have different ATEs and may have different casual outcomes, e.g. if the treatment has an effect in one population but none or the opposite effect in another. Our view is that the target of the trial, including the population of interest, needs to be defined in advance. Otherwise, almost any estimated number can be interpreted as a valid ATE for some population, we allow deviations from the design to define our target, and we have no way of knowing whether apparently contradictory results are really contradictory or are correct for the population on which they were derived. Differences in results, between different RCTs and between RCTs and observational studies, may owe less to the selection effects that RCTs are designed to remove, than to the fact that we are comparing non-comparable people, Heckman, Lalonde, and Smith (1999, p. 2082). Without a clear idea of how to characterize the population of individuals in the trial, whether we are looking for an ATE or to identify causality, and for which groups enrolled in the trial the results are supposed to hold, we have no basis for thinking about how to use the trial results in other contexts.

To illustrate some of the issues, consider a simple RCT in which a treatment T is administered to a trial sample that is split between a treatment group of size n and a control group of size n, but that only a fraction p of the treatment group accepts their assignment, with fraction

$(1-p)$ receiving no treatment. Suppose that the parameter of interest is the ATE in the original population, from which the trial sample was drawn randomly. Denote by β the hypothetical ideal ATE estimate that would have been calculated if everyone had accepted assignment; as we have seen, this is an unbiased estimator of the parameter of interest for both the trial sample and the parent population. β cannot be calculated, but there are various options.

Option one is to ignore the original assignment and calculate the difference in means between those who received the treatment and those who did not, including among the latter those who were intended to receive it but did not. Denote this ("as treated") estimate β_1. Alternatively, option two, is to compare the average outcome among those who were *intended* to be treated and those who were intended to be controls. Denote this estimate, the "intent to treat" (ITT) estimator, β_2. It is easy to show that one set of conditions for $\beta_1 = \beta$ is that those who were treated have the same ATE as those who were intended to be treated, and that those who broke their assignment have the same untreated mean as those who were assigned to be controls, conditions that may hold in some applications, for example where the treatment effects are identical.

The ITT estimator, β_2, will typically be closer to zero than is β, and it will certainly be so if the average treatment effect among those who break their assignment is the same as the overall ATE, in which case $\beta_2 = p\beta$. For these reasons, the ITT is often described as yielding a conservative estimate and is routinely advocated in medical trials even though it is an attenuated estimator of the ATE. A third estimator, β_3, the local average treatment estimator (LATE) is computed by running a regression of outcomes on an (actual) treatment dummy using the treatment assignment as an instrumental variable. In this case, the LATE is simply the ITT, scaled up by the reciprocal of *p*, so that $\beta_3 = \beta_2 / p$. From the above, the LATE is β if the average treatment effect of those who break their assignment is the same as the average treatment effect in general, so that the ITT estimator is biased down by counting those who should have been treated as if they were controls. More generally, and with additional assumptions, Imbens and Angrist (1994) show that the LATE is the average treatment effect among those who were induced to accept the treatment by their assignment to treatment status, which can be a very different object from the original target of investigation. These various estimators, the ATE, the ITT, and the LATE, are all averages over different groups; more formally, Heckman and Vytlacil (2005) define a marginal treatment effect (MTE) as the ATE for those on the margin of treat-

ment—whatever the assignment mechanism—and show that the other estimators can be thought of as averages of the MTEs over different populations.

In general, and unless we are prepared to say more about the heterogeneity in the treatment effects, the three estimators will give different results because they are averages over different populations. Economists tend to believe that people act in their own interest, at least in part, so it is not attractive to believe that those who break their assignments have the same distribution of treatment effects as do those who accept them. In Heckman's (1992) analogy, people are not like agricultural plots, which are in no position to evade the treatment when they see it coming. Such purposive behavior will generally also affect the composition of the trial sample compared with the parent population, with those who agree to participate different from those who do not. For example, people may dislike randomization because of the risks it entails, or people may seek to enter trials in the hope that they will receive a beneficial treatment that is otherwise unavailable. A famous example in economics is the Ashenfelter (1978) pre-program "dip," where those who enter trials of training programs tend to be those whose earnings have fallen immediately prior to enrolment, see also Heckman and Smith (1999). People who participate in drug trials are more likely to be sick than those who do not, or are likely to be those who have failed on standard medication. Another example is Chyn's (2016) evidence that those who applied for vouchers in the Moving to Opportunity experiment and were thus eligible for randomization—and only a quarter of those who were eligible actually did so—were those who were *already* making unusual efforts on their children's behalf. These parents had effectively substituted for part of the better environment, so that the ATE from the trial understates the benefits to the average child of moving. Similar phenomena occur in medicine. In the 1954 trials of the Salk polio vaccine in the US, the rates of infection, while lowest among the treated children, were higher in the control children than in the general population at risk, so that the parents of those who selected into the trial presumably had some idea that they might have been exposed, Hausman and Wise (1985, p. 193–4). In this case, the average treatment effect in the trial sample exaggerates the ATE in the general population, which is what we want to know for public policy.

Given the non-parametric spirit of RCTs, and the unwillingness of many trialists to make assumptions or to incorporate prior information, the only way forward is to be very clear about the purpose of the trial and, in particular, which average we are trying to estimate. For those who focus on internal validity in terms of establishing causality by finding an ATE significantly

different from zero, the definition of the population seems to be a secondary concern. The idea seems to be that if causality is established in *some* population, that finding is important in itself, with the task of exploring its applicability to other populations left as a secondary matter. For the many economic or cost–benefit analyses where the ATE is the parameter of interest, the population of interest is definitional, and the inference needs to focus on a path from the results of the trial to the parameter of interest. This is often difficult or even impossible without additional assumptions and/or modeling of behavior, including the decision to participate in the trial, and among participants, the decision not to drop out. Manski (1990, 1995, 2003) has shown that, without additional evidence, the population ATE is not (point) identified from the trial results, and has developed non-parametric bounds (an interval estimate) for the ATE. As with the ITT, these bounds are sometimes tight enough to be informative, though the interval defined by the bounds will often contain zero, see Manski (2013) for a discussion aimed at a broad audience. Faced with this, many scholars are prepared to make assumptions or to build models that give more precise results.

RCTs may tell us about causality, even when they do not deliver a good estimate of the ATE. For example, if the ITT estimate is significantly different from zero, the treatment has a causal effect for at least some individuals in the population. The same is true if the LATE is significantly different from zero; again the treatment is causal for some sub-population, even if we may have difficulty characterizing it or accepting it as the population of interest. From this, we also learn that, provided we had a population with the right distribution of $\beta_i's$ and governed by the same potential outcome equation, the treatment would produce the effect in at least some individuals there.

Section 2: Using the results of randomized controlled trials

2.1 *Introduction*

Suppose we have the results of a well-conducted RCT. We have estimated an average treatment effect, 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 sample population, up to the limits of statistical inference. What are such findings good for? How should we use them?

The literature in economics, as indeed in medicine and in social policy, has paid more attention to obtaining results than to whether and how they should be adapted for use, often as-

suming that findings can be used "as is." Much effort is devoted to demonstrating causality and estimating effect sizes in study populations, both in empirical work—more and better RCTs, or substitutes for RCTs, such as instrumental variables or regression discontinuity models—as well as in theoretical statistical work—for example on the conditions under which we can estimate an average treatment effect, or a local average treatment effect, and what these estimates mean. There is less 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 more than an estimate that was plucked out of thin air. If trials are to be useful, we need paths to their use that are as carefully constructed as are the trials themselves.

It is sometimes assumed that a parameter, once well established, is invariant across settings. The parameter may be difficult to estimate, because of selection or other issues, and it may be that only a well-conducted RCT can provide a credible estimate of it. If so, internal validity is all that is required, and debate about using the results becomes a debate about the conduct of the study. The argument for the "primacy of internal validity," Shadish, Cook, and Campbell (2002), is 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 is the default assumption failing arguments to the contrary. An invariance argument is often made in medicine, 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, and we are interested in using the result outside the original context. If "the same" result holds elsewhere, we say we have exter-

nal validity, otherwise not. External validity may refer just to the transportability of the causal connection, or go further and require replication of the magnitude of the average treatment effect. 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; it both overstates and understates the value of the results from an RCT. It directs us toward simple extrapolation—whether the same result will hold elsewhere—or simple generalization—whether it holds universally or at least widely—and away from possibly more complex but more useful applications of the evidence. Just as internal validity says nothing about whether or not a trial result will hold elsewhere, 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. Even successful replication by itself tells us little either for or against simple generalization or extrapolation. Without further understanding and analysis, even multiple replications cannot provide much support for, let alone guarantee, the conclusion that the next will work in the same way. Nor do failures of replication make the original result useless. We can often learn much from coming to understand why replication failed and use that knowledge to make appropriate use of the original findings, not by expecting replication, but by looking for how the factors that caused the original result might be expected to 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 provides an excellent example of the limitations to straightforward extrapolation from repeated successful replication. The bird infers, based on multiply 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 Christmas dinner. Of course, our chicken did not base her inference on an RCT. But had we constructed

one for her, we would have obtained exactly the same result that she did. Her problem was not her methodology, but rather that she was studying surface relations, and that she did not understand the social and economic structure that gave rise to the causal relations that she observed. So she did not know how widely or how long they would obtain. Russell notes, "more refined views as to the uniformity of nature would have been useful to the chicken" (1912, p. 44). We often act as if the methods of investigation that served the chicken so badly will do perfectly well for us.

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 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 (perhaps provocatively) be labeled 'anecdotal causality'. The operation of a cause generally requires the presence of support or helping 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 helping 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 the case where a set of causes are jointly but not separately sufficient for an effect. If we rewrite (3) in the form

$$Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij} = \left(\sum_{k=1}^{K} \theta_k w_{ik} \right) T_i + \sum_{j=1}^{J} \gamma_j x_{ij} \qquad (6)$$

where θ_k controls how w_{ik} affects individual *I*'s treatment effect β_i. The "helping" or "support" factors for the treatment are represented by the interactive variables w_{ik}, among which may be included some *x*'s. Since the ATE is the average of the β_i's, two populations will have the same ATE only if, except by accident, they have the same average for the support factors necessary for the treatment to work. 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 dif-

ferent development (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), or another where a long chain of ropes and pulleys causes the insertion of food into the mouth to activate a face-wiping napkin. These are causal machines, but they are specially constructed to give a kind of causality that operates extremely locally and has no general applicability. The underlying structure affords a very specific form of (6) 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 (6) 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 machines 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. Conditional cash transfers 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 disprove a general theoretical proposition to which it provides a counterexample. The test might be of the general proposition itself (a simple refutation test), or of some consequence of it that is susceptible to testing using an RCT (a complex refutation test). Of course, counterexamples are often challenged—for example, it is not the general proposition that caused the rejection, but a special

feature of the trial—but here we are on familiar inferential turf. 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. Once again, this is 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.

Another case where no transportation is called for is when an RCT is used for evaluation, for example to satisfy donors that the project they funded actually 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 the development of 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, p.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, and that the something needed 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 average treatment effect in a well-defined population from which the sample trial population—from which treatments and controls are randomly assigned—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 the public-health-style uses of RCTs is that the 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 common in health. There is also an issue of timing if the results are to be implemented some time after the trial.

In economics, a 'public-health-style' example is the imposition of a commodity tax, where the total tax revenue is of interest and we do not care who pays the tax. Indeed, theory can often identify a specific, well-defined magnitude whose measurement is key for the 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 interest of the sponsors is in the consequences for the budget of the state responsible for the program; 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 development 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 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," Kremer and Duflo (2008, p. 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) use a Kenyan trial as the basis for a blanket statement without context restriction, "Provision of free school uniforms, for example, leads to 10%-15% reductions in teen pregnancy and drop out rates." Kremer and Duflo (2008), 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."

Of course, the problem of generalization extends beyond RCTs, to both "fully controlled" laboratory experiments and to most non-experimental findings. For example, ever since

Alfred Marshall thought of it while sunbathing, economists have used the concept of an elasticity—as in the income elasticity of the demand for food, or the price elasticity of the supply of cotton—and have transported elasticities—which are conveniently dimensionless—from one context to another, as numerical estimates, or in ranges, such as high, medium, or low. Articles that collect such estimates are widely cited even though, as has long been known, the invariance of elasticities is not guaranteed in practice and is sometimes inconsistent with choice theory. Our argument here is that evidence from RCTs, like evidence on elasticities, is *not* automatically simply generalizable, and that its internal validity, when it exists, does not provide it with any unique invariance across context. We shall also argue that specific features of RCTs, such as their freedom from parametric assumptions, although advantageous in estimation, can be a serious handicap in use.

Most 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 only of whether health insurance increased expenditures, but how much it did so 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, Conlisk (1973). Experiments on time-of-day pricing for electricity had a similar structure, see Aigner (1985).

The MDRC experiments have also been analyzed across cities in an effort to link city features to the results of the RCTs within them, 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 assembles evidence on a large number of trials, and 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 those 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 these 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 almost any 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, causally interprets the difference in a way that allows at least some of the findings to survive.

How can we think about this more seriously? How can we do better than simple generalization and simple extrapolation? Many writers have emphasized the role of *theory* in transporting and using the results of trials, and we shall discuss this further 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 (6), suppose that the $\beta_i's$, the individual treatment effects, are functions of a set of *K* observable or unobservable support variables, w_{ik}, and that the non-vacuous *w*'s may even represent different features in different places. It is then clear that, provided the distribution of the *w* values is the same in the new circumstances as the old, then 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, the ATEs within each of the subgroups calculated, and then reassembled according to the configuration of *w*'s in the new context. For example, if the treatment effects vary with age, the age-specific ATEs can be estimated, and the age distribution in the new context used to reweight the age-specific ATEs to give a new, overall, ATE. 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. Of course, this method will only work in special cases; for example, if we only know some of the *w*'s, there is no reason to suppose that reweighting for those alone will give a useful correction.

Other methods also work when there are too many *w*'s for stratification, for example by estimating the probability of each observation in the population being 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).

There are yet further reasons why these methods do not always work. As with any form of reweighting, the variables used to construct the weights must be present in both the original and new context. If treatment effects vary by sex, we cannot predict the outcomes for men using a trial sample that is entirely female. 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. It also depends on assuming that the same governing equation (6) covers the trial and the target population. If they differ not only by what causal factors are present in what proportions but also in *how* (if at all) the causes contribute to the effects, re-weighting the effect sizes that occur in trial sub-populations will not produce good predictions about target population outcomes.

It should be clear from this that reweighting works only when the observable factors used for reweighting include all and only genuine interactive causes; we need data on *all* the relevant interactive factors. But as Muller (2015) notes, this 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.

Pearl and Bareinboim (2014) use Pearl's do–calculus to provide a fuller formal analysis for transportability of causal empirical findings across populations. They define transportability as "a license to transfer causal effects learned in RCTs to a new population, in which only observational studies can be conducted," Pearl and Bareinboim (2015, p. 1). They consider both qualitative causal relations, which they represent in directed acyclic graphs, and probabilistic facts, such as the conditional probability of the outcome on a treatment conditional on some third factor. They then provide theorems about what the relationship between the causal and probabilistic facts in two populations must be if it is to be possible to infer a particular causal fact, such as the ATE, about population 2 from causal and probabilistic information about population 1 coupled with purely probabilistic information about population 2. Not surprisingly, for many things we should like to know about population 2, knowledge of even the full structure on population 1 will not suffice. Inferences to facts about a new population require not only that the

facts we suppose about population 1—like an ATE—are well grounded, that the RCT was well conducted, that the statistical inference is sound—but that we have equally good grounding for other assumptions we need about the relation between the two populations. For example, using the result described above for directly transporting the ATE from a trial population to some other—simple extrapolation—we need good grounds to suppose both that the average of the net effect of the interactive factors is the same in both populations and also that the same governing equation describes both populations.

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. We need to think through the causal chain that has generated the RCT result, and the underlying structures that support this causal chain, whether that causal chain might operate in a new setting and how it would do so with different joint distributions of the causal variables; we need to know *why* and whether that *why* will apply elsewhere. While it is true that there exist general causal claims—the force of gravity, or that people respond to incentives—they use relatively abstract concepts and operate at a much higher level than the claims that can be reasonably inferred from a typical RCT, and cannot, by themselves, guarantee the outcomes that we are considering here. That transportation is far from automatic also tells us why (even ideal) RCTs of similar interventions can be expected to 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.

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 the fraction of people near the relevant cutoff. We also need to know that the same mechanism works in any new setting where we consider using small subsidies to increase school attendance.

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 necessarily 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 a range of empirical strategies including observational studies. We will also need to be able to characterize the population to which the original RCT and its ATE applied because how the population is described is commonly taken to be some indication of which other populations the results are likely to be exportable to and which not. 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, p14) 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." The problems of characterizing the population here goes beyond those we faced in considering a LATE. An RCT is conducted on a population of specific individuals. The results obtained, whether we think in terms of an ATE or in terms of establishing causality, are features of that population, 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 population satisfies. How is the description of the population that is used in reporting the results to be chosen? For choose we must—the alternative to describing is naming, identifying each individual in the study by name, which is cumbersome and unhelpful and often unethical.

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 population 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 significant substantial assumptions about what might be and what might not be relevant to the production of the outcome studied. (For example, the ICMJE guidelines 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, not just RCTs. But RCTs require

special conditions if they are to be conducted at all and especially if they are to be conducted successfully—local agreements, compliant subjects, affordable administrators, 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 the study population should have are not justified, made explicit, or subjected to serious critical review. This careful description of the study population is uncommon in economics, whether in RCTs or many observational studies.

The need for observational knowledge is one of many reasons why it is counterproductive to insist that RCTs are the unique 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 itself. 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 new findings must be able to explain previous results, even results that are now thought to be invalid; methodological prejudice is not an explanation.

These considerations can be seen in practice in the range of randomized controlled trials in economics, which we shall explore in the final subsection below.

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 long and 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; more recent examples are 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) and Duflo, Hanna and Ryan (2012). These structural models sometimes require formidable auxiliary assumptions on functional forms or the distributions of unobservables, which makes many economists reluctant to embrace them, but they have compensating advantages, including the ability to integrate theory and evidence, to make out-of-sample predictions, and to analyze welfare—which always requires some understanding of why things happen—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, which is what, as we keep repeating, the trial results tell us about) respond to some policy is rarely enough to tell whether or not they are made better off. What works is not equivalent to what should be.

In many papers, Heckman has developed ways to model how the beliefs and interests of participants affect their participation in, behavior during, and their outcomes in trials, for example using a Roy model of choice; see e.g. Heckman and Smith (1995), and more recently Chassang, Padró I Miguel, and Snowberg (2012) and Chassang et al (2015). The modeling of beliefs and behavior allows predictions about the results of trials that differ from the base trial, or where the risk and reward structures are different. Beyond that, and in line with a running theme of this Section, thinking about how to handle new situations can be incorporated into the design of the original trial so as to provide the information needed for transportation.

Light touch theory can do much 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 was 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, and 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, allows Arrow to use this 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 we have, a suitable combination of (elementary and transparent) theoretical assumptions and observational data in order adapt and use the trial results. Such analysis can also help design the original trial by clarifying what we need to know in order to be able 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) make this point and provide examples. In the labor supply case, an increase in the tax rate *t* has the same effect as a decrease in the wage rate *w*, 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 wage rates 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 think of 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 the population 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 re-*

fuse 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

$$(t_1 - t_0)w_0h_0 + \frac{1}{2}s_{00}(t_1 - t_0) > G_1 - G_0 \tag{7}$$

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 actually implemented. Both the ITT 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 (3) 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, see again Heckman, Lalonde and Smith (1999). When people are allowed to reject their randomly assigned treatment according to their own (real or perceived) individual advantage, 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*

A typical RCT, especially in the development context, is 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—that failure of the trial results to replicate at a larger scale is likely to be the rule rather than the exception—as well as to note once again that, as in failures of transportability, 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. Using RCT results is not the same as assuming the same results holds in all circumstances.

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 or insecticide 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 the basic point is that the scaled-up effect in this case is *opposite in sign* to the trial effect. The problem here 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 (1999) and Hsieh and Urquiola (2002). Educational or training programs may benefit those who are treated, but harm those left behind; if the control group is selected from the latter, the RCT may generate a positive result in spite of hurting some and helping none; 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. 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 an example, and immunization programs affect the dynamics of disease transmission through herd

immunity, so that the effects on an individual depend on how many others are vaccinated, Fine and Clarkson (1986), Manski (2013, p 52). The usual, if seldom correct, conception of an RCT in medicine is of a biological process—for example, the administration of aspirin after a heart attack—where the effect is thought to be similar across individuals, and where there are no interactions. Yet even here, the social and economic setting 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 actually (or potentially) included in the trial. A well-conducted RCT delivers an average treatment effect for a well-defined population but, in general, that average does *not* apply to everyone. It is not true, for example, as argued in *JAMA'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," Guyatt et al (1994). 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*, though such a conclusion would be better supported by a Fisher randomization test.

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). 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 they rely on to help them identify features in a particular patient that are likely to affect the effectiveness of a given treatment for that patient. But there is an odd balance being struck here. These judgments are deemed admissible in dealing with the individual patient, at least for discussion with the patient as possible considerations, 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, see Kahneman and Klein (2009).

Whether or not averages are useful to individuals raises the same issue in 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, or at least were eligible to be so. 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 simply 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 US 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 very similar to recommendations in development economics, is that the intervention be demonstrated effective through well-designed RCTs in more than one site of implementation, and that "the trials should demonstrate the intervention's effectiveness in school settings similar to yours" (2003, p. 17). No operational definition of "similar" is provided.

We note finally that these caveats, which apply to individuals (or schools) even if they were in the trial, provide another reason why the concept of "external" validity is unhelpful. The real issue is how to *use* the findings of a trial in new settings, including settings included in the trial; external validity in the sense of invariance of the ATE emphasizes simple replication, which guarantees nothing, while ignoring the possibility that lack of replication can be a key to understanding.

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. IThink 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, Levy (2001), nor in places where people strongly oppose education or vaccination.

Similarly, given that the helping 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 conditional cash transfers on school attendance; all but four show the expected positive effect, and the range runs from –8 to +38 percentage points. 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 should transport from one place to another—is not well defined. 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 ATE in the original RCT can differ depending on whether the outcome is measured in levels or in logs; the two ATEs could even have different signs.)

Deworming is surely more complicated than conditional cash transfers though not because anyone disputes the desirability of removing parasitical worms or the biological efficacy of the medicines, at least if they are repeatedly and effectively administered; that is the part of the causal process that is transportable from one place to another. Yet nutritional or school attendance outcomes depend on reinfection from one person to another—which depends on local customs about defecation (which vary from place to place and are subject to religious and cultural factors), particularly on the extent of open defecation and the density of population, on whether or not children wear shoes, and on the availability and use of public and private sanitation; this last was crucial in the elimination of hookworm in the southern states of the U.S. according to Stiles (1939). Temperature may also be important; indeed, such "macro" variables are likely to be important in a wide range of medical, employment, and production trials, Rosenzweig and Udry (2016). There are two prominent positive studies in the economics literature, one in Kenya, Kremer and Miguel (2000) and one in India, Bobonis, Miguel and Puri-Sharma (2006); these are often cited as examples of the power of RCTs to come up with the "right" answer, for example by Karlan and Appel (2008). Yet the Cochrane Collaboration review of deworming and schooling, Taylor-Robinson et al (2015), which reviews one trial (from India) covering more than a million participants, and 44 others covering 67,672 participants, including Kremer and Miguel (2004), conclude that there is "substantial evidence" that deworming shows no benefit in nutritional status, hemoglobin, cognition, school performance or death. The validity of this meta-analysis is disputed by Croke et al (2016). A replication, Aiken et al (2015) and reanalysis (using different methods) of Miguel and Kremer's original data by Davey et al (2015) concluded that the study "provided some evidence, but with high risk of bias," provoking a lengthy exchange, Hicks et al (2015) and Hargreaves et al (2015). Most of the differences in results come from different methodological choices, themselves largely based on disciplinary traditions, rather from the effects of mistakes or errors. In an impressive and clear reanalysis, Humphreys (2015) argues that one puzzling feature of Miguel and Kremer's results is the ab-

sence of any clear effect of deworming on health, as was the case in the large Indian RCT. Yet the effects of deworming on education, which are the main target of the paper, presumably work through health, so that the absence of health effects—a failure of expected mediators—is a puzzle, see also Miguel, Kremer and Hicks (2015), and Ahuja et al (2015). Recall too our earlier discussion of the difficulty of interpreting the standard errors of the original study in the absence of randomization.

It is not our purpose here to try to adjudicate these competing claims but rather to relate this work to our general argument. First, it is not clear that there is a right answer to be discovered; given the causal chains involved, deworming might be helpful in one place but unhelpful in another. Yet the focus of the debate is almost entirely on internal validity, on whether the original studies were correctly done. The Cochrane review, in line with this, and in line with much meta-analysis of trials, seems to suppose that there is a single effect to be uncovered that, once established, will be invariant to local and environmental differences. External validity, it seems, is implied by internal validity. Indeed, Chalmers, one of the founders of the Cochrane Collaboration, has explicitly argued (in response to one of us) that, in the absence of strong reasons to the contrary, results should be taken as applicable everywhere, Pettigrew and Chalmers (2011).

Second, the debate makes it clear that the practice of RCTs in economic development has done little to fulfill the original promise that their simplicity—how hard is it to subtract one mean from another?—would dispose of the methodological and econometric disputes that characterize so many observational studies and were thought to be one of their main flaws. While RCTs tend to take some contentious issues of identification off the table, they leave much to be disputed, including the handling of factors that interact with treatment effects, the appropriate level of randomization, the calculation of standard errors, the choice of outcome measure, the inclusion criteria for the sample, placebo and Hawthorne effects, and much more. The claim that RCTs cut through the usual econometric disputes to deliver to policymakers a simple, convincing, and easily understood answer is simply false. The deworming debates are perhaps the leading illustration.

Much of the development 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 head 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 effect sizes 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 superior—standardization from the effect size standardization discussed above.)

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 African or non-African 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 average treatment effects 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 average treatment effect 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, 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, so that many of the familiar caveats do not apply. 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 long 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 unclear 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 average treatment effect 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. In psychology, where the concept originated, there are endless disputes about what should and should not be pooled in a meta-analysis. Beyond that, as argued by Simpson (2016), 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 manipula-

tion by exclusion rules. It makes no sense to claim replicability on the basis of effect sizes, let alone to use them to rank projects.

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 entirely 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 such conditions exist, and that would likely be unnecessary when they exist. Finding out what works is not the magic key to economic development. Technical knowledge, though always worth having, requires suitable institutions 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 essentially 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 are susceptible to the critique that they 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, rather than because of the people who will be called to work it in reality.

There is also much to be learned from many years of economic trials in the United States, particularly from the work of the Manpower Demonstration Research Corporation (now known by its initials 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. MRDC'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 which are often the subject of the trial. Should there be work requirements? Should there be remedial educational before work requirements? 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 faith that RCTs can reveal the truth. At the Federal level, prospective policies are vetted by the non-partisan

Congressional Budget Office, which makes its own estimates of the budgetary implications of the program. Ideologues whose programs score poorly by the CBO have an incentive to support an RCT, not to convince themselves, but to convince their opponents; once again, RCTs are especially 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 always give the right answer, at least for the budgetary implications, which, rather than the wellbeing of the recipients, were often the primary (and indeed sometimes the only) 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). Note that 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 for each poor person treated, something that a good RCT conducted on a representative sample of the target population is equipped to deliver, at least in the absence of general equilibrium effects, timing effects, etc.

These RCTs by MDRC and other contractors deserve much credit. They 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 of those who were originally opposed. There are also limitations; the trials appear to have had at best a limited influence on scientific thinking about behavior in labor markets. 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 very beginning to peer into the black boxes. Without such mechanisms, transportability is always in doubt, it is impossible for policymakers or academics to purposively improve the policies, and the contributions to cumulative science are severely limited.

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 the question of whether more generous insurance would cause 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 account points, once again, to the central importance of transportability for the usefulness and long-term usefulness 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

RCTs are the ultimate in credible estimation of average treatment effects in the population being studied because they make so few assumptions about heterogeneity, causal structure, choice of variables, and functional form. They are truly nonparametric. And indeed, this is sometimes just what we want, particularly where we have little credible prior information. 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 excessive heterogeneity in responses, and 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 all we get are means. Yet, in the presence of outliers, means themselves do not provide the basis for reliable inference. 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 by permitting factors other than the treatment to affect the outcome, akin to a failure of exclusion restrictions in instrumental variable analysis.

The lack of structure can become 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, we are in trouble. We cannot use the results to help make predictions elsewhere 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 elsewhere, or to use the estimates out of the original context. Simple generalization and simple extrapolation just 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.

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 or not the intervention is actually welfare improving, see Harrison (2014) for a vivid account that sharply identifies this and other issues. Without knowing *why* things happen and *why* people do things, we run the risk of worthless casual ("fairy story") causal theorizing and have essentially 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.

In the specific context of development that has concerned us here, RCTs 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 by governments, whose motives and operating principles are different from the NGOs who typically run trials. While it is true that abstract knowledge is always likely to be beneficial to economic development, successful development depends on institutions and on politics, matters on which RCTs have little to say. In the end, RCTs are one of the many external technical fixes that have meandered off and on the development stage since the Second World War, including building infrastructure, getting prices right, and service delivery, none of which have faced up to the essential domestic political foundations for development.

Citations

Ahuja, Amrita, Sarah Baird, Joan Hamory Hicks, Michael Kremer, Edward Miguel, and Shawn Powers, 2015, "When should governments subsidize health? The case of mass deworming," *World Bank Economic Review*, 29, S9–S24.

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.

Aiken, Alexander M., Calum Davey, James R. Hargreaves and Richard J. Hayes, "Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication," *International Journal of Epidemiology*, 0(0), 1–9.

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.

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, "Estimating the effect of training programs on earnings," *Review of Economics and Statistics*, 60(1), 47–57.
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.
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 Columbia," 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, 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.
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
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.
Bobonis, Gustavo, Edward Miguel, and Charu Puri-Sharma, 2006, "Anemia and school participation," *Journal of Human Resources*, 41(4), 692–721.
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. RandMcNally.

Cartwright, Nancy, 1994, *Nature's capacities and their measurement*. Oxford. Clarendon Press.

Cartwright, Nancy, and Jeremy Hardie, 2012, *Evidence based policy: a practical guide to doing it better*, Oxford. Oxford University Press.

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.

Chalmers, Iain, 2003, "Fisher and Bradford Hill: theory and pragmatism?" *International Journal of Epidemiology*, 32, 922–24.

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

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.

Croke, Kevin, Joan Hamory Hicks, Eric Hsu, Michael Kremer, and Edward Miguel, 2016, "Does mass deworming affect children's nutrition? Meta analysis, cost effectiveness, and statistical power," Cambridge, MA. NBER Working Paper No. 22382 (July.)

Cronbach, Lee J., S. R. Ambron, S. M. Dornbusch, R. D. Hess, R. C. Hornick, D.C. Phillips, D. F. Walker, and S. S. Weiner, 1980, *Towards reform of program evaluation*, San Francisco, Jossey-Bass.

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, Calum, Alexander M. Aitken, Richard J. Hayes, and James R. Hargreaves, 2015, "Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped wedge trial," *International Journal of Epidemiology*, 0(0), 1–12.

Deaton, Angus, and John Muellbauer, 1980, *Economics and consumer behavior*, New York. Cambridge University Press.

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, 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., 2006, "Statistical models for causation: what inferential leverage do they provide?" *Evaluation Review*, 30: 691–713.

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

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.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch, *Impact evaluation in practice*, Washington, DC. The World Bank.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz, 2004, "Retrospective vs. prospective analyses of school inputs: the case of flip-charts in Kenya," *Journal of Development Economics*, 74, 251–68.

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.

Hargreaves, James R., Alexander M. Aiken, Calum Davey, and Richard J. Hayes, 2015, "Authors' response to: deworming externalities and school impacts in Kenya," *International Journal of Epidemiology*, 0(0), 1–3.

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.

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., 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., 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 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.
Hicks, Joan Hamory, Michael Kremer, and Edward Miguel, 2015, "Commentary: deworming externalities and schooling impacts in Kenya: a comment on Aiken et al (2015) and Davey et al. (2015)," *International Journal of Epidemiology*, 0(0), 1–4.
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.
Humphreys, Macartan, 2015, "What has been learned from the deworming replications: a nonpartisan view," Columbia University, Aug. http://www.columbia.edu/~mh2245/w/worms.html
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 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.)
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*, 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
Kendall, Maurice G., 1959, "Hiawatha designs an experiment," *American Statistician*, 13(5), 23–4.
Kramer, Peter, 2016, *Ordinarily well: the case for antidepressants*, 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.
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., 1990, "Nonparametric bounds on treatment effects" *American Economic Review*, 80(2), 319–23.
Manski, Charles F., 1995, *Identification problems in the social sciences*, Cambridge, MA. Harvard University Press.
Manski, Charles F., 2003, *Partial identification of probability distributions*, New York. Springer.
Manski, Charles F., 2013, *Public policy in an uncertain world: analysis and decisions*, Cambridge, MA. Harvard University Press.
Metcalfe, Charles E., 1973, "Making inferences from controlled income maintenance experiments," *American Economic Review*, 63(3), 478–83.
Miguel, Edward, and Michael Kremer, 2004, "Worms: identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 72(1), 159–217.
Miguel, Edward, Michael Kremer, and Joan Hamory Hicks, 2015, "Comment on Macartan Humphreys' and other recent discussions of the Miguel and Kremer (2004) study," Berkeley, Dec. http://emiguel.econ.berkeley.edu/assets/miguel_research/63/Worms-Comment_2015-12-21.pdf
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, 2009, *Causality: models, reasoning, and inference*, 2nd edition, Cambridge. Cambridge University Press.
Pettigrew, Mark, and Iain Chalmers, 2011, "Use of research evidence in practice," *Lancet*, 378(9804), 1696.
Rodrik, Dani, 2006, personal email communication.
Rosenzweig, Mark and Christopher Udry, 2016, "External validity in a stochastic world," Cambridge, MA. NBER Working Paper 22449 (July).
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.
Scriven, Michael, 1974, "Evaluation perspectives and procedures," in W. James Popham, ed., *Evaluation in education—current applications*, Berkeley, CA. McCutchan Publishing Corporation.
Sen, Amartya K., 2011, *The idea of justice*, Cambridge, MA. Harvard University Press.
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, 2016, "Comparing and combining standardized effect sizes: the misdirection of public policy," Working Paper, University of Durham (July).
Singer, Burton H., and Steve Pincus, 1998, "Irregular arrays and randomization," *Proceedings of the National Academy of Sciences of the USA*," 95, 1363–8.
Stiles, Charles Wardell, 1939, "Early history, in part esoteric, of the hookworm (uncinariasis) campaign in our southern United States," *Journal of Parasitology*, 25(4), 283–308.
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.
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
Taylor-Robinson, David C., Nicola Maayan, Karla Soares-Weiser, Sarah Donegan, and Paul Garner, 2015, "Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance (review)," The Cochrane Collaboration. Wiley. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000371.pub6/abstract
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.

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.

Worrall, John, 2002, "*What* Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69, S316-S330.

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

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.