2

A DEFENSE OF OBSERVATIONAL RESEARCH

Susan C. Stokes

1. Introduction

Experimental research is quickly gaining ground in the social sciences. Building on a rich tradition of experimentation going back to Gosnell's field experiments on voter turnout in Chicago in the 1920s, political scientists have devised laboratory experiments to study (among other topics) media effects on voters, attitudes about race, and distribution rules on collective action, while field experiments cover an increasingly wide range, from voter turnout to voting behavior to corruption and the rule of law. Beginning in the 1990s experimental techniques have become widespread in development and labor economics. Economists have conducted experiments to test the effect of a wide range of interventions, from varying interest rates on the repayment of microloans, to deworming on school attendance, to charging for antimosquito bed nets on the incidence of malaria. Psychology has long been an experimental discipline.[1]

Experiments have contributed to our basic knowledge of causal effects in the social world. When they are feasible, ethically acceptable, and cost-effective they are clearly a valuable research tool. Criticisms of various aspects of experimental research—problems of implementation such as compliance and spillover, problems of external validity, and the scope of the questions that can be addressed—have come from experimentalists and from outsiders. My main objective here is not to criticize social

science experimentation. Instead it is to describe and criticize a set of beliefs that a growing number of social scientists hold about observational research. I contend that if these beliefs were applied evenhandedly to experimental studies, we would give up on observation and experiments alike as contributing to the building and testing of theories about the social world.

In the section that follows, I characterize the beliefs entailed in *radical skepticism* of observational research. I do not prove these beliefs to be incorrect, but I do offer reasons why they are unlikely to be warranted. In the section entitled "Radical Skepticism and Experimental Research" I argue that, if one were to embrace radical skepticism and apply it evenhandedly to experimental research, one would despair of the possibility that such research could contribute to the building and testing of social science theories. Yet we should draw back from the abyss and abandon radical skepticism of observational and experimental research. We should replace it with skepticism disciplined by alternative explanations. In the final section, "Observation, Experiments, and Theory," I discuss the relationship of observational and experimental research to the building and testing of theory.

2. Characterizing Radical Skepticism of Observational Research

Fueling the rise of experimental methods in social research has been growing awareness of the pitfalls of observational research. Awareness of these pitfalls is not new; it formed part of the backdrop to the invention of randomized experiments in the 1920s and 1930s.[2] What's new in the social science community is a spreading pervasiveness of radical skepticism about observational studies.[3] This skepticism involves the following logic. Consider a linear model of individual i's outcome on Υ:[4]

$$\Upsilon_i = \alpha_i + \beta_0 X_{i0} + \sum_{n=1}^{j} \beta_j \mathbf{X}_{ij} + \sum_{m=1}^{k} E_k \mathbf{Z}_{ik} + \mu_i$$

in which (dropping the i subscripts) X_0 is the key explanatory variable. The β_j coefficients relate $n = 1,2, \ldots, j$ X variables to Υ, but these factors are unrelated to X_0. Then $\mathbf{Z}$ represents $m=1,2, \ldots, k$ other factors

related to Υ and to X_0; these are confounders. μ represents unobserved causes of Υ.[5]

Among the $\mathbf{Z}_k$ confounders, some are not observed at the point at which the research is reported. But assume that some could be observed—if, for instance, the researcher's critics suggested them as confounders—whereas some are unobservable. The confounding variables can be thought of as a set composed of observed, currently unobserved but observable, and unobservable vectors:

$$\boldsymbol{Z} = \{\boldsymbol{Z}_{\text{observed}}, \boldsymbol{Z}_{\text{observable}}, \boldsymbol{Z}_{\text{unobservable}}\}.$$

The core belief of radical skepticism is that unobservable confounders always exist. $\mathbf{Z}_{unobservable}$ is never itself an empty set. No matter how diligent and inventive the observational researcher, she will never be able to overcome the bias imposed by the presence of unobserved—because unobservable—correlates of the key causal variable. As we shall see, there are two reasons why skeptics conclude that some confounders are likely always to remain unobserved. One has to do with the very large number of ways in which units can vary. This high dimensionality of units, the reasoning goes, means that it is basically impossible for the researcher to consider, much less control for, all confounders. The second reason is that some dimensions of variation are inherently difficult to measure. Given the inevitable existence of unobservable covariates, in the view of radical skeptics, observational researchers will rarely be able to identify, without bias, causal effects. Unobserved heterogeneity inescapably frustrates causal inference from observational data.

Regarding the first reason, one could make sense of the radical skeptic's belief that unobservable confounders always exist by noting the large numbers of ways in which human and social life varies. The social unit, whether a person, an institution, or a case of an event, generally varies on a very large number of dimensions. The dimensions of individuals include their income, schooling, attentiveness, physical characteristics . . . the list could be extended endlessly. So too for other social units. Were this dimensionality much smaller, the problem would appear less intractable. For instance, imagine a study involving people who vary only on three dimensions: income, handedness, and hair color. To test the hypothesis that handedness influences income, the researcher would have to assure

himself only that hair color and handedness are unrelated or that hair color exerts no influence on income, and that income has no reciprocal effect on handedness. He would not say to himself that he might be led astray by failing to control for some other confounder; by assumption, in this example, there are no other dimensions along which units vary. The degree of dimensionality of units increases the chances of omitted-variable bias. Because units in the social world tend to have high dimensionality, without a plausibility constraint observational research would indeed be basically incapable of detecting causal effects. Yet, I will argue later, this same high dimensionality would—again, in the absence of plausibility constraints—undermine experimental research as well.

The belief that observational research can never exhaustively introduce controls or make adjustments for all confounding factors tends simply to be asserted. This assertion is in sharp contrast to a more traditional (though embattled) approach. In this approach, the researcher begins with plausible alternative explanations—ones suggested by theory and by logic—that could vitiate his or her causal claim, devises measures of confounders implied by this alternative, and then examines the effect of the key explanatory variable in the presence of controls. A textbook description of such a procedure is offered by King, Keohane, and Verba. They give a hypothetical example in which the investigator seeks to estimate the effect of residential segregation in the Israeli-occupied West Bank on conflict between Israelis and Palestinians. Ideological extremism might be a confounding factor, leading people both to live in segregated communities and to be more prone to conflict. The solution is to control for ideological extremism.

We might correct for the problem here by also measuring the ideology of the residents explicitly and controlling for it. For example, we could learn how popular extremist political parties are among the Israelis and PLO affiliation is among the Palestinians. We could then control for the possibly confounding effects of ideology by comparing communities with the same level of ideological extremism but differing levels of residential segregation.[6]

From the standpoint of radical skeptics, controlling for one potential confounder is not a satisfactory fix. If unobservable covariates always lie just over the horizon, controlling for one or even several does not exhaust the problem of unobserved heterogeneity.

A sense of the intractability of omitted-variable bias in observational research comes through in the methodological reflections of many social scientists. In a thoughtful essay aimed at improving the quality of natural experiments, Dunning writes that "the strong possibility that unobserved differences across groups may account for differences in average outcomes is *always omnipresent* in observational settings."[7] Referring to Posner's study of ethnic relations in Malawi and Zambia, Przeworski writes, "While Posner provides persuasive arguments that members of each of the two groups do not differ otherwise than by being on different sides of the border, rival hypotheses entailing unobserved differences are *always plausible*."[8]

Gerber, Green, and their coauthors are impressed with the high dimensionality of human and social variation and infer from it that observational research tends to produce biased results. In the first chapter of this book, Gerber, Green, and Kaplan construct a Bayesian framework to compare the increments to knowledge provided by observational and experimental studies.[9] The authors concede that there is a risk of bias in both experimental and observational research but contend that it is typically much greater in observational research.[10]

The main source of bias on which they focus is omitted covariates. Whereas experimental researchers control the assignment of units to treatment and control, in observational studies "the data generation process by which the independent variables arise is *unknown* to the researcher."[11] This ignorance means that the observational researchers can never be confident that an unobserved factor has not shaped both their favored explanatory variable and the outcome. Gerber and coauthors' "Illusion of Observational Learning Theorem" rests on the fact that "if one is *entirely uncertain* about the biases of observational research, the accumulation of observational findings sheds no light on the causal parameter of interest."[12] Though this uncertainty is stated in the conditional tense—"*if* one is entirely uncertain"—the illustrations that Gerber and his coauthors offer represent it as irreducible.

With mediating confounding factors as with other confounders, these skeptics are impressed with their near-limitlessness. Green, Ha, and Bullock write that "as a practical matter, it is impossible to measure all of the possibly confounding mediating variables. Putting measurement aside,

it is rare that a researcher will be able to think of all of the confounding mediators."[13]

Like Gerber, Green, and their coauthors, Banerjee and Duflo (chapter 4) view omitted-variable bias as inherent in observational research. And like the political science skeptics, these economists also despair of the possibility of accumulation of knowledge from observational studies:[14] "If we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates. In contrast, there is no comparable statement that could be made about observational studies. . . . with observational studies, *one needs to assume non-confoundedness* . . . of all the studies to be able to compare them. If several observational studies give different results, one possible explanation is that one or several of them are biased . . . and another one is that the treatment effects are indeed different."[15]

Do observational researchers "know nothing" about the processes that generate independent variables and are they hence "entirely uncertain" about bias? Is the "strong possibility" of unobserved confounding factors "always omnipresent" in observational research? Are rival hypotheses "always plausible"? Can one do nothing more than "assume non-confoundedness"? To the extent that the answers to these questions are no, radical skepticism is undermined.

Let us consider the first claim, that observational researchers know nothing about the processes that generate their independent variables. The claim elides the undisputed fact that observational researchers do not *control* assignment of units to treatment and controls with the more questionable one that they *cannot understand* the process by which this selection takes place. If nonexperimental researchers can know nothing about the processes that generate their independent variables, they could not take advantage of natural experiments, in which a clearly exogenous event—for example, a natural disaster, a geographic feature, perhaps the drawing of a border—produces "as-if" randomization. Rather than thinking of observational researchers as necessarily in the dark about the processes producing their key explanatory variables, we should think of them as more or less constrained by the fact that they do not control this process. Observational researchers must learn all they can about the pro-

cess that selects units (individuals, groups, countries) into the kind that will be treated by the presumed cause and about the ways in which this process may also shape the outcome. They must always reassure themselves and their critics that they are dealing adequately with potential confounders and reverse causation.

Turning to the second claim of the omnipresence of unobserved confounding factors, a reason to doubt that they will always undermine observational research is that even though dimensionality of social units tends to be high, the number of plausible alternative explanations for any outcome of interest may not be so large. In fact, the number of plausible rival explanations may, in any given context, be relatively small. But the scope of the problem—are we dealing with myriad possible confounders or a handful?—remains unknown unless the researcher and his critics discuss specific alternative accounts and their plausibility. Instead of offering such specifics, the radical skeptic typically makes a blanket claim of the presence of unobserved covariates. The generality of the claim leaves the often-misleading impression that the number of plausible rival accounts must be very large. And, as Gerber et al. explain (in a paper which, as we have seen, is skeptical of observational research), the smaller the number of plausible rival explanations, the more confident one can be of causal inference in the absence of randomized tests.[16]

A belief in the omnipresence of unobserved confounders informs criticisms of the modeling of observational data. The Neyman model builds on the idea that a treatment effect is the difference in potential outcomes between a unit under treatment and that same unit under control. Though we observe individual units only in one state or the other, an experimental design seeks to ensure that treatment assignment is independent of all baseline variables. When assignment is not random, the researcher attempts to achieve conditional mean independence, which posits that potential outcomes should be, on average, identical between two groups of units under treatment or control, conditional on their having identical covariates. Observational studies, in which the assignment of units is not controlled by the researcher and is often not random, seek to balance units on their *observed* covariates. As Przeworski notes, "Having reached a satisfactory balance, [observational studies] then invoke [the] mean independence assumption, thus assuming either that balancing on the observed

covariates is sufficient to balance on the unobserved ones or that unobserved factors do not affect the outcome."[17] Whether the estimation technique is regression or nonparametric matching, the radical skeptic will be unsatisfied since there is no matching on unobserved covariates.

The problem may be overstated, for reasons similar to the ones laid out earlier. To the extent that currently unobserved potential confounders can be shifted into the category of the observed, the problem is mitigated. We cannot know whether unobservable confounders vitiate causal inference in any particular case unless we explore plausible rival explanations and seek out additional information and measures that will help us to evaluate them. What's more, this perspective implies a research process in which one gathers all information that is readily available about units, the "low fruit" among covariates, matches on them, and declares the task complete. A more effective approach is to begin with conjectures about the causes of an explanandum—conjectures informed both by empirical observation and by deductive reasoning—observe patterns relating key causes to the outcome of interest, interrogate oneself and be interrogated by others about possible confounders or reverse causation, and seek out information that would allow matching on these covariates, shifting them from the category of unobserved to observed.

There are plenty of ways in which observational researchers can go astray, and the challenges are not easy. Yet they appear more insurmountable in the abstract than they often prove to be in concrete cases. Observational researchers who self-consciously lay out potential alternative explanations often find the number of plausible ones to be small and the confounders that they imply to be observable. The process by which they present findings and open themselves to alternative explanations is, as Rosenbaum explains, a crucial part of the research process.[18] They must take these steps because their interlocutors will suggest alternative explanations, explanations which in turn suggest confounders that must be taken into account. The researcher will either find that her original hypothesis survives analyses that take into consideration potential confounders or it does not. Whether plausible rival explanations exist that would require adjustments for truly unobservable confounders, as opposed to observable ones which the community of researchers has simply not yet measured, is an open question. In any case, the ratio of unobserv-

able to currently unobserved confounders may be smaller than the radical skeptic supposes.

Rather than indicating specific alternative explanations, thus setting off the effort to shift confounders from the category of unobserved to observed, radical skeptics are prone simply to assert that, in the absence of random assignment of units to treatment and controls, unobserved covariates must be making mischief. Radical skepticism replaces a discussion of specific potential confounders and the alternative explanations to which they are attached with blanket complaints about the absence of an identification strategy. This failure to posit specific rival explanations leaves the impression that myriad alternatives must exist. The lack of specificity often masks the fact that the number of plausible alternatives is tractably small. Radical skepticism thus remains ungrounded, in Wittgenstein's sense.[19] Not the failure to test rival explanations by observing potential confounders, but the failure to randomize assignment to treatment and controls, is what—in the radical skeptic's view—vitiates observational studies.

Przeworski's sense of the intractability of omitted-variable bias stems from the second reason mentioned earlier: that some confounders simply cannot be observed. To illustrate the point, he posits, hypothetically, that democracy promotes economic growth whereas dictatorship slows it. Yet, he claims, any observational study of national growth rates under democracy and dictatorship would be frustrated by an omitted, unmeasurable confounder: the quality of political leadership. But his example of an unobservable confounder seems more illustrative of the wisdom of exploring plausible alternatives than of the inevitable unobservability of confounders:

> Suppose that leaders of some countries go to study in Cambridge, where they absorb the ideals of democracy and learn how to promote growth. Leaders of other countries, however, go to the School of the Americas, where they learn how to repress and nothing about economics. Dictatorships will then generate lower growth because of the quality of leadership, which is "Not Available" [i.e., not measured in this hypothetical exercise and presumably unknowable]. . . . Since this is a variable we *could not observe systematically*, we cannot match on it.[20]

There may be ways of conceptualizing "quality of leadership" that leave it unobservable. But coding the postsecondary educational careers of third world leaders sounds like a laborious task, not an impossible one.

Przeworski laments the "subjectivity" of plausibility assessments of rival hypotheses, seeming to wish for an objective test, a kind of t-test for plausibility. And the language he chooses is heavy with a sense of improvisation and looseness in the absence of random assignment. Evaluating the quality of instrumental variables necessitates "conjuring and dismissing stories" about their effects on outcomes; justifying them entails "rhetoric: one has to tell a story"; the amount of information that can be squeezed out of historical data is "a matter of luck."[21]

The title of his essay is "Is the Science of Comparative Politics Possible?" In light of the difficulty of applying experimental techniques to such questions as do democracies grow economically more rapidly than dictatorships? or do independent central banks promote growth?, his answer is no. Comparative politics is a science if (all) this means is "following justifiable procedures when making inferences and examining evidence" and "agreeing to disagree." We are capable of generating "reproducible results, arrived at through reasonable procedures." But "to identify causal effects, we must rely on some assumptions that are untestable."[22] Here again is the key to his frustration with the limits of comparative politics as a science: the absence of tests for the assumptions we must make.

Yet, as we shall see, to produce meaningful results, experimentalists make assumptions, and some are not testable. We should be wary, furthermore, of a cartoon character of the natural or physical scientist whose work is free of improvisation, intuition, interpretation, and reliance on procedures that are reasonable rather than testable.

None of the foregoing is to say that the problems of unobservable covariates and potential reverse causation are not real—some important covariates simply cannot be observed, proxies are problematic, and good instruments are elusive. But experimental researchers face equivalent sorts of challenges.

3. Radical Skepticism and Experimental Research

Experiments allow us to test the null hypothesis that the average effect of a presumed cause is zero and to estimate the average size of the

effect. Random assignment of subjects to treatment and control, with a sufficiently large sample, ensures balance on observable and unobservable covariates, avoiding the problem of omitted-variable bias.[23] Freedman notes that the key parameter of interest is the difference between the average response if all subjects were assigned to treatment and the average response if all subjects were assigned to controls. An unbiased estimator of this difference is the difference between the average response of all subjects assigned to treatment and that of those assigned to the control.[24]

Experimental researchers are well aware of a number of problems that can afflict their work. My focus here is on the phenomenon of subsets of experimental subjects' responding differently to a treatment, known as an *interaction* (i.e., the treatment interacts with traits of subjects) or as a *heterogeneous treatment effect.*

The history of medical research is littered with examples of interactions. Consider recent research into the effectiveness of cholesterol-lowering drugs on heart attacks. It was well established that low-density lipids (LDL) increased the risk of coronary events and that statin therapy lowered both LDL levels and the risk of these events. Additional research suggested health benefits from statin therapy even among subjects with LDL levels considered normal. But researchers suspected an interaction: that statin therapy improved health among people whose LDL level was normal but whose level of c-reactive protein (CRP), a marker for inflammation, was elevated, while having little beneficial effect among those with normal levels of both.[25] The public health implications were important: not all people with normal LDL levels, only those with elevated CRP, would benefit from statin therapy. It is not hard to find less innocuous examples of drugs whose beneficial average effect masks small benefits for a majority of the sample and highly deleterious ones for a subpopulation.[26]

Interactions do not threaten the step from experimentally uncovered average treatment effects to causal inference. If an experimental study is large, well designed, and well implemented, random assignment of units to treatment and control allows one to infer that the treatment is the cause of any observed average difference in outcomes between treatment and control groups.[27] Experimental design ensures that no unobserved covariate is the real cause in differences in outcomes and that apparent differences are not the result of reverse causation. These are no mean

feats. The problem posed by interactions is that they can change the meaning of experimental results in a broader sense.

For example, consider Wantchekon's field experiment on clientelism in Benin.[28] With the cooperation of four major political parties in the run-up to a national election in 2001, he studied the impact of alternative campaign strategies in eight of Benin's eighty-four electoral districts. In each of the eight districts he selected one noncompetitive village to receive a "clientelist" treatment and one a "public policy" treatment; the remaining villages were controls.[29] In the clientelist treatment, campaign workers promised local public goods or trade protections for local producers, should their party be elected. In the public policy treatment, campaign workers made promises that were national in scope: alleviating poverty, advancing national unity, and eradicating corruption, among others.[30] Wantchekon then compared aggregate voting patterns in the following election across treatment and control villages and conducted a postelection survey in which respondents were asked how they voted.

Wantchekon finds that the average effect of the clientelist treatment was to increase electoral support for the party associated with this message. The average effect of the public policy message was to reduce support. In the villages of one district, however, there was no positive effect of the clientelism treatment, and in the villages of two districts there was no negative effect of the public policy treatment. It is not clear from Wantchekon's presentation whether the reported average treatment effects are across the full sample.

The results indicate many and complex interactions. Viz: "there is a significant and negative public policy treatment effect for northern candidates, regional candidates, and incumbent candidates. By contrast, there is a positive treatment effect for southern candidates. A direct comparison of the treatment effects—that is, of clientelism versus control . . . reveals that clientelism is more effective for northern candidates."[31] Wantchekon also uncovers interactions between gender and the treatments, with women on average responding more favorably than men to the public policy treatment and men more favorably than women to the clientelism treatment (though here again he cites a somewhat dizzying set of caveats related to the region and incumbency status of the candidate).

The author makes a vigorous effort to explain these interactions, perhaps less so to use them to evaluate theories of clientelism. As mentioned,

from his presentation it is not entirely clear whether the average treatment effect holds even without controls for interacting variables or whether positive (clientelism) or negative (public policy) average treatment effects emerge only when one disregards the regions in which these effects were missing. Assuming the latter is the case, we would have no average treatment effects but potentially theoretically relevant interactive effects of treatment with region, incumbency status of candidates, and gender. Should the big news of this study be "Clientelism has no significant effect on voting behavior?" or should it be (for example) "Men are more susceptible to clientelist appeals than women?"

Note, furthermore, that units have not been randomly assigned to values of the interacting factors. Individuals whom Wantchekon surveyed were not randomly assigned to gender, region, or—for candidates—incumbency status; nor could they be. Without follow-up studies, a version of the unobserved heterogeneity problem creeps back in. For instance, is it women who are more susceptible to public policy appeals? or people engaged in interregional trade, who hence have wider exposure to national problems? (Most such people, Wantchekon notes, are women.)

In light of suspected interactions, researchers can undertake a number of research design fixes and statistical adjustments. With a sufficiently large sample, they can calculate the difference between treatment and control groups within the relevant subsample. But with small samples and multiple interactions, the number of units will be used up quickly. Other strategies are statistical, such as regressing the outcome variable on the assignment variable (a dummy registering assignment to treatment or control groups), a control for the trait in question, and an interaction between these two main effects.[32] Research design can be crafted with an eye toward suspected interactions; for instance, researchers can randomize within strata of observable factors that are suspected to interact with the treatment. Hence medical researchers who suspect that cholesterol-lowering drugs have a differential impact on people with high and low levels of c-reactive protein can conduct a new study, this time of people with high levels of CRP, randomly assigning them to treatment and control groups.[33] Wantchekon or others could undertake a follow-up study exclusively of women, assigning them randomly to clientelism and public policy treatments and to a control group; presumably long-distance traders and nontraders would be balanced among the groups.

But from the standpoint of the radical skeptic, no research design can dispose of all potential interactions. Setting plausibility aside, if units have high dimensionality and if some confounders are unmeasurable, some unobserved trait is always likely to interact with the treatment. Faced with an experimental study that uncovers a causal effect, the radical skeptic should posit some unspecified subset of units whose response to treatment is at odds with the average response, potentially changing the theoretical implications of the study's findings. If interactions can change the interpretation of experimental results, then the radical skeptic should be unnerved by their implication for experimental research. Because one can test only for interactions between treatments and *observed* factors, ungrounded skepticism implies that we will remain in the dark regarding the real findings of experimental studies.

Unobserved interactions play—or ought to play—the same role in the radical skeptic's view of experimental results as do unobserved covariates in her view of observational research: both are omnipresent and inevitably limit the contribution of research to knowledge. We may be able to rule out (or in) the possibility that not just people with high CRP levels but other subpopulations who differ on some other dimension, specified or not, will be helped by drugs, or that not just women and southerners but some other subpopulation is resistant to clientelism. But if some key traits (like some covariates) will always remain unobservable, then additional experiments and statistical adjustments cannot fix the problem.

To see the parallels between the problem of unobserved confounders and unobserved heterogeneous treatment effects, consider a model of an experimental subject's response on outcome Υ:

$$\Upsilon_i = \alpha + \beta_0 T_i + \sum_{n=1}^{j} \gamma_j \mathbf{Z}_{ij} + \sum_{n=1}^{j} \theta_j \mathbf{Z}_{ij} * T_i + \mu_i,$$

β_0 relates an assignment variable, T_i, indicating treatment status, to the expected outcome for individual *i*. The $\mathbf{Z}$ matrix represents $n = 1, 2, \ldots, j$ traits particular to each observational unit (gender, age, race, income, regime type, what have you). γ_j coefficients relate these traits to the dependent variable (the main effect); with random assignment and large samples, we expect these traits to be balanced, whether or not the researcher observes them. $\mathbf{Z}_{ij}$ picks out the ith row of the $\mathbf{Z}$ matrix, that is, the set

of j characteristics for person i. θ_j coefficients relate the effects of treatment, conditional on these traits: the effect, for instance, of a drug on mortality among people with higher or lower levels of a blood protein or of the region a person lives in on their susceptibility to electoral appeals.

As with confounding covariates, the **Z** traits can be conceived as a set composed of observed, unobserved but observable, and unobservable elements:

$$\boldsymbol{Z} = \{\boldsymbol{Z}_{\text{observed}}, \boldsymbol{Z}_{\text{unobserved}}, \boldsymbol{Z}_{\text{unobservable}}\}.$$

The radical skeptic should believe that the set of unobservable interacting factors is never empty. Her keen sense of human and social variability and hence of the high dimensionality of units, as well as of the unmeasurability of some of these key dimensions, should lead her to believe that unobservable interactions always threaten the meaningfulness of causal inference based on experimental data.

The case of statin therapy illustrates the difficulties experimental researchers would face if radical skepticism were warranted. First, note that detailed biochemical knowledge, not prior double-blind testing, led researchers to suspect that CRP played an interactive role between statin therapy and health outcomes. Second, the interacting trait (elevated levels of CRP) was readily observable: experimental subjects could be given a simple test to detect its level. Had either ingredient been missing—had researchers been unaware of a likely interaction or had they been aware of it but unable to measure levels of CRP—they would not have been able to assess the difference between the statin treatment's effect on the average recipient and on those with elevated CRP. As a consequence, they would have been led far astray in their assessment of the health benefits of statin therapy on people with normal cholesterol levels.

Another example, this one also from a study of voter mobilization, illustrates how radical skepticism challenges the meaningfulness of experimental research. To gauge the impact of norms of civic duty and social pressure on electoral participation, Gerber, Green, and Larimer conducted an ingenious field experiment on a sample of 180,000 people registered to vote in Michigan in 2006.[34] They mailed letters containing messages intended to elicit shame and, they conjectured, to induce higher rates of turnout to four treatment groups. The outcome variable

was turnout in the primary election in August of that year. The difference in turnout rates between the control group, who received no mailing, and the treatment group that received the strongest shaming cue was 8.1 percentage points—37.8 percent for the treatment versus 29.7 percent for the control.[35]

The authors tested for two kinds of interactions. One was between a person's internalized sense of civic duty and the social shaming treatment. Their measure of civic duty was the individual's voting propensity. Their likelihood-ratio test for the interactions failed to reject the null hypothesis of equal treatment effects among high- and lower-propensity voters.[36] They also tested for interactions between treatments and the probability of voting Democratic, which they estimated with demographic information. They found no evidence of interaction. Because the authors had excluded most Democratic voters from the experimental sample, this test is less persuasive.[37] Hence the study's findings can be summarized in this way: among some categories of registered voters in Michigan—and probably more broadly—shame increases turnout. Shame has the same effect across registered Michigan voters with varying levels of civic duty and (perhaps) partisan orientations.

There are other interactions that a student of political mobilization might reasonably want to consider, particularly in light of the implications that Gerber, Green, and Larimer draw from their experimental results. Perhaps not people (or Michigan residents) in general but those already involved in organizations are primed to suffer such shame. A follow-up study that stratified by level of organizational participation might lead to a fairly substantial revision of the study's results, viz: in Michigan and probably more broadly social shame encourages turnout among organizationally active citizens but not among nonparticipants.

More generally, one would like to know not just the average effect of treatment but its heterogeneous effects on subpopulations because these parameters can change the interpretation of why a treatment has the effect it does, with implications for theory. If people from all regions of Benin are susceptible to clientelist mobilization, the explanation might be that clientelism boosts consumption, the marginal utility of consumption diminishes as income rises, and Benin is a poor country.[38] Yet if mainly men were persuaded by such offers, perhaps not poverty but particular

labor market experiences and, behind them, a parochial versus national perspective drive voters' responses to clientelist electoral appeals. If U.S. voters in general are susceptible to shaming, the implication is that very basic social sensibilities to shame induce people to undertake costly behavior such as voting. If only people already involved in organizations are susceptible, the implication might be that a susceptibility to shame needs to be elicited by experiences outside of the political sphere, as in, say, religious communities or unions, before it is available for political activists to take advantage of. In the latter settings, contacts tend to be more intimate and ongoing than in party politics in advanced democracies. Heterogeneous treatment effects imply quite different theories of electoral participation.

On the radical skeptic's view, one could not hope to test all such hypotheses. The problem could not be fixed by piling experiment upon experiment, as in Banerjee and Duflo's contention that "if we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any set of covariates."[39] Green, Ha, and Bullock's less sanguine view is more consistent with the implications of radical skepticism. Regarding the problem posed by interactions, they write, "The bottom line is that when subjects are governed by different causal laws, analyses that presuppose that the same parameters apply to all observations may yield biased results. . . . a single experiment is unlikely to settle the question of heterogeneous treatment effects. In order to ascertain whether different subjects transmit the causal influence of X in different ways, multiple experiments—maybe decades' worth—will be necessary."[40] Yet if, like unobserved confounders, unobserved interactive traits always lie just over the horizon, in fact we could never learn as much as we want to know from more experiments. The truly consistent radical skeptics' position would be that the diversity of human and social units makes the observation of all interactions impossible; hence there can be no stopping rule, no point at which one has tested for all possible interactions.

Of course, observational researchers cannot be complacent in the presence of unobserved interactive effects. Imagine that we were to gather data on vote shares in response to clientelist appeals, without random

assignment to treatment and control. Imagine, furthermore, that measures were readily available of all plausible covariates, and that an effect of higher vote shares survived the introduction of these controls. If, nevertheless, we suspected an interactive effect of some other factor, and if this factor were unmeasurable, we could not be confident of our causal inferences about the effect of the campaign strategy.

I have made the case that radical skepticism implies equally devastating consequences for observational and experimental research. Yet the implications are not identical. Assume for the sake of argument that the radical skeptics are right. What are the consequences? The observational researcher could never be certain that a putative causal effect in fact operates. We could never be sure that A causes B, rather than C causing both A and B; there will always be some unobserved C's out there that may actually be explaining the apparent causal effect. Or, in the presence of ungrounded doubt, we can never be certain that A causes B rather than B causing A. For experimental research, confidence remains high that, on average, A causes B; random assignment, given a sufficient sample size, ensures balance on all covariates, the observed, the unobserved, and the unobservable.[41] And well-designed experiments also dispose of the problem of reverse causation. Yet, if the radical skeptic is both right and consistent in her criticisms of observational and experimental research, we must always doubt whether the experimentally elicited response of some subgroups is very different from that of others. If so, the average treatment effect may mislead us as to the study's meaning and theoretical implications.

Radical skeptics might reject the equating of unobserved covariates and unobserved interactions on the following grounds. At least the finding of a significant average treatment effect tells us that a causal effect is present. Follow-up studies can always be conducted to refine our understanding of any interactive effects, whereas one cannot be sure that an observational study has identified any real effect.

But if we really believe that some unknown and unobservable factor could be interacting with the treatment, we should always be uncertain about the meaning of experimental results. The Benin example points to the possibility of false negatives: experimental evidence against any effect when in fact there are significant and theoretically meaningful effects on subpopulations. Or, returning to the study of shaming and turnout,

imagine that in reality shaming increased the probability of voting only among people who are already involved in organizations, whereas those not involved were not affected by it. And imagine that researchers carried out an experiment like the shaming study but did not entertain the possibility that organizational involvement was a required condition for shaming to boost turnout. The average effect of the treatment could be entirely due to its very large impact on the organizationally active. The researchers would draw the wrong conclusion from the study: that shame has a universal effect on people, whereas in fact it has an effect only on organizational participants. The theoretical implications are quite different.

In sum, even when experiments turn up significant treatment effects, if we really believe that some crucial interacting factor remains unobserved, then our explanation of this effect is unlikely to be accurate or complete. A causal effect that cannot be explained cannot be identified, in any meaningful sense of that term.

Yet radical skepticism of experimental research is unwarranted, just as it is unwarranted of observational research; and for similar reasons. The range of plausible interactions is not infinite but finite. Blanket challenges such as "there is undoubtedly, among the population sampled, some heterogeneity that makes the average treatment effect irrelevant" are unpersuasive. More persuasive are concrete challenges of the form, "Is it not likely that subgroup *X* responds in a distinctive way to this intervention?" especially when knowledge of subgroup *X* and its reactions motivates the question. To be more concrete, returning to the turnout study, a blanket claim of unobserved interactions should give way to conjectures such as "It may be that organizationally active people react to shaming cues by turning out more, but the uninvolved won't be responsive or might even turn out less." In sum: challenges to both observational and experimental research need to be disciplined by specific and reasonable alternatives accounts.

Are heterogeneous treatment effects a potential challenge to internal validity or to external validity? The answer is, both. We saw in the voter mobilization studies that uncertainty about the effects of treatments on subsamples opened up a great deal of ambiguity about the fundamental findings that particular experiments had unearthed and potentially undercut their role in constructing and testing theory. These are problems of

internal validity. But interactions are probably also a common obstacle to external validity. Consider a hypothetical randomized experiment that reveals an average treatment effect of β_t, but—unbeknownst to the researcher—the treatment interacts with variable X_j in such a way that, for a small subpopulation with high values on X_j, the average (subgroup) treatment effect is $\beta_j = -\beta_t$. If the identical experiment were repeated in another location in which a majority of the population had high values on X_j, we would expect the average treatment effect to be very different from that of the original study.

These are the kinds of problems that lead Rodrik, Deaton, and other development economists to doubt that individual studies close the case on the effectiveness of policy interventions. Rodrik's example is research into the effectiveness of various methods for distributing insecticide-treated bed nets, which are helpful in preventing malaria infection. The debate is whether free distribution or charging a nominal price for the nets is more effective in getting people to use them. A single study that strongly favored free distribution, carried out in western Kenya, would not necessarily generalize to other African settings because the treatment might have interacted with factors specific to this region, such as a lot of prior social marketing of the nets.[42] Interactions are not the only hindrance to generalizability or external validity, but they are an important one.

4. Observation, Experiments, and Theory

A central role of social science is the building and testing of broad theories of social phenomena. A powerful role that experiments can play is to test whether basic causal claims that theories rely on or imply in fact can be sustained, especially when there are good reasons to think that observational research is hobbled by endogeneity problems (good specific reasons, not ungrounded skepticism).

Skeptics are sometimes tempted to set aside not just all prior (observationally based) causal claims but all received theories. They argue for a tabula rasa. Their reasoning is that observational research produces unreliable causal claims, and these causal claims are the foundation of much theory. Gerber, Green, and their coauthors sometimes advocate for this tabula rasa stance, such as when they write,

> Lest one wonder why humans were able to make so many useful discoveries prior to the advent of randomized experimentation, it should be noted that physical experimentation requires no explicit control group when the range of alternative explanations is so small. Those who strike flint and steel together to make fire may reasonably reject the null hypothesis of spontaneous combustion. When estimating the effects of flint and steel from a sequence of events culminating in fire, σ^2_B [the uncertainty associated with the degree of bias] is fairly small; but in those instances where the causal inference problem is more uncertain because the range of competing explanations is larger, this observational approach breaks down.[43]

Green, Ha, and Bullock are more strident, arguing in favor of "black box experimentation," by which they mean experimental studies that do not attempt to explain why a cause has the effect it does:

> Experimenters have good reason to be cautious when encouraged to divert attention and resources to the investigation of causal mechanisms. First, black box experimentation as it currently stands has a lot going for it. One can learn a great deal of theoretical and practical value simply by manipulating variables and gauging their effects on outcomes, regardless of the causal pathways by which these effects are transmitted. Introducing limes into the diet of seafarers was an enormous breakthrough even if no one at the time had the vaguest understanding of vitamins or cell biology. Social science would be far more advanced than it is today if researchers had a wealth of experimental evidence showing the efficacy of various educational, political, or economic interventions—even if uncertainty remained about why these interventions work.[44]

The idea that theories, in social or natural sciences, are built inductively from the accumulation of shreds of cause-and-effect relationships is a misconception. Theory building always involves a great deal of deduction as well as inductive testing. The late-eighteenth-century innovation of feeding foods rich in ascorbic acid to sailors was an "enormous

breakthrough" in public health but not, on its own, in biological science. The discovery—supported, in fact, with experimental evidence—of dietary means to fight scurvy, along with other like innovations, certainly contributed to biological theory but so did deductive reasoning and observational investigations.

Indeed, in contemporary medical practice the vast majority of accepted treatments and procedures have never been subjected to double-blind clinical trials. Among them are appendectomies for appendicitis, cholecystectomy for gallstones, penicillin for tuberculosis, and diuretics for heart attack patients.[45] Doctors use these treatments even though they have never been and are unlikely ever to be tested experimentally. In this connection, critics of the evidence-based medicine movement argue for a restored sense of the value of clinical experience and biochemical research as sources of objective evidence on which medical practice is based. Perhaps, one might counter, doctors persist in these untested practices even though the evidence in their favor is indeed weak only because it would be ethically unacceptable to expose appendicitis patients, for example, to the risk of foregoing an appendectomy. Yet, as Worrall points out, if we really believe that the evidence is weak, we would not be convinced that the risk is great.[46] Though one can cite many examples of treatments thought to be therapeutic and turning out not to be, in the history of medicine most have been abandoned on the grounds of nonexperimental evidence—think bloodletting.

Social scientists should resist the temptation to cast all nonexperimental research as flawed, to reject all prior theorizing as based on flawed evidence, and to conceive of theory building as the incremental accretion of shreds of knowledge about cause-and-effect relations, narrowly construed. We need rich and variegated evidence, rigorously developed and analyzed, and considered in light of theories—which, as in all fields of science, are in part deductive in nature—if we are to gain knowledge about the workings of the social world.

Acknowledgments

I am grateful to Chris Achen, Thad Dunning, Don Green, William Hennessey, Gary King, Matt Kocher, Noam Lupu, Steve Pincus, Frances Rosenbluth, Ian Shapiro, and Dawn Teele for their comments.

Notes

1. For experiments on voters, see Iyengar and Kinder (1987) and Ansolabehere and Iyengar (1995); and on voting behavior Wantchekon (2003). For attitudes about race, see Nelson, Sanbonmatsu, and McClerking (2007) and Valentino, Hutchings, and White (2002). On distribution rules on collective action, Dawes et al. (1986). On corruption and the rule of law, see Fried et al. (2010). For recent reviews, see Druckman et al. (2006) and Humphreys and Weinstein (2009). For an account of early political science field experiments, see Green and Gerber (2003b). Brady (2009) and Brady, Collier, and Box-Steffensmeier (2009) review experiments in the broader context of political science methodologies. For experiments in labor economics, see Sackett et al. (2000). Experiments on microloans, Karlan (2005). On deworming on school attendance, Miguel and Kremer (2004); and on bednets and malaria, see Cohen and Dupas (2007). For recent reviews of experimental research in development economics, see Imbens and Wooldridge (2009) and Banerjee and Duflo (2009). For recent critical assessments of experimentation in development economics, see Deaton (chapter 6 in this book) and Rodrik (2009). See Morton and Williams's (2010) discussion of disciplinary shifts to experimentalism.

2. See Fisher (1925). Rosenbaum (2002) and Freedman (2006) provide historical background to this development.

3. By no means are all experimentally oriented social scientists radical skeptics. Morton and Williams (2010), for instance, are much more circumspect in their criticisms of observational research.

4. In this illustration the effects are assumed to be constant, but the same point holds with nonlinear effects.

5. Even if all factors systematically related to Y were measured, we might expect unexplained variation in Y, due to fundamentally unexplainable variability.

6. King, Keohane, and Verba (1994: 95).

7. Dunning (2008: 289), emphasis added.

8. Przeworski (2007: 287). His reference is to Posner 2004. Referring to an earlier draft of Dunning 2008, Przeworski adds that rival hypotheses "may be more or less plausible but, as Dunning [2005] emphasizes, assessments of their plausibility are inevitably subjective" (2007: 153–54). In fact, Dunning seems less troubled than Przeworski about the "subjectivity" entailed in posing and evaluating alternative explanations. About Posner's same study, Dunning writes that his "investigation of the plausibility of the relevant counterfactuals provides an example of 'shoe leather' research (that is, walking from house to house to find nuggets of evidence and rule out alternative explanations)" (2007: 287).

9. Gerber, Green, and Kaplan (chapter 1 in this book, p. 12).

10. Ibid.

11. Ibid., p. 10, emphasis added.

12. Ibid., p. 15, emphasis added.

13. Green, Ha, and Bullock (2010: 203). The context of this discussion is mediating variables, ones that transmit the causal effects of other variables. Confounding

mediating variables covary with the variable which the researcher believes is the key mediator.

14. Banerjee and Duflo (chapter 4 in this book, p. 94)

15. Ibid.

16. Gerber, Green, and Kaplan (chapter 1 in this book, p. 28).

17. Przeworski (2007: 153).

18. Rosenbaum (2002).

19. Wittgenstein (1969).

20. Przeworski (2007: 161), emphasis added.

21. See ibid., 162, 163, 167.

22. Ibid., 169.

23. It is relatively straightforward to show that observed covariates are balanced across treatment and control groups, but some philosophers of science question whether randomization ensures balance on unobserved covariates. See Urbach (1985).

24. Freedman 2006. Cook and Campbell recommend that researchers report their results in a cautious manner, such as that, in this test, a given effect was found to be of magnitude X and "in tests similar to the one conducted, the effect in 95 percent of the cases would be an increase" of this same magnitude. But they lament the general lack of circumspection. Cook and Campbell (1979: 41).

25. And further clinical trials showed that people with elevated CRP levels benefited from statin therapy, independent of the therapy's effect on LDL levels; see Ridker et al. (2005).

26. Worrall (2007b: 995) discusses clinical trials of benoxaprofen, developed to treat arthritis, in which the subjects included only eighteen to sixty-five year olds. The drug was approved for the market but later was associated with kidney failure among elderly patients, leading to a number of deaths. In this case, the treatment seemed to interact with a factor (age) in a nonlinear fashion, and the experimental sample excluded subjects with values of this factor over which the interaction occurred. In this case the average treatment effect was probably biased by the exclusion of part of the relevant population, at least if the outcome variable included overall quality of health rather than, narrowly, a relief of pain from arthritis.

27. Large sample size is important, Urbach (1985) reminds us, because small samples may produce imbalance in unobserved covariates, simply by chance.

28. Wantchekon (2003).

29. It is not clear from Wantchekon's report whether the sets of villages were randomly assigned to treatments or control groups.

30. There is a tension between the construct of clientelism used by most theorists and Wantchekon's treatment. The former stresses parties' distribution of minor material goods to voters, whereas Wantchekon's treatment entails promises of local public goods.

31. Wantchekon (2003: 413–14).

32. Some are queasy about estimating multivariate regression models with experimental data. Freedman explains that in a regression model of the outcome variable,

the error term is not a random variable but is fixed by assignment to treatment or control; the only random variable in the model is the assignment variable. Hence "the assignment variable . . . and the error term in the model will generally be strongly related" (2008: 2). For a response, see Green (2009). Furthermore, the inclusion into successive regression specifications of controls for covariates (subgroup identifiers) and interaction terms raises questions of data mining.

33. Ridker et al. (2008).

34. Gerber, Green, and Larimer (2008).

35. The strongest, "neighbors" treatment was a letter reminding them of their own voting history and reporting that of their neighbors.

36. Gerber, Green, and Larimer (2008: 39).

37. Only 2.7 percent of their sample was composed of Democrats. The remaining Democrats, what's more, were ones who lived in households with likely Republicans. The reason for the exclusion was that there were no competitive Democratic races; therefore Democratic voters would be unlikely to open campaign mailings.

38. See, e.g., Dixit and Londregan (1996).

39. Banerjee and Duflo (chapter 4 in this book).

40. Green, Ha, and Bullock (2010: 206.)

41. However, random assignment does not ensure the identification of an average treatment effect in the experimental population that is an unbiased estimate of the average effect in populations not included in the experiment. For this we need random sampling from the broader population of interest into the experimental subject population.

42. Rodrik (2009). The study he discusses is by Cohen and Dupas (2007)

43. Gerber, Green, and Kaplan (chapter 1 in this book, p. 27).

44. Green, Ha, and Bullock (2010: 207–8)

45. These and other examples are offered by Worrall (2007b: 986).

46. Ibid., 986–87.