Causal Empiricism in Quantitative Research

Cyrus Samii, New York University

Quantitative analysis of causal effects in political science has trended toward the adoption of "causal empiricist" approaches. Such approaches place heavy emphasis on causal identification through experimental and natural experimental designs and on characterizing the specific subpopulations for which effects are identified. This trend is eroding the position of traditional regression studies as the prevailing convention for quantitative causal research in political science. This essay clarifies what is at stake. I provide a causal empiricist critique of conventional regression studies, a statement of core pillars of causal empiricism, and a discussion of how causal empiricism and theory interact. I propose that the trend toward causal empiricism should be welcomed by a broad array of political scientists. The trend fits into a broader push to reimagine our discipline in terms of collective research programs with high standards for evidence and a research division of labor.

This essay discusses the move toward "causal empiricism" in quantitative political science. Causal empiricism refers to an approach to quantitative research that emphasizes nonparametric causal identification as well as design-based inference methods (Dunning 2012; Freedman 1991; Rosenbaum 1999; Rubin 2008). Causal empiricist research designs leverage the identifying power of experiments or natural experiments to establish specific causal facts for well-defined subpopulations (Imbens 2010). In empirical economics, the advance of causal empiricism is said to have ushered in a "credibility revolution" (Angrist and Pischke 2010).

Causal empiricism is often understood in terms of deep consideration—some might say an obsession—with "causal identification" and clear definition of counterfactual comparisons (Angrist and Pischke 2009, chaps. 1–2; Imbens and Rubin 2015, chap. 1; Morgan and Winship 2015, chap. 2). Identification refers generally to establishing that conditions sufficient for drawing an unbiased conclusion from data hold (Manski 1995), and causal identification applies this notion to causal effects. Such "identifying conditions" include random assignment, conditional random assignment, or discontinuous assignment of "treatment" variables—that is, variables for which we are interested in estimating causal effects. Causal empiricism is associated with "identification strategy" research designs. An identification strategy is the "combination of a clearly labeled source of identifying variation in a causal variable and the use of a particular econometric [or other statistical] technique to exploit this information" (Angrist and Krueger 1999, 1282). Such techniques include simple comparisons with randomized experiments, instrumental variables estimation with a valid instrument, regression discontinuity estimation with a valid discontinuity, and conditioning strategies like regression and matching under conditional independence assumptions.

Moves toward causal empiricism are evident in the increased attention that researchers put on identification strategies in applied research. The move toward causal empiricism is a clear departure from the prevailing convention in political science. The prevailing convention, consolidated during the 1990s, is what we might call mass production of quantitative "pseudo-general pseudo-facts" through multiple regression analyses.[1] In quantitative research, pseudo-facts are statistical results that are interpreted erroneously in terms of their causal implications, and pseudo-general findings are ones that are erroneously described as applying to a more general class of units than is immediately warranted.[2] This characterization

Cyrus Samii (cds2083@nyu.edu) is assistant professor, Politics Department, New York University, 19 West 14th Street, New York, NY 10012.

1. Econometrician Angus Deaton is even more colorful in characterizing such work in economics in terms of "the magic regression machine" (Ogden 2015).

2. A pseudo-fact is not the same thing as a "stylized fact"—rather, a pseudo-fact could be understood as a statistical finding presented as a stylized fact on the basis of an erroneous interpretation. For example, in an example below I will propose that the claim of a statistical nonrelationship between ethnic fractionalization and civil war onset is a pseudo-fact and therefore should not enjoy the status of a stylized fact.

The Journal of Politics, volume 78, number 3. Published online May 17, 2016. http://dx.doi.org/10.1086/686690

of prevailing conventions is harsh, but I will argue below that it is justified. At the turn of the millennium, the modal quantitative research design was one in which researchers assembled data on theoretically interesting dependent and independent variables for the "universe" of cases of interest. Researchers then assessed the presumably causal relationships in these data using regressions with informally motivated sets of control variables to reduce the potential for confounding.[3] The question of whether one was supposed to "believe" in the regression specification was rarely addressed, and so it is rarely clear whether the regression model, as an object, should be construed as a structural model of outcomes or as an agnostic tool for achieving causal identification.

This convention in quantitative causal research appears to be breaking down, and more quantitative causal research is moving toward causal empiricism. This does not represent a major change in general goals—researchers have always been interested in causal inference. Rather, it represents a major change in what researchers believe are credible ways of doing causal inference and immediately justifiable ways of describing their results.

Take, for example, quantitative research on the causes and effects of civil conflict. This subfield hosts one of the most cited quantitative empirical papers in recent decades (Fearon and Laitin 2003), but it is also a subfield where one might expect that causal empiricist approaches would be difficult to employ. I searched all quantitative papers on civil conflict that make causal claims published in *American Political Science Review*, *American Journal of Political Science*, and the *Journal of Politics*. Thirty of 34 papers (88%) published between 2000 and 2010 relied on the conventional multiple regression design.[4] Those four standouts went further in trying to estimate causal relations by using either instrumental variables, matching, or panel fixed effects methods to try to improve causal inference. Between 2011 and 2015, conventional regression studies accounted for 32 of 48, or 67%, a marked drop. Those 16 other studies applied methods such as instrument variables, regression discontinuity, difference in differences, or matching combined with explicit discussions of identification.[5] This indicates a shift in an area where causal factors of interest are quite stubborn in terms of their manipulability. Causal empiricist research in conflict studies is substantively rich and diverse, focusing on the effectiveness of counterinsurgency strategies, effects of foreign aid, institutional and economic roots of civil war, and postwar consequences of exposure to wartime violence, among other topics. This goes to show that causal empiricist research does not have to be narrow or focus on small questions (e.g., Deaton 2010; Huber 2013). In other subfields, such as the study of voting behavior (de Rooij, Green, and Gerber 2009; Green, McGrath, and Aronow 2013) or representation and accountability (Grose 2014), the move toward causal empiricism has been more thorough.

The sections below offer three considerations related to the move toward causal empiricism. First is a causal empiricist critique of what is still the prevailing convention: loosely specified and heroically interpreted regression studies. I will show that in terms of generalizability, conventional regression studies possess no special evidentiary advantage over experiments or natural experiments that estimate effects for well-defined subpopulations. Moreover, conventional regression studies are often highly compromised in terms of their internal validity. The declining prevalence of conventional regression studies relative to experiments and natural experiments is, therefore, welcome on internal validity grounds and does not represent a loss in terms of the generalizability of the findings.

Second is a statement of core pillars of causal empiricism. The goal here is to clear misconceptions and to show how causal empiricism puts emphasis on both statistical rigor and in-depth knowledge of specific cases. In doing so, causal empiricist research aims to establish credible causal facts understood for their specificity. Whether or not such facts generalize is not a question to be addressed definitively by a single study. Rather, these are questions to be addressed in research programs that consider collections of credible, specific facts in light of theoretical models. If journal editors want to be realistic in their promotion of credible causal research, they should expect quantitative studies to do less in terms of generalization and theory development and more in terms of identification. Generalization and theory development are better left to synthesis studies.

Given such realism about identification and specificity, where does this leave generalization? The section Pillars of Causal Empiricism addresses this question in terms of the relationship between causal empiricism and theoretical modeling. In debates about causal empiricism, a refrain among skeptics is that "theory is being lost" in the so-called identification revolution (Huber 2013). I address this concern by describing how causal empiricist research can fit into broader

3. I say informally motivated because in only very rare cases do researchers motivate control specifications on the basis of a structural model. See Morton (1999, 130–31) for a related discussion in political science.

4. The list of studies is available on the author's website.

5. The tally does not include field experimental studies on post-conflict development programs, such as Avdeenko and Gilligan (2015), Beath, Christia, and Eniolopov (2013), and Fearon, Humphreys, and Weinstein (2015).

research programs that also pursue theoretical modeling. Theoretical models provide lenses for interpreting specific empirical results in terms that are generalizable.

This essay focuses on conceptual issues and does not go into identification strategies and techniques employed in causal empiricist research. For reviews covering such techniques, readers should consult Imbens and Wooldridge (2009) or Keele (2015b). Textbook treatments are given by Angrist and Pischke (2009), Dunning (2012), Hernan and Robins (2013), Imbens and Rubin (2015), and Morgan and Winship (2015). Nor do I discuss the relationships between quantitative and qualitative research, which for political science applications are covered in the contributions to Brady and Collier (2010). Finally, the focus here is on quantitative causal studies. Alternative modes of quantitative analysis include development of quantitative measures of latent or otherwise hard-to-observe phenomena, pure prediction problems, for which machine learning has contributed to recent advances (Kleinberg et al. 2015), and descriptive characterizations of trends or equilibrium relationships. My focus on quantitative causal research does not imply any disregard for these other modes of empirical research. The discussion of research programs below takes such research to be complementary to causal research. At the same time, Ashworth, Berry, and Bueno De Mesquita (2015) show that many points raised by causal empiricists are relevant for these other modes of quantitative analysis as well, and so even those who do not engage in causal research may find the discussion below interesting.

A CAUSAL EMPIRICIST CRITIQUE OF PREVAILING CONVENTIONS

I begin with a critique of the prevailing convention in quantitative causal research in political science, which is to use multiple regression methods to estimate causal relations on "general" data sets that are meant to be representative of a universe of cases of interest, whether in the form of a data set exhaustive of all units of interest (e.g., a cross-national study with all available country data) or a representative survey sample. The methodological literature has given much more attention to threats to internal validity for conventional regression studies, and these threats serve as a primary motivation for the turn to a causal empiricist approach. And yet, conventional regression studies still dominate empirical practice in quantitative political science. This may be because researchers seek the comfort of methods that have the veneer of generalizability or seem to be reliable enough for estimating multiple causal effects at once. This is a false comfort. This section explains why, both with respect to generality and internal validity.

The pseudo-generality problem

The predominant approach to quantitative research in political science is a regression study on a data set representative of a universe of cases of interest. Researchers describe the findings from such studies in general terms, and they use summary statistics for the sample at hand when reasoning about scope conditions. Consider the following very typical excerpt, from the abstract of Hartzell and Hoddie (2003, 318): "Employing the statistical methodology of survival analysis to examine the 38 civil wars resolved via the process of negotiations between 1945 and 1998, we find that the more dimensions of power sharing among former combatants specified in a peace agreement the higher is the likelihood that peace will endure." Or consider this more recent excerpt from the abstract of Prorok (2016, 70): "These propositions are tested on an original data set identifying all rebel and state leaders in all civil conflict dyads ongoing between 1980 and 2011. Results support the hypothesized relationships between leader responsibility and war outcomes." These particular authors are by no means the exception—they operate well within the predominant mode of quantitative empirical inference. But that is exactly the problem. This type of generalizing reflects a presumption that by using a data set representative of some target population ("the 38 civil wars . . .", "all rebel and state leaders in all civil conflict dyads . . ."), the study will produce findings generalizable to that population. Such a presumption is sometimes used to justify the use of the conventional regression study design over a research design based on an experiment or natural experiment that is clearly limited to a specific subpopulation.[6]

Statistically speaking, this line of reasoning is completely misguided and has led to problematic judgments about the merits of studies using "general" data relative to experiments or natural experiments using data from more tightly defined subpopulations. It is the identifying variation that determines which units in a sample contribute to an effect estimate, not the mere presence of a unit in a sample. The key concept for characterizing identifying variation is "positivity" (Hernan and Robins 2013, 29–30; Petersen et al. 2011), also known as the condition of "overlap" in covariate (i.e., control variable) distributions over values of the treatment variable (Imbens 2004). Supposing for a set of units indexed by i, one is interested in the effect of some treatment of interest T_i that takes values defined by the set $\mathcal{T}$. Each unit in the population is characterized by "potential outcomes"

6. Exemplary instances of such arguments include Bardhan (2013), as cited in Aronow and Samii (2016), and Huber (2013), who compares a "traditional regression-type paper" to a hypothetical natural experiment in Sweden.

corresponding to each treatment value, denoted by a random variable $Y_i(t)$ for all $t \in \mathcal{T}$. For each unit, the observed outcome, Y_i, equals the value of $Y_i(t)$ corresponding to the treatment that the unit received, t. Finally, suppose one controls for a vector of covariates X_i. The data "alone" only support causal comparisons where, in the neighborhood of a given value $X_i = x$, the sample includes overlapping units with different treatment values. It is in such neighborhoods that positivity holds.[7] The values of X for which positivity holds are the locations in the covariate space where one has identifying variation in the treatment.

Grasping positivity and overlap is easy, as the following shows.[8] Suppose at time 1 we randomly assign households in one county in northern California either to receive pamphlets on income inequality or to receive nothing, and we want to estimate the effect of the pamphlets on household members' attitudes toward redistribution. Then, at time 2 we survey not only households in that one county, but in all counties in the United States, even though none of the other counties received pamphlets. Would this research design provide credible evidence on the average effect of the pamphlets for all US households? Clearly not, because the identifying variation is limited to but a small and specific segment of the US population. This example may seem contrived, but as I show below it resembles what occurs in conventional regression studies.

Where there is no overlap, one can only make comparisons with interpolated or extrapolated counterfactual potential outcomes values. King and Zeng (2006) brought the issue to political scientists' attention, characterizing interpolations and, especially, extrapolations as "model dependent," by which they meant that they were nonrobust to modeling choices that are often indefensible. By pointing out how common such model dependent estimates are in political science research, King and Zeng raised troubling questions about the validity of many generality claims in quantitative causal research in political science. They provided an algorithm for determining whether counterfactual comparisons are within the convex hull of the data, and thus in areas where positivity likely holds, although few studies seem to have applied these methods. Among those who take positivity and overlap seriously, the common reaction, and the one endorsed by Ho et al. (2007), has been to resort to other estimation methods like matching estimators (e.g., Gilligan and Sergenti 2008). Matching estimation forces the researcher immediately to confront the reality of limited overlap. By dropping cases in areas with no overlap (e.g., by using matching calipers), one consciously limits the scope of one's inference (Banerjee and Duflo 2009, 162–63).

Aronow and Samii (2016) take these points further. The conventional research design, which marries regression with a representative sample of some target population (all countries in the world, the population of the United States, etc.), typically fails to yield results that generalize to the target population or even to natural subpopulations where positivity holds. That is, conventional regression studies can be worse than our toy example of the pamphlet experiment. To see why, first suppose that one uses ordinary least squares (OLS) to estimate the average effect of T on Y with a regression of the form

$$Y = \alpha + \beta T + X'\gamma + \varepsilon,$$

and that the regression satisfies specification requirements such that the OLS estimate for β indeed estimates a causal effect.[9] This is even more generous than King and Zeng, for whom the dangers arose from model misspecification in areas of no overlap. Let τ_i denote the i-specific effect of changes in T_i on Y_i. This embeds the very reasonable assumption that effects are heterogeneous over the units. (Without such heterogeneity, the issue of generalizability would be moot anyway!) Then, the OLS estimator, $\hat{\beta}$, obeys

$$\hat{\beta} \xrightarrow{p} \frac{\mathrm{E}\left[w_i \tau_i\right]}{\mathrm{E}\left[w_i\right]}, \quad \text{where} \quad w_i = \left(T_i - \mathrm{E}\left[T_i | X_i\right]\right)^2.$$

The w_i weights are referred to as the multiple regression weights, and they characterize the extent to which a given observation contributes to $\hat{\beta}$.[10] The weights are largest in areas of the covariate space where the treatment is poorly explained—that is, for units where the treatment assignment is unpredictable given observables. The multiple regression weights are particular to linear regression methods, and Aronow and Samii go further to characterize general conditions needed to produce generalizable estimates. Under such conditions, various interaction-model, response-surface modeling, and weighting techniques are capable of producing effects that generalize to the population for which positivity holds.

The upshot is that the effective sample that gives rise to an effect estimated in a regression study can be quite dif-

7. That is, positivity or overlap holds for a population of interest and a set of treatment values, $\mathcal{T}' \subseteq \mathcal{T}$ if $0 < \Pr[T \in \mathcal{T}' | X_i = x] < 1$ for all x with positive measure in the population of interest (Hernan and Robins 2013, 30; Imai and van Dyk 2004, 855).

8. I credit Peter Aronow for this toy example.

9. Aronow and Samii (2016, thoerem 1) provides a precise statement of the assumptions.

10. Very similar results obtain for coefficients estimated via generalized linear models (logit, probit, etc.) and random coefficient models (Aronow and Samii 2016, 256–57).

ferent from the nominal sample with which one started. The effective sample is the transformation of the nominal sample after reweighting by the multiple regression weights. Aronow and Samii demonstrate with a study by Jensen (2003) on the effects of levels of democracy on foreign direct investment. Figure 1 reproduces the figure from Aronow and Samii (2016) for the Jensen example. The map on the left shows Jensen's nominal sample, which is representative of most of the world. The map on the right shows the effective sample that was the basis for the main result reported in the paper. The shading indicates the weight that each country receives in the respective sample. The first thing that strikes the eye is how radical the difference is between the nominal sample and effective sample. It really does resemble our pamphlet study example. The effective sample gives nonnegligible weight to only a few of the countries that were in the nominal sample. Here is the first indication that the results are not immediately general to the world that the nomimal sample is intended to represent. The top 12 contributing countries, accounting for more than half of the total weight applied for the main estimate in the paper, are (in descending order of their weights) Uruguay, Hungary, Niger, Philippines, Argentina, Madagascar, Pakistan, Zimbabwe, Poland, Peru, Lesotho, and Belarus. At first glance this appears to be an odd grab bag of countries. Upon further consideration one notices that many of these countries had rapid regime shifts due to military coups d'etat, while others had rapid regime shifts associated with the end of the Cold War.[11] As such, the study is based primarily on effects associated with these-specific types of transitions.

Suppose someone were to present studies that carefully sought to estimate the consequences of coups or post-Communist transitions on foreign investment. Journal reviewers operating on the basis of current conventions would, I suspect, criticize such studies for their lack of generalizability. I hope the hollowness of such criticisms is now clear: such criticisms draw an implicit comparison to either a fallacious interpretation of how conventional regression studies work or some unattainable ideal.

Conventional wisdom in political science about tradeoffs between generalizability and internal validity for different research designs is based on faulty foundations. There is no clear ordering of experiments, quasi-experiments, and observational studies that use regression or other control methods in terms of the generality of their findings. In observational studies, positivity is out of the control of the researcher, and it is typically limited to an idiosyncratic subset of the population (Dunning 2008, 291). Once we isolate areas of positivity, what looks on the surface like a "general" empirical analysis is often, in reality, a comparison within a highly specific subpopulation. What is disturbing is that authors of conventional regression studies typically have no clue about where positivity holds or how regression weights this subsample, and they make no effort to be transparent about it. In experiments, by construction, one controls positivity, although the reach of experiments is limited to causal factors available for direct manipulation and to subpopulations where we can run the experiments. But at least sample summary statistics will accurately portray the effective sample for experiments. For natural experiments, researchers have become accustomed to characterizing areas of identifying variation, as with the complier subpopulation for instrumental variables studies (Abadie 2003), the subpopulation near the cutoff for regression discontinuity studies (Lee and Lemieux 2010), or the comparison cases for difference-in-differences studies (Abadie 2005; Abadie, Diamond, and Hainmueller 2010; Abadie and Gardeazabal 2003). The same scrutiny should be applied to conventional regression studies: journals should insist that their research design sections report characteristics of the effective sample—this requires only examining the residual variance in the treatment after partialing out the controls.

The problem of pseudo-facts

We now turn to issues of internal validity—that is, questions of whether the causal "facts" that conventional regression studies produce are actually reliable. For causal identification, conventional regression studies rely on the assumption that the control variables, X, are adequate to account for confounding in the relationship between a causal factor of interest, T, and the outcome Y.[12] Moreover, such studies rely, to some extent, on getting functional forms correct. Many things can go wrong, and when they do the results produced from a conventional regression study are better thought of as "pseudo-facts." Below I review two important issues related to internal validity of conventional regression studies: misspecification and determination of control variables.

11. I thank Ali T. Ahmed for this astute observation.

12. One way to formalize this identifying assumption is in terms of conditional mean independence for a nonempty set of treatment values, $\mathcal{T}$, and a nonempty subset of the covariate space, $\mathcal{X}$:

$$\mathrm{E}\,[Y(t)|T = t,\ X = x] = \mathrm{E}\,[Y(t)|X = x] \quad \text{for all } t \in \mathcal{T}, x \in \mathcal{X}, \text{ and } 0 < \Pr[T \in \mathcal{T}|X = x] < 1.$$

This formulation is weaker than the more common assumption of the full distribution of potential outcomes, $Y(t)$, being independent of treatment, T, conditional on X, although as Imbens (2004) notes the case for conditional mean independence is rarely more compelling than the case for the stronger conditional independence assumption.

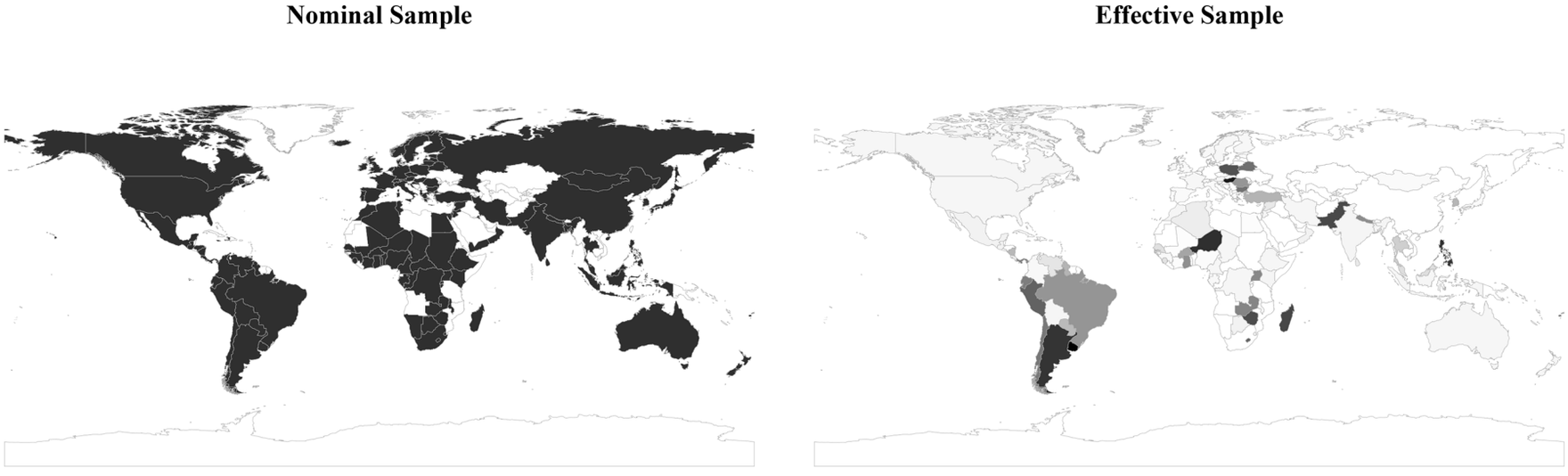

Figure 1. Nominal and effective samples from Jensen (2003), reproduced from Aronow and Samii (2016)

The first issue is associated with bias due to mispecification. In a paper that is well cited in political science, Achen (2005) demonstrated how misspecification for control variables undermines estimates of coefficients on treatment variables. The solution he proposed was that researchers should use formal theory and specification checks to make more deliberate functional form choices and also to define more homogeneous subpopulations within which to conduct one's analysis. But as Ho et al. (2007) explain, the problem with this "solution" is that it grants considerable latitude to researchers. As they put it, there is little credible basis to believe that conventional regression studies "are not merely demonstrations that it is *possible* to find a specification that fits the author's favorite hypothesis" (199; emphasis in original). Ho et al. propose matching as a more credible solution. Matching orthogonalizes treatment variables relative to control variables and thereby limits the extent to which control variable specifications affect coefficients estimates on treatment variables. Following on that work, there have been numerous methodological contributions that help to free researchers from the problems of misspecification, including advances in matching (as reviewed by Sekhon 2009) and nonparametric regression and machine learning methods (Hainmueller and Hazlett 2014; Hill 2011; Van der Laan and Rose 2011). Researchrs' increasing use of such methods to make the case for the robustness of their findings is a welcome development by the standards of causal empiricism.

The second issue concerns determination of control variables, a problem for which political scientists tend to rely on faulty heuristics in spite of guidance from causality theory (Angrist and Krueger 1999, 1291–93; Imbens 2004; Pearl 2009, chap. 3; Rosenbaum 1984). Conventionally, researchers use informal substantive arguments to motivate their sets of controls, often appealing to some notion of a "standard" set of controls for an outcome of interest. The underlying statistical motivation is typically based on the concept of "omitted variables" as taught in conventional regression textbooks. Unfortunately, such textbooks provide vague guidance leading to highly problematic decisions. Two commonly referenced textbooks by Greene (2008, 133–34) and Wooldridge (2009, 87–90) define omitted variable bias in terms of omitting control variables that (i) should appear in the "correct" or "true" specification for the outcome variable and (ii) are also correlated with the causal factor of interest. What this definition omits is that the "correct" specification depends on the effect that one wants to estimate. These differences are based on the causal ordering of the treatment and control variables (Elwert and Winship 2014; Pearl 2009, 17; Rosenbaum 1984). We need to ask, are we interested in a "total" effect, or some kind of "partial" effect (Pearl 2009, 126–32; VanderWeele 2015, chap. 2)? If it is a partial effect, is it causally identified under the given specification and given assumptions that we are really willing to believe? Rather, researchers still tend to take the textbook characterization of omitted variable bias to mean that all variables correlated with treatment variables and outcomes should be controlled in order to obtain unbiased causal estimates. Rote application of the textbook characterization of omitted variables contributes to the problem of "bad control" (Angrist and Pischke 2009, 64–68)—that is, causally incoherent control for "posttreatment" variables, meaning variables that are causal descendants of the treatment of interest. The result is vagueness, if not horrendous bias and inconsistency, in the estimated causal effects.

Take Fearon and Laitin (2003), who use a regression analysis to challenge the idea that ethnic structure affects civil war risk. To do so, they examine the relationship between ethnic fractionalization and civil war onset. A headline finding of the study—one of the supposed "facts" that it establishes—is that "factors that explain which countries have been at risk for civil war are not their ethnic or religious characteristics" (75, abstract). Column 1 of table 1 replicates Fearon and Laitin's main results, showing a small and statistically insignificant coeffient on ethnic fractionalization. As measured for this study, ethnic fractionalization is an unchanging characteristic of a country.[13] Now, Alesina and La Ferrara (2005) review studies showing a strong negative relationship between ethnic fractionalization and social and economic development. Thus, how should we interpret a coefficient produced from a model that includes economic and social factors that are widely believed to be affected by ethnic fractionalization? In fact, the unconditional correlation between ethnic fractionalization and civil war onset is really strong. This is shown by the bivariate regression in column 2, as well as in columns 3 and 5, which account for country-level clustering (given that ethnic fractionalization does not vary from year to year) and then also the "prior war" variable (to mimic Fearon and Laitin's approach to handling dynamics). Only when we "control" for per capita income do we get the insignificant result, as shown by columns 4 and 6.[14] Now, Fearon and Laitin acknowledge this point.[15] But it still raises important questions. These data exhibit the same pattern that Alesina and Ferrara summarize—a very strong negative correlation between ethnic fractionalization and income, as shown in column 7. Thus,

13. Ethnic fractionalization is constant for all countries in the data set except USSR/Russia and Yugoslavia, owing to those countries' break-ups.

14. Introducing any of the other control variables, on their own, does little to change the large, significant coefficient on ethnic fractionalization.

15. In their abstract, they are clear that "after controlling for per capita income, more ethnically or religiously diverse countries have been no more likely to experience significant civil violence."

Table 1. Replication and Auxiliary Analyses for Laitin and Fearon (2003)

	Outcome						
	Civil War Onset						Per Capita Income
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimator	Logit	Logit	Logit	Logit	Logit	Logit	OLS
Prior war	−.95 **				−.24	−.38	
	(.31)				(.23)	(.25)	
Per capita income	−.34***			−.29***		−.29***	
	(.07)			(.07)		(.07)	
Ethnic fractionalization	.17	1.12***	1.12**	.35	1.16**	.40	−4.14***
	(.37)	(.33)	(.42)	(.39)	(.43)	(.40)	(.90)
Observations	6,327	6,610	6,610	6,373	6,610	6,373	6,373
Country-clustered SEs			Y	Y	Y	Y	Y

Note. Regression coefficients with standard errors in parentheses. To save space the table omits from column 1 coefficients for the following control variables: log(population), log(% mountainous), noncontinguous state, oil exporter, new state, instability, democracy, religious fractionalization, and the constant term.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

it is not at all clear that ethnic fractionalization is unrelated to conflict, at least in terms of its "total" effect. That such an effect may operate via effects on income does not change this basic conclusion.[16]

The Fearon and Laitin paper is over a decade old, but problems of bad control remain ubiquitous, meaning that we should doubt a tremendous amount of the purported facts established by quantitative political scientists. Acharya, Blackwell, and Sen (2016) find that over half of quantitative papers published in top political science journals since 2010 suffer from "bad control." They review methods developed by Robins (1997) and VanderWeele (2015) for getting at what many researchers seem to really want—a type of partial effect known as the "controlled direct effect." Generally speaking, valid estimation of such an effect requires more than just plopping posttreatment variables into a linear regression specification. Because of this, along with other endogeneity concerns, there is no good reason to think that the coefficient in column 1 of table 1 captures a meaningful partial effect either.

Other "omitted variables" fallacies arise in interpreting the consequences of changes to the set of control variables. Suppose we have a study suggesting that T affects Y on average, controlling for X. Another researcher comes along and suggests that, actually, we need also to control for the variable W in addition to X, and in doing so, the estimated effect of T is now very small. The conventional interpretation would invoke the logic of "omitted variables," concluding that the original study probably did a poor job of estimating the average effect of T and the second study provides an improvement. Is this a reasonable conclusion? The results from Aronow and Samii (2016), discussed above, would have us wonder whether inclusion of W may have merely shifted the effective sample toward a subpopulation for which the effect of T is weak. In that case there may have been nothing wrong with the first study. The analysis by Achen (2005) would have us wonder whether the change is a result of misspecification for X or W. A third possibility is bias amplification: there was residual confounding in the first regression, but the second has only amplified the bias and made things worse (Clarke 2005; Pearl 2010). Once we consider these possibilities alongside the conventional "omitted variables" interpretation, it is clear that the change in the coefficient on T has at least four explanations for it, each being difficult if not impossible to distinguish!

A better conclusion is that the conventional regression studies are deeply problematic in terms of their causal content. Reflecting on the indeterminacies that plague the search for control variables in quantitative political science research, Clarke (2005) argued for "substituting research design for control variables" and to "test broad theories in narrow, focused, con-

16. At the same time, we should be skeptical about whether table 1 conveys any meaningful causal relationships, given that we have no reason to believe that any of the regressions succeed in either identifying the causal effect of fractionalization or applying proper functional forms.

trolled circumstances" (349–50). This is precisely the reorientation that the causal empiricist turn is trying to establish.

PILLARS OF CAUSAL EMPIRICISM

Causal empiricism is an approach to quantitative empirical analysis that pursues well-identified and specific causal facts. The pillars of causal empiricism that distinguish it from the prevailing convention include (i) realism about whether a research design is adequate to identify a causal effect and (ii) realism about the specificity of empirical results. Neither statistical technique nor the goal of causal inference distinguishes causal empiricism from the prevailing convention that the previous section examined: causal empiricist research is sometimes based on regression techniques, and conventional regression studies regularly aim to make causal inferences. If someone asks, "What is it that makes a causal empiricist study special?," the answer should be "careful use of an identification strategy research design and interpretation of the specificity of the results." The following subsections develop these ideas about the pillars of causal empiricism.

Identification by design

Conditions for causal identification are easy to state (positivity, conditional independence) but realizing them is not easy. Sekhon (2009, 503) writes poignantly, "Without an experiment, a natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive. This conclusion has implications for the kind of causal questions we are able to answer with some rigor. Clear, manipulable treatments and rigorous designs are essential. And the only designs I know of that can be mass produced with relative success rely on random assignment. Rigorous observational studies are important and needed. But I do not know how to mass produce them." Thus, neither the data-analysis technique, sample, nor control variables make the identification but rather the research design and the way that it exploits identifying variation (Freedman 1991; Rosenbaum 1999). Using an instrumental variables estimator does not imply causal identification if exclusion does not hold. Using a matching estimator does not imply causal identification if there is no plausible basis for conditionally exogenous treatment assignment. We must be able to answer the question, For two units of causal observation that are identical in terms of all important background characteristics, how could it be that they might differ in the treatments they receive?[17] This requires understanding the processes through which treatment values are determined—what Imbens and Rubin (2015, 34) describe as the "assignment mechanism." Multiple regression is frequently used to analyze randomized experiments. The causal credibility is qualitatively higher than a multiple regression study in which the regressor of interest is not randomly assigned.

But nature rarely provides sources of identifying variation, and experiments require considerable effort. For this reason, causal empiricism demands that empirical studies give extraordinary attention to analyzing and characterizing sources of identifying variation. The arguments should be careful about what kinds of effects are identified and they should draw on intimate knowledge and evidence regarding the topic (Titiunik and Sekhon 2012). Substantively rich identification debates should be welcomed, such as the debates between Albouy (2012) and Acemoglu, Johnson, and Robinson (2001, 2012) over the use of settler mortality as an instrument for colonial era institutional investments, or between Fewerda and Miller (2014, 2015) and Kocher and Monteiro (2015) over the exogeneity of the Vichy-German administrative border placement. Credibility of empirical findings depends on being able to stand up to critiques based on in-depth knowledge.

The logic of causal empiricism is sometimes described as the study of "the effects of causes" rather than the "causes of effects" (Holland 1986). This is based on realism about the difficulty of causal identification. To hope, much less demand, that a single paper investigate the effects of multiple treatments is a very tall order. In terms of identification, what would be required are factorial experiments or whatever analogues there may be among natural experiments. For natural experiments, each treatment would need its own specific, in-depth evidence to make the identification credible.[18] This will be a hard idea to accept for those steeped in the prevailing convention, where researchers regularly attempt the heroic feat of trying to evaluate the causes of multiple effects in one analysis—usually in one regression. These analyses may turn up intriguing correlations. But what the causal empiricist asks is that audiences recognize the large gulf in the credibility of causal facts established via strong identification research designs and those produced through conventional regression studies.

Specific causal facts

Causal empiricism is an approach that is realistic about the specificity of the causal estimates that we can obtain. This is

17. Implicit here is the assumption that the "units of causal observation" are ones for which the "stable unit treatment value assumption" (SUTVA) holds (Aronow and Samii 2015; Imbens and Rubin 2015, 11–12).

18. Adjudicating between causal mechanisms is, however, much more in line with the "effects of causes" approach.

an implication of the fact that causal identification is difficult to obtain. The local average treatment effect (LATE) theorem (Angrist, Imbens, and Rubin 1996) is a formal expression of such realism about specificity. The LATE theorem states that under a set of basic identifying conditions, an instrumental variable identifies the average causal effect for the subpopulation of units whose treatment status is in fact moved by the instrument. Summary statistics describing this subpopulation can be computed using the kappa-weighting results of Abadie (2003). The result from Aronow and Samii (2016) described above is a LATE-type result, showing that under the relevant identifying assumptions, linear regression estimates are consistent for the average causal effect local to a subpopulation whose traits can be characterized by reweighting the nominal sample by the multiple regression weights. Similarly, regression discontinuity identifies effects local to the relevant cut points, matching with calipers identifies effects local to the region of common covariate support, experiments identify effects local to the typically nonrepresentative sample of experimental subjects, and so on. Only under highly ideal circumstances, which are unlikely to apply in political science research, will we obtain empirical estimates that are immediately general to some "global" population. The realistic conclusion to draw is that all quantitative empirical results that we encounter are "local" (Angrist and Pischke 2010, 23–24).

Such realism is sure to make many political scientists uncomfortable. But under prevailing conventions, generalization from highly specific results is rampant with little recognition that this is actually what is going on. As such, much research is blind to the assumptions needed for such generalization. This blindness is harmful in that causal questions are not given any more scrutiny than can be explored in a single contribution due to the sense that causal estimates from one study actually answer the causal question generally. It leads journal editors to reject studies that focus on obtaining well-identified, if specific, estimates of important causal quantities for lack of novelty or for their specificity. Armed with a better understanding of how empirical results are always local and assumptions necessary for effect homogeneity are often dubious, we should welcome the pursuit of opportunities to estimate important, but not necessary novel, causal effects for new subpopulations. This is how one tests and bounds the scope of theoretical claims.

CAUSAL EMPIRICISM AND THEORY

Some have charged that "theory is being lost" amid the turn to causal empiricism and that causal empiricism leads to the pursuit of only narrow questions (Huber 2013). In a way, this is a valid concern, at least when it comes to evaluating what an individual paper aims to accomplish. But to evaluate causal empiricism in these terms misses the point. Causal empiricism forces realism about what we should expect from paper-length research contributions and therefore forces us to think in terms of research programs. Above, I explained why we need to disabuse ourselves of two fantasies: (i) a single, paper-length empirical analysis is likely to yield nonspecific causal facts that generalize without strong assumptions, and (ii) there are ready techniques to produce such facts at will for populations of one's choosing. In a causal empiricist paper, the absence of novel theorizing does not have to mean that "theory is being lost" but rather that theory is being held constant as we go about the difficult business of trying to do credible causal inference. This orientation respects not only the difficulty of causal inference, but it also respects the difficulty of developing good theory that is capable of explaining a variety of facts.

Causal empiricism emphasizes research design in pursuit of causal identification. Causal empiricist observational studies expend considerable space to justify identification. Experimental studies typically have less to prove on this front. Nonetheless, experimental papers tend to use considerable space to describe the research design. Naturally this will leave less room to do other things, such as presenting new theoretical models. But even if extra space were granted, any model building should be done under realism about specificity. A model that makes claims consistent with the evidence from one study is limited in its generality by the study's effective sample. This suggests that model building is typically more compelling when it synthesizes results from numerous studies, hence the proposal to work with existing models at times rather than always proposing a new model (Angrist and Pischke 2010, 23).

The point is that there is no inherent tension between causal empiricism and theoretical modeling. An empiricist research program builds up to general knowledge through incremental accretion of credible findings across a diversity of settings (Keele 2015a, 104). This can happen either organically as researchers happen to discover new opportunities for empirical work, or consciously by deploying, across a variety of contexts, a set of studies designed to allow for comparative analysis of causal estimates. A recent issue of the *American Economic Journal: Applied Economics* focusing on six field experiments on micro-credit is exemplary (Banerjee, Karlan, and Zinman 2015), as are the "Metaketa" research programs managed by the Evidence in Governance and Politics network (http://egap.org/metaketa). Theoretical modeling could be conducted in synthesis pieces in which more evidence can be assessed than what is contained in a single paper-length empirical analysis.[19]

19. Dehejia, Pop-Eleches, and Samii (2015) discuss the current state of the art in the statistical synthesis of experimental and natural experimental results.

Such syntheses could consider not only credible quantitative causal research, but other types of empirical research such as descriptive statistical analyses, ethnographies, and other qualitative studies. Such a collection of credible facts establishes the puzzles and contours that theorists can use to assess the usefulness of behavioral models and develop new hypotheses to guide further empirical research. As Gehlbach (2015) discusses, it is unreasonable to think any individual or any single paper could do all of these things well, which motivates the need for division of labor in a research program.

An experiment or natural experiment is especially interesting if it provides an opportunity to assess the value of competing models of causal mechanisms. Empirical analyses do not "prove" or "disprove" models—as Clarke and Primo (2012) discuss, to take this as the goal of empirical work is generally nonsensical, and even more so once we appreciate that all estimates are "local." Rather, credible empirical work clarifies situations where one or another model is useful. When certain models tend to guide policy or other decisions, it is crucial for empirical research to demarcate scope conditions, expose areas where model propositions fail, and establish the need for richer models. In labor economics, the research program on minimum wage laws has followed such an evolution, driven by causally well-identified studies and prompting new models (see Schmitt [2013] for a review).

An excellent venue for theoretical framing is a research design and analysis plan, where one can specify how a research design and empirical analyses allow one to assess competing models. At present, research design and analysis plans tend to focus mostly on statistics.[20] They should do much more theoretical framing, answering the question, what is at stake for competing models in the analyses being proposed? At conferences and seminars researchers should be spending much more time discussing these kinds of model-framing research design and analysis plans—arguably, it is at this stage that broad feedback is more critical than after the results are in.[21]

A second theory-related critique comes from "structuralists" who find that causal empiricist research puts too little effort into interpreting experiments and natural experiments on the basis of fully specified behavioral models (Deaton 2010; Heckman and Urzua 2010; Wolpin 2013). The structural approach to causality is indeed different in its goal of using data from specific settings to estimate parameters of general models of behavior (that is, "causes of effects" models; Heckman 2010, 361). The debate about causal inference between those working in the structuralist versus empiricist (or "reduced form") traditions is mostly among economists; in political science journals, structural estimation is still almost exclusively applied in latent factor measurement (e.g., voter ideal points; Quinn, Martin, and Whitford 1999) or addressing confounding due to strategic interaction (following Signorino 1999). Nonetheless, my expectation is that as the prevailing convention described above continues to wither, that the relationship between causal empiricism and structural causal analysis will become more important in political science.

I have sympathy for the structuralist view and believe that there are fruitful ways to bridge these two approaches. First, we should be clear on where the two approaches tend to agree. Current structural analyses, in economics at least, also emphasize identification and clear definition of counterfactual comparisons (Heckman 2010). Gone are the days when someone could get away with relying heavily on structural assumptions to identify an average causal effect in data that are plagued by endogeneity problems.[22] The key difference today, I would say, is in the two approaches' respective treatments of the specificity issue.[23] With structural estimation, identifying variation defines the opportunity to estimate model parameters (or combinations of such parameters), which are presumed to be invariant and therefore permit simulation of counterfactuals and generalization to new settings. For causal empiricists, counterfactual comparisons are limited to what the data identify directly and generalization occurs only after a set of facts are obtained. Moreover there is typically no a priori reason to believe invariance assumptions for structural models. The logic of the LATE theorem and other "localness" results extend immediately to attempts to estimate parameters in structural models, as Angrist, Graddy, and Imbens (2000) show in a nonparametric analysis of a simultaneous-equations supply

20. See Humphreys and de la Sierra Raul Sanchez de la Sierra (2013) and Monogan (2013) other contributions to the 2013 *Political Analysis* symposium on research registration.

21. Nyhan (2015) takes this idea even further, proposing that journals could make publication decisions on the basis of research design and analysis plans, so that contributions are assessed on the basis of their theoretical framing and methods, rather than on the basis of whether they find "significant" results. The cognition and neuroscience journal *Cortex* has begun to apply this model in their "registered reports" section. A few inter-institutional research working groups, including the Working Group in African Political Economy (WGAPE), Experiments in Governance and Politics (EGAP), and Northeast Workshop in Empirical Political Science (NEWEPS) regularly devote space on conference agendas to research designs and analysis plans.

22. That ship began to sail as early as the publication of Lalonde (1986). Of course, instrumental variables have their origin in work on identifying structural models. But until recently exclusion restrictions were assumed in a manner that was very fast and loose and, by current standards, quite unconvincing (Angrist and Pischke 2010).

23. Angrist and Krueger (1999, 1280) draw a similar distinction.

and demand model. So even if one identifies structural parameters from a given study, generalizability remains an open question. What is nice about the localness results is that they provide ways to characterize the sets of units that contribute to parameter estimates.

A bridge between the two approaches is to view structural analysis as a tool for theoretical framing and interpretation that can inform the evolution of the research program. Consider the analysis by Wolpin (2013, 127–33) of the Project STAR class size experiment (Krueger 1999). His analysis shows that the causal relations identified by the experiment may be too coarse to make predictions about what would happen if class size reductions were applied on a larger scale. This analysis defines what further research is necessary to answer questions about consequences of scaling up. An example of structural analysis for theoretical framing from political science is by Brollo and Nannicini (2012), who unpack a causal effect identified by a regression discontinuity design in a study of political alignment and federal transfers. More ambitious would be efforts toward "structural synthesis." Causal empiricist research would deliver a set of credible empirical results for which contextual conditions are clearly stated. Then, one would assess the restrictions these findings imply on parameter values for behavioral models. In a 2011 special issue of *Journal of African Economies*, Fafchamps (2011), Harrison (2011), McKenzie (2011), and Wantchekon and Guardado (2011) discuss structural analysis of randomized controlled trials. Wolpin (2013) gives structural interpretations of a variety of experimental and natural-experimental results for examples in labor economics. Chetty (2009) discusses ways to derive "sufficient statistics" from experimental and natural experimental results to inform model-based welfare analysis. Tamer (2010) reviews methods for using empirical results to bound parameter combinations for structural models. Keniston (2011) gives a nice example of using structural methods to reanalyze experimental results to develop richer counterfactual implications. Current methods training should prepare students to analyze how behavioral models and causal identification relate.

CONCLUSION

Our journals continue to churn out conventional regression studies that try to estimate causal effects and then interpret them in general terms.[24] This reflects how most political scientists (including myself) were trained. Perhaps most importantly, it reflects how journal editors tended to be trained. I worry that it also reflects beliefs that such studies can generate credible and generalizable facts at will. The first aim of this essay was to make clear that such beliefs are generally false. Those of us trained in conventional regression methods have much to unlearn. Conventional regression studies rely on identifying variation that is out of the researcher's control, and as such they generate estimates that are specific only to certain subpopulations. And this they only do if the regression methods are applied sensibly. Conventional practice does not lead to sensible use of regression and so even some of the most seminal findings from recent political science research are dubious.

Causal empiricism represents a more realistic approach to quantitative causal research, emphasizing the importance of good research design for causal identification and the specificity of the causal facts that are obtained even in the best of circumstances. The second aim of this essay was to clarify these pillars of causal empiricism.

Empirical contributions need to devote more space to research design and characterization of the subpopulations for which effects are identified. As such, a single empirical contribution should only devote a limited amount of space to theory development. The gains from such revisions to the concept of an individual empirical contribution should be the accumulation of more credible findings. Once a set of such findings accumulates, we should be in a much better position to evaluate theoretical models in terms of the scope of their usefulness. The fascination with theoretical novelty in every empirical paper should be replaced with more appreciation of work that brings increasingly refined empirical scrutiny to bear on existing theoretical models. This should excite modelers as well because it would provide for them a richer set of facts to use when considering new directions. The third aim of this essay was to propose that credible empirical research should interact with theoretical models as part of research programs and a division of labor. The next chapters in the "credibility revolution" may very well be in further synthesis of causal inference and behavioral modeling.

My focus on quantitative causal research does not imply a disregard for descriptive quantitative research or qualitative research. Descriptive regression studies and analyses of trends can define puzzles that establish research programs. For example, the negative relationship between ethnic diversity and development described above was the product of important scientific contributions to measurement and establishes an intriguing puzzle. The argument here is that in trying to move from intriguing relationships to causal statements, credibility demands of researchers much more

24. As I was writing this essay, the most recent issue of the *American Political Science Review* (November 2015) contained four quantitative causal studies, examining causal effects of discrimination, government transparency, remittances, terrorism. All four were conventional regression studies in the style described above.

effort and care in establishing sets of well-identified empirical results and interpreting the specificity of their findings than is the case under the prevailing convention.

ACKNOWLEDGMENTS

Thanks to Pablo Argote for excellent research assistance and to Jeffery Jenkins, Peter Aronow, Jonathon Baron, Neal Beck, Kevin Bryan, Andrew Clarke, Michael Gilligan, Macartan Humphreys, Helen Milner, Elizabeth Levy Paluck, Brenton Peterson, and Sam Plapinger for helpful discussions.

REFERENCES

Abadie, Alberto. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 113:231–63.

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72 (1): 1–19.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490):493–505.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93 (1): 113–32.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91 (5): 1369–1401.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Reply." *American Economic Review* 102 (6): 3077–3110.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* (forthcoming).

Achen, Christopher H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22 (4): 327–39.

Albouy, David Y. 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment." *American Economic Review* 102 (6): 3059–76.

Alesina, Alberto, and Eliana La Ferrara. 2005. "Ethnic Diversity and Economic Performance." *Journal of Economic Literature* 43 (3): 762–800.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In Orley C. Ahsenfelter David Card, eds., *Handbook of Labor Economics*. Vol. 3 Amsterdam: North Holland.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies* 67:499–527.

Aronow, Peter M., and Cyrus Samii. 2015. "Estimating Average Causal Effects Under General Interference." Unpublished manuscript, Yale University and New York University.

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60 (1): 250–67.

Ashworth, Scott, Christopher R. Berry, and Ethan Bueno De Mesquita. 2015. "All Else Equal in Theory and Data (Big or Small)." *PS: Political Science and Politics* 48 (1): 89–94.

Avdeenko, Alexandra, and Michael J. Gilligan. 2015. "International Interventions to Build Social Capital: Evidence from a Field Experiment in Sudan." *American Political Science Review* 109 (3): 427–49.

Banerjee, Abhijit V., Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics* 7 (1): 1–21.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1:151–78.

Bardhan, Pranab. 2013. "Little, Big: Two Ideas about Fighting Global Poverty." *Boston Review* (May/June), online journal.

Beath, Andrew, Fotini Christia, and Ruben Eniolopov. 2013. "Empowering Women through Development Aid: Evidence from a Field Experiment in Afghanistan." *American Political Science Review* 107: 540–57.

Brady, Henry E., and David Collier, eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. 2nd ed. Lanham, MD: Rowman & Littlefield.

Brollo, Fernanda, and Tommaso Nannicini. 2012. "Tying Your Enemy's Hands in Close Races: The Politics of Federal Transfers in Brazil." *American Political Science Review* 106 (4): 742–61.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge between Structural and Reducted Form Methods." *Annual Review of Economics* 1:451–87.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22 (4): 341–52.

Clarke, Kevin A., and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford: Oxford University Press.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48:424–55.

Dehejia, Rajeev H., Cristian Pop-Eleches, and Cyrus Samii. 2015. "From Local to Global: External Validity in a Natural Fertility Natural Experiment." NBER Working paper 21459, National Bureau of Economic Research, Cambridge, MA.

de Rooij, Eline A., Donald P. Green, and Alan S. Gerber. 2009. "Field Experiments on Political Behavior and Collective Action." *Annual Review of Political Science* 12:389–95.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61 (2): 282–93.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40:31–53.

Fafchamps, Marcel. 2011. "Randomised Controlled Trials or Structural Models (or Both . . . or Neither . . .)?" *Journal of African Economies* 20 (4): 596–99.

Fearon, James D., and David Laitin. 2003. "Ethnicity, Insurgency and Civil War." *American Political Science Review* 97:75–90.

Fearon, James D., Macartan Humphreys, and Jeremy M. Weinstein. 2015. "How Does Development Assistance Affect Collective Action Capac-

ity? Results from a Field Experiment in Post-Conflict Liberia." *American Political Science Review* 109 (3): 450–69.

Fewerda, Jeremy, and Nicholas L. Miller. 2014. "Political Devolution and Resistance to Foreign Rule: A Natural Experiment." *American Political Science Review* 108 (3): 642–60.

Fewerda, Jeremy, and Nicholas L. Miller. 2015. "Rail Lines and Demarcation Lines: A Response." SSRN Working paper 2628508.

Freedman, David A. 1991. "Statistical Models and Shoe Leather." *Sociological Methodology* 21:291–313.

Gehlbach, Scott. 2015. "The Fallacy of Multiple Methods." *Comparative Politics Newsletter* 25 (2): 11–12.

Gilligan, Michael J., and Ernest J. Sergenti. 2008. "Does Peacekeeping Cause Peace? Using Matching to Improve Causal Inference." *Quarterly Journal of Political Science* 3:89–122.

Green, Donald P., Mary C. McGrath, and Peter M. Aronow. 2013. "Field Experiments and the Study of Voter Turnout." *Journal of Elections, Public Opinion and Parties* 23 (1): 27–48.

Greene, William H. 2008. *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Pearson.

Grose, Christian R. 2014. "Field Experimental Work on Political Institutions." *Annual Review of Political Science* 17:355–70.

Hainmueller, Jens, and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22 (2): 143–168.

Harrison, Glenn W. 2011. "Randomisation and Its Discontents." *Journal of African Economies* 20 (4): 626–52.

Hartzell, Carolyn, and Matthew Hoddie. 2003. "Institutionalizing Peace; Power Sharing and Post Civil War Conflict Management." *American Journal of Political Science* 47:318–32.

Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48 (2): 356–98.

Heckman, James J., and Sergio Urzua. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156 (1): 27–37.

Hernan, Miguel A., and James M. Robins. 2013. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC.

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20 (1): 217–40.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Huber, John. 2013. "Is Theory Getting Lost in the 'Identification Revolution'?" *The Money Cage*. http://themonkeycage.org/2013/06/is-theory-getting-lost-in-the-identification-revolution/.

Humphreys, Macartan, and Raul Sanchez de la Sierra Raul Sanchez de la Sierra. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21 (1): 1–20.

Imai, Kosuke, and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99 (467): 854–66.

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 (2009)." *Journal of Economic Literature* 48:399–423.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press Press.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment." *International Organization* 57:587–616.

Keele, Luke. 2015a. "The Discipline of Identification." *PS: Political Science and Politics* 48 (1): 102–6.

Keele, Luke. 2015b. "The Statistics of Causal Inference: A View from Political Methodology." *Political Analysis* 23 (3): 313–35.

Keniston, Daniel E. 2011. "Experimental vs. Strucural Estimates of the Return to Capital in Microenterprises." Unpublished manuscript, Yale University.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14 (2): 131–59.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review: Papers and Proceedings* 105 (5): 491–95.

Kocher, Matthew Adam, and Nuno P. Monteiro. 2015. "What's in a Line? Natural Experiments and the Line of Demarcation in WWII Occupied France." SSRN Working paper 2555716.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.

Lalonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4): 604–20.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48:281–355.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

McKenzie, David. 2011. "How Can We Learn Whether Firm Policies Are Working in Africa? Challenges (and Solutions?) for Experiments and Structural Models." *Journal of African Economies* 20 (4): 600–625.

Monogan, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21 (1): 21–37.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research, Second Edition*. Cambridge: Cambridge University Press.

Morton, Rebecca B. 1999. *Methods and Models: A Guide to the Emprical Analysis of Formal Models in Political Science*. Cambridge: Cambridge University Press.

Nyhan, Brendan. 2015. "Increasing the Credibility of Political Science Research: A Proposal for Journal Reforms." *PS: Political Science and Politics* 48 (S1): 78–83.

Ogden, Timothy. 2015. "Experimental Conversations: Angus Deaton." *Medium.com* (Accessed Oct 13, 2015).

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference, Second Edition*. Cambridge: Cambridge University Press.

Pearl, Judea. 2010. "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates." In Peter Grunwald and Peter Spirtes, eds., *Proceedings of UAI*. Corvallis, OR: AUAI, 417–24.

Petersen, Maya L., Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. Van der Laan. 2011. "Positivity." In Mark J. Van der Laan and Sherri Rose, eds., *Targeted Learning: Causal Inference for Obervational and Experimental Data*. New York: Springer, 162–86.

Prorok, Alyssa K. 2016. "Leader Incentives and Civil War Outcomes." *American Journal of Political Science* 60 (1): 70–84.

Quinn, Kevin M., Andrew D. Martin, and Andrew B. Whitford. 1999. "Voter Choice in Multi-Party Democracies: A Test of Competing Theories and Models." *American Journal of Political Science* 43 (4): 1231–47.

Robins, James M. 1997. Causal Inference from Complex Longitudinal Data. In M. Berkane, ed., *Latent Variable Modeling and Applications to Causality*. New York: Springer-Verlag, 69–117.

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, Series A* 147 (5): 656–66.

Rosenbaum, Paul R. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science* 14 (3): 259–304.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2 (3): 808–40.

Schmitt, John. 2013. *Why Does the Minimum Wage Have No Discernible Effect on Employment?* Washington, DC: Center for Economic and Policy Research Reports.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12 (1): 487–508.

Signorino, Curtis S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93 (2): 279–97.

Tamer, Elie. 2010. "Partial Identification in Econometrics." *Annual Review of Economics* 2:167–95.

Titiunik, Rocio, and Jasjeet Sekhon. 2012. "When Natural Experiments Are Neither Natural Nor Experiments." *American Political Science Review* 106 (1): 35–57.

Van der Laan, Mark, and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Obervational and Experimental Data*. New York: Springer.

VanderWeele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.

Wantchekon, Leonard, and Jenny Guardado. 2011. "Methodology Update: Randomised Controlled Trials, Structural Models and the Study of Politics." *Journal of African Economies* 20 (4): 653–72.

Wolpin, Kenneth I. 2013. *The Limits of Inference without Theory*. Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2009. *Introductory Econometrics: A Modern Approach*. 4th ed. Mason, OH: South-Western.