Getting Off the Gold Standard for Causal Inference

Robert Kubinec

Niehaus Center for Globalization and Governance

Princeton University

January 29, 2019

Abstract

I argue that social scientific practice would be improved if we abandoned the notion that a gold standard for causal inference exists. Instead, we should consider adopting three silver standards relating to distinct theories of causation: experimental approaches with the counterfactual theory, large-N observational studies with Humean causation, and qualitative process-tracing with mechanistic causation. Under ideal conditions, any of these three theories provide strong evidence of an important part of what we mean by causality. I also present the use of statistical entropy, or information theory, as a possible yardstick for evaluating research designs across the silver standards. The concept of entropy shows that the weight given to any of these approaches depends on the relative reduction in our uncertainty of a given causal system.

You shall not crucify mankind upon a cross of gold.

– William Jennings Bryan, July 8, 1896

The renewed attention to causal identification in the last twenty years has elevated the status of the randomized experiment to the sine qua non gold standard of the social sciences. Nonetheless, research employing observational data, both qualitative and quantitative, continues unabated, albeit with a new wrinkle: because observational data cannot, by definition, assign cases to receive causal treatments, any conclusions from these studies must be descriptive, rather than causal. However, even though this new norm has received widespread adoption in the social sciences, the way that social scientists discuss and interpret their data often departs from this clean-cut approach. Analyses with observational data continue to be performed in a way that suggests causal interpretations, and even qualitative evidence is cited as valid explanations for human actions. This disjuncture between the clean logic of counterfactual causal inference and actual practice has created considerable confusion among researchers who must decide whether to devote precious resources to observational data collection that is fatally doomed regardless of its promise or a (quasi-)experiment even if that experiment would not answer their research question.

This paper inverts the meaning of *gold standard* by considering the use of the phrase in heated monetary debates of the 19th century. The quotation above by William Jennings Bryan referenced devaluation occurring in the United States as a consequence of pegging the dollar to gold reserves. When the price of gold increased, the money supply constricted and borrowers would have to pay more than they initially agreed to. Bryan and his confederates wanted to switch to the silver standard, of which there was a much more plentiful supply, in order to maximize much-needed specie for the United States' quickly growing economy.

The analogy made in this paper is that employing a gold standard in causality can cause a similar deflation of the evidence available to answer a research question. The underlying problem is the artificiality of value. The price of gold should not be able to determine the value of all commodities in an economy, and similarly experiments and the potential outcomes framework should not be used as the only yardstick for evaluating empirical research. The difficulty is not that experiments can have serious flaws, although they do, but rather that scientists have yet to come up with a single non-tautological definition of *what causality in fact is*. The social sciences are struggling not only with the estimation of uncertainty in methods, research design and the amount of data, but with knowing what it is exactly that we aim to accomplish with all this endeavor. Instead, we commit a subtle logical fallacy by assuming that the beneficial properties of certain statistical paradigms make them the definition of causality.

I reduce the literature on causality to three basic theories of causal inference that closely match the variety

of research designs social scientists employ. I define these as follows: the counterfactual theory of inference (which I also link to the closely-related manipulationist theory), the Humean theory of inference, and the mechanistic theory of inference. Each of these theories can be seen as a kind of *silver* standard, incorporating much of what we mean by the term causal, but failing at the same time to encompass all of what we mean (if we could somehow express it). Social scientists have been operating in these paradigms or moving between them without necessarily being aware of the distinction, but a greater appreciation for the diversity of meaning in causality would help explain our continued divergence in research practice. Pretending that this uncertainty does not exist–that there exists enough gold to supply our research economy–yields distortions and inefficiencies in the research process as we do not take advantage of all possible avenues of inquiry.

I propose that a helpful, possibly unifying, criterion for social science research is the concept of reducing entropy. Entropy is a concept widely used in the natural sciences and has important applications in statistics in the form of information theory. It is more of the latter that I turn to as an aid for evaluating the payoff of research designs across the three silver standards. Entropy has its own ambiguities of meaning, unfortunately, but it does provide a relatively neutral framework for discerning what kind of research will reduce entropy in our understanding of social relations. The contribution of the entropy criterion is to construct a framework by which to evaluate the relative utility of divergent causal paradigms over the same causal system.

I show that entropy when applied to causal graphs returns intuitive results that more directly follow research practice. In a given causal graph, it is entirely possible that an observational analysis that can only show associations will reduce the entropy of a causal graph as much as, or more than, an experiment that is truly "causally identified." As the aim of the social sciences should be to reduce our uncertainty in understanding how the social world operates, we would benefit from applying the word causal more liberally rather than force research designs into a hierarchical mold of descriptive vs. causal inference.

1 The New Intellectual Battlefield: Causality

The credibility revolution of the past fifteen (or so) years has produced a sea change in how political scientists, economists and increasingly others conceive of research and measure its success. The theories behind the credibility revolution, notably the potential outcomes framework, date back to the 1970s or even earlier (Fisher 1935; Rubin 1974; Holland 1986), but for whatever reason, the practice of formal randomized experiments did not take off in fields besides psychology until the 2000s (D. P. Green and Gerber 2003; Morgan and Winship 2007; Levitt and List 2008). More recently, a second credibility revolution has swept through social scientific disciplines that long employed experiments, particularly psychology. This second revolution has questioned

the use of discretized decision rules, i.e. p-values, as a source of inferring causal inference, and shown that many published experiments fail to replicate even if the original experiment had statistically significant results (Open Science Collaboration 2015; Gelman and Loken 2013). This revolution has emphasized pre-registering research questions (Nosek et al. 2018), sharing data so that conclusions can be replicated or reproduced (S. N. Goodman, Fanneli, and Ioannidis 2016), and even more radical changes, such as fundamentally altering the standards used to judge statistical inference (Benjamin et al. 2017). While the first revolution has dramatically elevated the status of experimental research designs, the second has ironically pointed to deep problems in how RCTs and quasi-experimental methods have been implemented and evaluated.

While this second revolution has just reached disciplines late-adopting disciplines like political science and economics, the first revolution has still barely reached its zenith in these same disciplines. As of this writing, many if not most political science, sociology and economics departments have causal inference as a central part of their methods instruction, though familiarity and comfort with counterfactual causal inference varies widely across researchers. Increasingly, presenting experimental results, whether in labs, the field or surveys, is no longer considered remarkable or ground-breaking, and to some it has become the standard for what social science research should be.

There remain, however, significant pockets of resentment at the success of the experiment, though many of only appear in private conversations using epithets like "barefoot experimentalist" and "causal inference Taliban."[1] The success of experiments appears to endanger the role of observational studies, whether qualitative or quantitative, as these studies can never meet the stringent criteria imposed by their randomized kin (Beck 2006; Gerber, Green, and Kaplan 2014; Gerstein, McMurray, and Holman 2019). As a result, previously popular methods like large-N time-series cross section models have come under criticism for failing to either estimate *average treatment effects* (ATEs) (Samii 2016; Arkhangelsky and Imbens 2018; Gibbons, Serrato, and Urbancic 2017), the causal criterion of the potential outcomes framework, or to account for missing variables and over-time dynamics (Plümper and Troeger 2019). Applied statisticians continue to emphasize the value of observational methods (Gelman 2011; Imai, King, and Stuart 2008), but researchers nonetheless are tempted to assign research designs into a hierarchy.

This tension has boiled over into published debates, most recently in a remarkably broad and heated discussion in the 2018 August issue of *Social Science and Medicine*. On one side, Deaton and Cartwright (2018) argue that the emphasis on RCTs as a cure-all for causal inference is over-blown because researchers often ignore the known limitations of their samples by reference to randomization. While some support Deaton and

1. These two epithets are those I have overheard in conversation with other scholars; I do not know if they are the most representative or common labels so applied.

Cartwright, including Gelman (2018) and Sampson (2018), others argue that recent research on understanding treatment heterogeneity and the application of experimental results to novel problems mitigate Deaton and Cartwright's concerns (Arkhangelsky and Imbens 2018; Ioannidis 2018). This brewing dispute has all the hallmarks of a noteworthy battle of the minds, but it is questionable as to whether all the discourse will do more for applied researchers than create more methodological minefields they must be wary of.

The main problem, I maintain, underlying these disagreements is an acute problem for social scientists: while we all want to obtain causal knowledge, we do not in fact know what causal knowledge is. At the very least, we struggle to define it precisely in a way that is non-tautological. Existing research shows that causal thinking is deeply connected to human thought processes (Sloman and Lagnado 2015). Perhaps because it is so foundational to how we process the world, we also have trouble encompassing precisely what we mean when we say a relation is causal versus spurious.

The difficulties in defining causation are nothing new as causality has been a subject of intense philosophical debate for several centuries, if not millennia. Rather, my point is that it is easy for scientists to ignore this source of uncertainty in discussions of causal inference, which leads to unrealistic standards for what a particular framework of causality can achieve. Indeed, some causal paradigms can become so entrenched that adherents no longer see them as paradigms but rather as the definition of the subject. It is when this conflation is reached–a certain methodology is defined as the gold standard–that inevitable distortions arise in evaluating research streams.

To establish this important though often-overlooked discrepancy, I briefly examine what I call the three silver standards of causality: counterfactual/manipulationist, correlational, and mechanistic inference. The intention of this overview is not to be exhaustive, as that would likely require a book-length treatment, but rather to emphasize how each of these causal paradigms captures a part, though not all, of what we mean by the term causality. I depart from existing writing on these subjects by treating these three research paradigms as distinct expressions of causal thinking rather than discrete stages in a research process (measurement and description vs. causal inference).

2 Counterfactuals and Manipulations

While its role in the social sciences was traditionally minor, manipulationalist and counterfactual causal inference has become the go-to reference for understanding causal relations. I consider these two paradigms, though conceptually distinct, to be grouped together as they are both fundamentally discussing the same thing. In brief, the counterfactual inference imagines an alternate world in which the causal factor is not

present, while the manipulationist account emphasizes human (or possibly divine) efforts to force causal factors to take on certain values (Woodward 2003; Morgan and Winship 2007).

While counterfactual causal thinking has been around for some time, it has received its strongest expression in the potential outcomes literature associated with Rubin and Holland. We are told to imagine that a unit i could exist simultaneously in one of two states: a treatment state $Y(1)$ in which i receives a treatment and a control state $Y(0)$ where i does not receive the treatment. Unfortunately, we cannot observe unit i simultaneously in both states, both having and not having the treatment, which is known as the fundamental problem of causal inference. While this fundamental problem would seem to be just that, a fundamental problem, instead Rubin argued that randomly assigning units to receive the treatment $D = 1$ or remain in control $D = 0$ would provide an average estimate of this counterfactual difference:

$$\hat{ATE} = E[Y(1|D = 1) - Y(0|D = 0)]$$

The reasoning is straightforward: if treatments are assigned randomly to subjects, then if we have a sufficiently large subject pool and take the average between the treatment and control groups, any one *individual* counterfactual is equally likely to occur in the treatment or control group. This brilliant formulation is what launched the "credibility revolution" (Angrist and Pischke 2010; Imbens and Rubin 2015) and compels widespread support for RCTs across the social and biomedical sciences.

It is not necessary, of course, to use randomization within a counterfactual or manipulationist theory of inference: it is simply more difficult to know whether an intervention affected the outcome. Pearl (2000) significantly expanded the definition and possibilities of manipulationist inference by his introduction of network relations in the form of directed a-cyclic graphs (DAGs) to stand for causal relationships. Pearl's main insight is that formulations of causal relations based on conditional probabilities alone cannot capture all of what is meant by causality because variables can be connected to each other without having any causal relationship. By providing a very rigorous definition to the notion that correlation and causation are separate phenomenons, he came up with a framework that precisely relates manipulationist ideas of inference to statistical methods of estimating relationships between variables. As I show later, Pearl's framework can extend far beyond manipulationist inference, but because he explicitly defines causality as interventions on causal graphs, I group him in this paradigm.

To summarize Pearl's approach to causality in brief, imagine that there are a set of variables Z, and a variable of interest X, all of which jointly cause an outcome Y. To come up with a directed acyclic graph (DAG), all of the variables that are causal factors must be pointing towards the outcome Y or on some chain to Y, as in

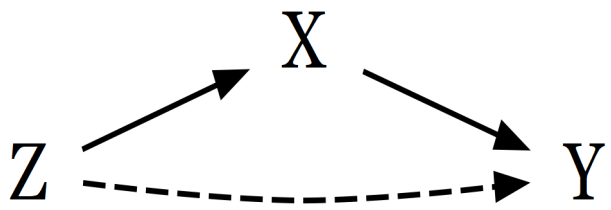

Figure 1: Confounded Example of a Directed Acyclic Graph

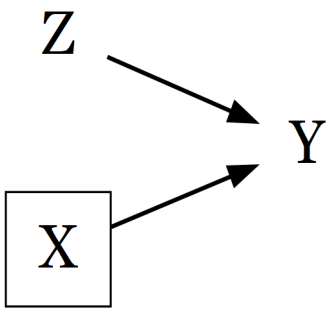

Figure 2: Identified Example of a Directed Acyclic Graph

Figures 1 and 2.

Because all the variables are pointing in the same direction, these graphs are acyclic, or each graph can never return to its origin. This stipulation reflects Judea's view of causality as involving manipulation: in one time period, an action was taken, and in the following time period, a response was observed. In fact, a causal relation is defined explicitly as fixing one of the causal factors (the *do* operator) so that only the manipulated variable affects the outcome, helpfully removing the confounding association between X and Z in Figure 1 as shown by the box around X in Figure 2.

This manipulation–forcing X to a specific value–removes the influence of Z and the "backdoor path" by which changes in Z cause changes in both X and Y. By visualizing these relationships, and providing an algorithm to convert the graph into observable probabilities, Pearl did a great service to causal inference practitioners. Yet even as powerful as his framework is, it does not capture all of what we mean by causality, at least so far as Pearl defines it. Even though it is possible to use these diagrams for analysis other than direct manipulation, as I show later, it is Pearl who restricts the definition of causality to physical interventions on

his networks. While manipulation is certainly a core part of causality, it does not exhaust the subject.

In addition, there have been substantial critiques as of late of RCTs, the strongest form of causal manipulation, as a means of causal inference. While RCTs are not the sole expression of the potential outcomes framework, but the framework does give special preference to RCTs as these represent statistically identified manipulations. As Deaton and Cartwright (2018) point out, the RCT only proposes that observable and unobservable characteristics of the treatment and control groups are balanced *in expectation*. The E operator in Rubin's formula is not harmless. It suggests that only with infinite data will uncertainty in the ATE ever disappear. Unfortunately, the very nature of RCTs sometimes makes generating substantial amounts of data a difficult goal to achieve. Furthermore, the formula assumes a straightforward relationship between the treatment D and the outcome Y. If there are other variables that interact with the treatment, so-called post-treatment variables, then the formula no longer holds. Unfortunately, in the social sciences in particular, these post-treatment issues are very likely to happen as human subjects are free to drop out of studies or to manipulate the treatment to their own ends.

Second, what is also not easily apparent from the formula is that the treatment D is measured without error. In the biomedical sciences with pharmaceutical treatments it is relatively easy to meet this standard, but in many other situations, it is not. Often times social scientists will use some stand in for the actual variable they are interested in, such as a hypothetical situation or a laboratory game to stand in for manipulating real conflict processes. While it is often described as a limitation of "external validity", it is more accurate to state that the problem is one of measurement error (Flake and Fried 2019). Many RCTs in the social sciences do not manipulate X, but some other variable Z that is presumed to correlate with X.

Rather than cast aspersions on RCTs and other forms of counterfactual/manipulationist inference, it is best to think of these methods as providing a very helpful and important component of what is meant by causality. RCTs are also the only means available for addressing unmeasurable selection attributes, such as socially undesirable preferences people may be unwilling to report truthfully. Under ideal conditions, such as with large samples, low attrition rates, and cleanly measured treatments, RCTs provide very strong knowledge about *certain types* of causal relations.

3 Correlation Could Equal Causation

Although at one time it was the dominant approach to causal inference in the sciences, the Humean conception of "constant conjunction" has fallen out of fashion. Hume supposed that we could not know why any two events occurred together and to infer any of this knowledge was a fallacy. Rather, all we could know was

whether events tended to occur together. This correlational theory of inference was codified by Pearson (Pearl 2018, 53–91) and has remained a staple of statistical analysis: checking for associations between variables, or looking for risk factors, as in medicine (Boyko and Alderman 1990; Gershman, Guo, and Dahabreh 2018). While today's statistical education emphasizes that correlation does not equal causation, that tendency has not always been as strong among statistical practitioners.

The reason that correlations still matter even if we do not know the direction of causality or the mechanisms behind the correlation is because causation *does* imply correlation. We may not be able to record the association between two variables due to confounders, but if a causal relationship exists, then we can infer that at some level, somewhere, correlation must be happening. Conversely, if we can prove that two variables are perfectly uncorrelated, we have reasonable evidence that a causal relationship does not exist. Again, this evidence is not a smoking gun or a gold standard as there remains the possibility of a perfectly balanced unobserved variable obscuring the correlation.

For this reason, observational analysis does have a close relationship to causality. When we observe a correlation among human-generated outcomes, we know that a causal process is likely at work, although we cannot rule out that the correlation was due to chance. Statistical methods only give us a way to quantify the uncertainty in estimating the correlation, not in determining whether the association is "truly" causal. But if we abandon the notion that there is a gold standard for causality, then observational analysis, or searching for correlations, still matters for establishing causality even if we cannot learn everything we want to learn.

While some are willing to dismiss traditional statistical methods without randomization or at least manipulation of treatments, entire fields of science are based on purely observational analysis, including astronomy, forensic anthropology, paleontology, and so on. We have never manipulated giant bodies of gas to force them to explode, but we are still fairly sure we know what supernova are (Bethe 1990). The reason for this is that as a silver standard, observational analysis does work if ideal conditions are met, as is true of randomized experiments. In particular, observational methods provide evidence of–though never fully determine–causal relations when we have data on all the variables that could be relevant to the outcome, the data closely match the research question, and we have as much data as we might want.

Researchers also often use syllogisms to interrogate models (Gelman et al. 2013) and reason their way through situations where they cannot collect all the data they want, such as astronomers' disputes over the location and number of planets. If an orbit reflects certain instabilities, then a planet can be hypothesized to exist even if there is no direct evidence for it (Smith 1989).

The denigration of observational models has also led to the rise of quasi-experimental methods. These methods,

including difference-in-difference and regression discontinuity designs, are confusing to define because they make reference to experimental treatments but treatment assignment is never manipulated by the researcher. Instead, these models' true claim to fame is that they have been clearly expressed in Rubin's potential outcomes notation in a way that makes their assumptions easy to intuit. Although the authors of these methods were honest about their practical limitations, practitioners seem to expect that these models are more likely to be "causally identified" than other observational statistical models, which is impossible to know a priori without empirical evidence to justify assumptions.

In fact, these methods are special cases of already well-known statistical models. Difference-in-differences is an interactive fixed effects model with panel data (Kropko and Kubinec 2018). Regression discontinuity design is a form of non-parametric regression where the predictor is evaluated at a single point (Lee 2008). The causal identification conditions for these models are no more or less likely to be true than those for other statistical models without direct randomization of treatments, leading some to argue that these methods have become over-employed and poorly understood (Kahn-Lang and Lang 2018; Caughey and Sekhon 2011; Grimmer et al. 2011).

There are situations in the social sciences where ideal conditions for observational analysis are met, and in these situations, observational models can provide strong causal knowledge without randomized treatments. One oft-cited example is the correlation between smoking and lung cancer. Another, although it is not often expressed this way, is the application of polling aggregation models to predict electoral outcomes (Campbell 2016). While election forecasting models are not perfect and sometimes over or under-predict, the underlying causal process is usually undisputed: if a person is asked who they will vote for the day before the election, they will very likely cast a vote for that person in the ballot box. In this situation, there is plenty of data, we can measure most of what we want to know, and the measures are fairly direct of the underlying outcome.

4 What Lies Beneath: Mechanistic Causation

While the previous two approaches are often contrasted as the only ways of thinking about causality in the sciences, there is a third philosophy that is gaining traction among qualitative researchers, especially in sociology and political science. There has been much work in these disciplines in recent years on establishing how case study research can identify mechanisms that link causal variables (Hedström and Ylikoski 2010). Again, instead of considering this line of inquiry to be a subsidiary issue to the question of causality, I propose that mechanisms are a part of what we mean by causality, and hence are their own silver standard. When ideal conditions are met, we can infer causality with reasonable, though never perfect, confidence.

Several definitions of mechanisms have been proposed in the literature (Mahoney 2012; Gerring 2017). I use for this paper the standard of Waldner (2014) that mechanisms are invariant processes connecting causal variables to each other. The example he uses is very instructive: the person who discovered the mechanisms through which aspirin relieved heart pressure, Sir John Vane, received a Nobel Prize in 1982, long after it had been conclusively established that aspirin had these lasting effects. This Nobel prize is puzzling under either the observational or experimental approaches to causality: if the strength of association was indisputable and random assignment had provided an estimate of all possible counterfactuals, then how could this person receive a Nobel prize for an entirely distinct discovery?

The answer is that when humans conceive of causality, we imagine there to be some kind of process linking cause and effect. Defining exactly what this is process is can be difficult, but the invariant standard is a helpful step. We are looking for processes that are so low-level that they operate similarly in all contexts. In a social scientific setting, we might think of basic emotions like fear, anger and happiness (Pearlman 2013), or the rational actor model (Elster 1994).

Waldner's theory of causation is particularly useful in this paper as it directly compares the mechanistic thinking of causation to Pearl's use of network diagrams for causal structures. To integrate mechanisms, we can introduce labels on the edges that identify which mechanism is in operation. Importantly, *these mechanisms are not variables*, as causal mediation analysis pre-supposes (Imai, Keele, and Tingley 2010). Rather, they are root processes that are at the limit of our observational capacity and are so fundamental as to be nearly determinate. Mechanistic causation is a distinct facet or dimension of causation along with observational and manipulationist inference.

In addition, like the other two silver standards, under ideal conditions it can provide strong evidence of causality. In the social sciences this entails collecting exhaustive evidence on a person's decision-making in what has come to be called process-tracing. Process-tracing methodologists refer to the idea of a "smoking gun" as the kind of evidence that establishes causality within this framework, such as obtaining a private diary of an important leader that describes in detail why they made their decisions (Evera 1997; Collier, Brady, and Seawright 2010; Bennett 2014; Humphreys and Jacobs 2015). These ideal conditions are relatively rare, but like experimental and observational approaches, we can have confidence in inferring *a kind* of causality when we have all the information we might want about how X changed into Y.

For example, imagine if we wanted to know whether the U.S. bombing of Dresden influenced Hitler's strategies during World War II. In a qualitative framework, we could obtain a smoking gun if we had a private diary by Hitler's own hand that described verbatim his thoughts on the U.S. bombing of Dresden and whether or not

he shifted war tactics in response to the bombing. If we then also had detailed evidence, such as transcripts of meetings and official orders, carrying out these changes in tactics prescribed in the journals, then we could state with a high degree of confidence that we know that the bombing of Dresden had a *causal* impact on Hitler's decision-making.

Of course, qualitative inference does lend itself to single-act causation as opposed to establishing general trends between variables. However, to the extent that the variables under study are the same across conditions, then we could presume that our knowledge of the mechanisms linking these variables is similarly invariant. If we can establish what the mechanisms are then we can be more confident that the variables are causally associated in addition to any relationship we estimate using observational or experimental statistics.

5 Synthesis

> "Of course it is happening inside your head, Harry, but why on earth should that mean that it is not real?"
>
> – Professor Albus Dumbledore from J.K. Rowling's *Harry Potter and the Deathly Hallows*

The point of presenting each of these ways of inferring causality is to promote a realist conceptualization of causal inference for the social sciences. It could well be that there exists one true definition of causality yet to be discovered, but even this cursory reading of research and theory suggests there is much more to the concept than we can contain within one mathematical or empirical framework. Rather than dismiss any one of these research conceptions, my intention is to show how each of them captures a distinct part of what humans mean by causality. Causality may have no true definition apart from what we choose to give it, and as extensive research has shown, it is not easy to conclusively identify the way humans reason with any of the theories presented. In an exhaustive summary, Sloman and Lagnado (2015) argue based on extensive experimental evidence that while the network perspective of Pearl solved many issues in formalizing causal thinking, it still "has not offered a silver bullet that answers all questions about human thought" (p. 240). What causality exactly means is a subject for debate, even if we have made remarkable strides in terms of precisely detailing certain causal logics in the past three decades.

This latent source of uncertainty over causality is never discussed in the papers cited above regarding the study of causal processes. Rather, frameworks are presented as a way of defining causality, but over time the framework becomes synonymous with the term. In common parlance in the social sciences, a "causal inference" model implies either a randomized control trial or the usage of certain statistical methods that can be expressed using Rubin's potential outcomes notation. This slippage in terminology puts the cart before

the horse: causality does not inherit from counterfactuals; rather, counterfactuals arose as a way of expressing what is meant by causality.

This realist point is not made to belittle or dismiss the voluminous research on the science of philosophy, causation and statistical science. Over time we have developed new ideas and frameworks to express causality, which have clarified issues and reduced our latent uncertainty. Judea Pearl's work, for example, developed a novel way to express causal relations that unified previously disparate strands in causal thinking. Furthermore, cross-fertilization of ideas can create new methods of analysis, such as the application of structural causal models and counterfactual theories of inference to traditional observational spatial (Egami 2018), time-series (Imai and Kim 2016) and even qualitative process-tracing (Humphreys and Jacobs 2015). Over time, we have learned more about what we mean by the word causal, and continued research should help us more astutely understand the inferences we make and the reasons why we make them.

Nonetheless, causality will remain a latent concept: we cannot observe it, and our uncertainty over the term will never completely go away. Permitting this remaining uncertainty, and allowing the use of the word "causal" in more settings than just randomized control trials or potential outcome models, will help avoid mis-perceptions and researcher frustration. Ultimately, by understanding that we are trying to find the answer to a question we cannot fully form, we should be a bit more charitable in assessing our incomplete yet very important frameworks for doing research.

5.1 Entropy as a Neutral Standard

Rather than end with a paean to methodological pluralism, this paper also proposes a framework for critically thinking about the role and suitability of applying these causal models to real research questions. To do so I present the concept of entropy as a holistic metaphor for understanding our uncertainty about causal relations. Entropy has been widely used to describe physical phenomena such as heat transfer, although social scientists are more familiar with statistical or Shannon entropy, which is also the way I employ the term. In general, entropy describes the decay of a system, such as gas molecules moving farther and farther apart to fill a sphere. Statistical entropy applies the same concept to probability, providing a measure of the "information" in a random variable (Shannon 1948). In general, as probability distribution becomes more equal or uniform, entropy increases because all outcomes are equally likely, whereas when a probability distribution becomes more degenerate or peaked, entropy decreases as some outcomes are more certain than others.

Shannon entropy S is defined as a simple formula for any distribution of probabilities that sum to 1:

$$S = -\sum_{n=1}^{N} p_n \log p_n$$

The formula is unfortunately not intuitive, but its appeal is in meeting certain qualifications for determining an entropy measure of probability, including that it increases as it moves away from neutral probabilities (such as $\frac{1}{N}$) and reaches a minimum at 0 when any of the probabilities is equal to 1. The units of entropy are determined by the type of logarithm employed. Because I am interested in entropy as a framework rather than with a particular empirical application, I use an unconventional logarithmic base of 1.01:

$$S = -\sum_{n=1}^{N} p_n \log_{1.01} p_n$$

A base of 1.01 means that every unit increase in entropy equals a one percent increase in entropy. Figure 3 plots entropy calculations for probability distributions with varying levels of total uncertainty or spreadout-ness. What is important to note is that all of these distributions have the same expected, or average, value, but are nonetheless very different statements about underlying uncertainty.[2] Roughly speaking, the uniform distribution has 100 percent more entropy than the normal distribution, which has 30 percent more entropy than the student's T and Laplace distributions. These plots show why entropy is a powerful heuristic: it captures our sense of how certain we are of the empirical possibilities underlying a distribution of probability that is independent of the form of the distribution.

While entropy has been applied successfully to many statistical problems, my intention in defining it here is to think of it as a way to understand the relative value of the causal paradigms previously discussed. Ultimately, the goal of the social sciences should be to reduce entropy whenever possible in terms of our understanding of how the social world operates. To do so, we have to produce new propositions that explain human behavior and allow us to make judgments about what is more or less likely to occur in the future.

A principle that summarizes this endeavor is that of maximum entropy. Jaynes (2003) defines the maximum entropy principle as always preferring a distribution of higher entropy conditional on including all known facts in the distribution. For example, suppose we wanted to predict stock market prices. Lacking any special knowledge into stock prices, we would want our uncertainty to reflect the fact that all we have to analyze are the movements of individual stocks over time–we would want to maximize entropy, or uncertainty, given the data we have. But if we knew that the Federal Reserve intended to increase interest rates, we could include that information in our model and consequently produce a lower entropy distribution.

2. Because these are continuous distributions and entropy is a measure of discrete random variables, the continuous variates were first binned and then converted to probabilities.

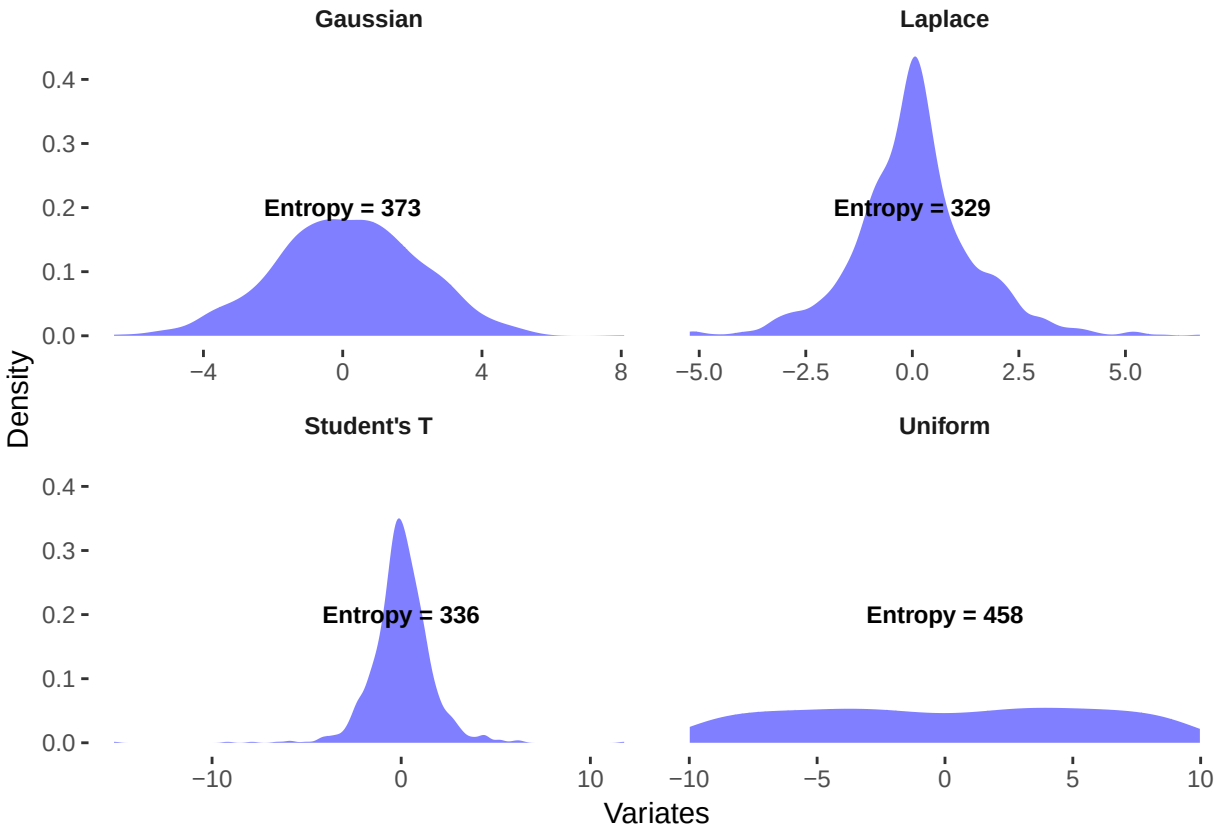

Figure 3: Entropy Calculations Based on Empirical Densities of Statistical Distributions

In other words, we want to learn new facts about the world such that we reduce our entropy in understanding causal relations. At the same time, we want to maximize entropy given what we know to reduce blind spots and over-confidence. Causal inference involves striking this delicate balance between assuming too much and assuming too little.

This framework helps resolve some inconsistencies in how models are incorporated in the social sciences. On the one hand, more complex models are seen as more sophisticated and thus more likely to be true (Clarke and Primo 2007). On the other hand, there has been considerable push back at models that appear to be baroque and less easy to explain than the beloved ordinary least squares (OLS) regression (Angrist and Pischke 2008; Dunning 2012). Maximum entropy helps explain these mixed feelings: we should prefer more complex models over simple models because our over-arching aim should be to reduce entropy, and more complex models have less entropy. On the other hand, we do not want to reduce entropy without a good reason lest we over-state our certainty (Frank 2009).

To use entropy to understand causal paradigms, I return to Pearl's causal diagrams. My intention is not to suggest that Pearl's theory is the final take on causality, but rather that network relations are deeply intuitive representations of human thought patterns. As such, they are a helpful starting block for comparing very different representations of causality.

Let us imagine that we launched ourselves in a spaceship and landed in a foreign world. We have almost no prior knowledge of how people on this planet relate to each other, but we want to understand how the world's residents select their leaders. All we can do is come up with a list of plausible factors that might affect leader selection. Given our experience of such occurrences on earth, we come up with the following list of variables: political ideology (I), economic benefits (E), ethnic affinity (A), leader personal qualities (Q), and the risk of conflict (C). We can think of these variables as nodes in a network all connected with the outcome of leader selection (Y) as is shown in Figure 4.

Each edge in this graph is labeled with the dual expected probabilities that a link exists in either direction, with each one labeled $\frac{1}{36}$ to represent our current ignorance, i.e., any link between any of the nodes is equally likely in either causal direction. While the uncertainty in this figure is extreme, it comes closer to the actual state of social science research, where substantial uncertainty exists over even defining the space within which causality happens.

Given the previous discussion, the question now is to reason about which method of causal inference to apply to Figure 4. The easiest way to answer this question, and one often chosen, is simply to choose whichever method best fits the researcher's skills and experiences. While very practical, it poses a chicken-and-egg

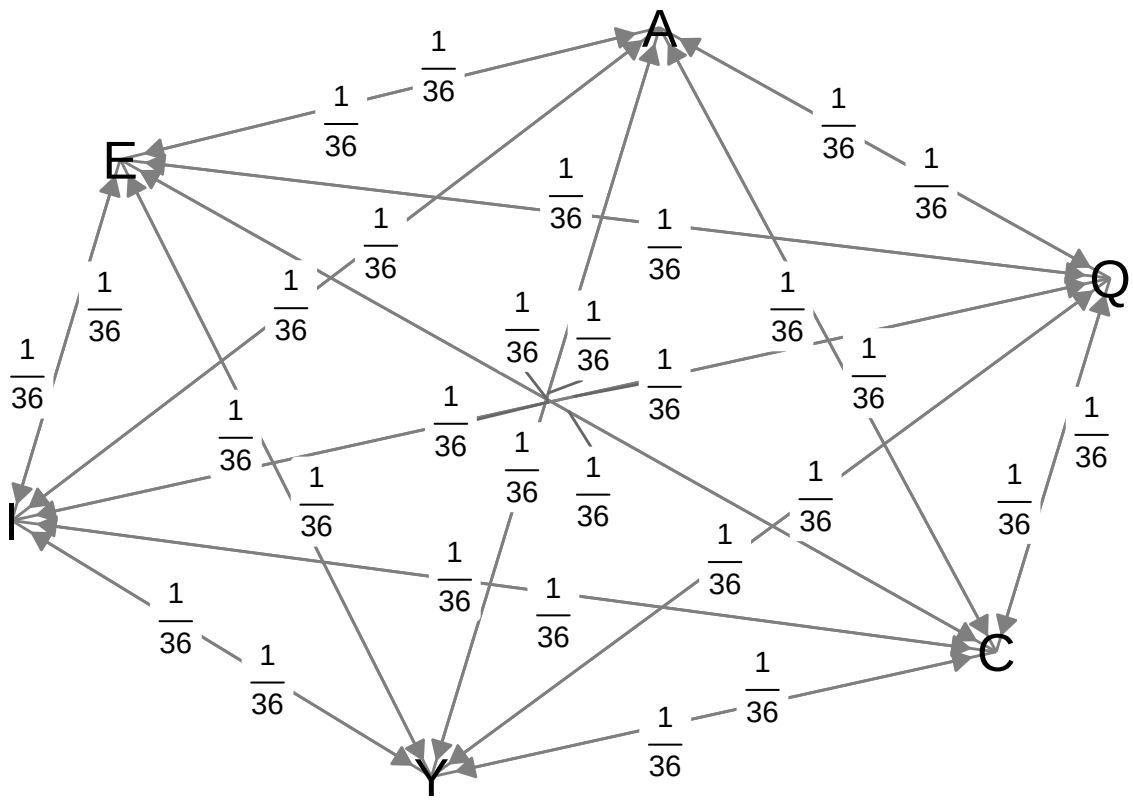

Figure 4: Causal Diagram with Complete Uncertainty

problem, and only shifts the question to which paradigm researchers should invest in to gain experience and skills.

I propose that a better heuristic is to ask what would reduce entropy in the causal graph. There are very complicated entropy statistics for graphs, but to simplify matters I apply the entropy formula to each edge probability in Figure 4:

$$-\sum_{1}^{36}\frac{1}{36}\log_{1.01}\frac{1}{36}=360$$

We start with the considerably high number of 360. At this amount of entropy, we are not likely to be wrong, but we also cannot say much of value about our study of this foreign planet's society. Initially, let us consider a choice between an experimental and an observational analysis. Suppose than with an experiment we can determine directly the probability of the connection between ethnic affinity (A) and leader selection (Y).[3] If we pull off a quality experiment, we can double our confidence in the link from A to Y and change our assessment of the link from Y to A (reverse causality) to 0. We can consider the experiment to be *causally identified* because it is possible, given enough experimental data, to either rule out or establish the relationship between these two nodes (Keele 2015). Then we can re-calculate our entropy measure, which shows a decrease in entropy of 4 percent:

$$-[(\sum_{1}^{34}\frac{1}{36}\log_{1.01}\frac{1}{36})+\frac{2}{36}\log_{1.01}\frac{2}{36}]=356$$

However, suppose that if we conducted an observational data analysis, we could increase or decrease the probability of the links between leader selection and economic benefits (E), leadership quality (Q), ideology (I), ethnic affinity (A) and conflict (C) from $\frac{1}{36}$ to $\frac{1.5}{36}$ while lowering the opposite links to $\frac{0.5}{36}$. This research design is not causally identified because we cannot be sure, i.e. we cannot know with probabilities approaching one, what effect these variables have on the outcome. This analysis would result in the following change in entropy:

$$-[\sum_{1}^{26}\frac{1}{36}\log_{1.01}\frac{1}{36}+\sum_{1}^{5}\frac{1.5}{36}\log_{1.01}\frac{1.5}{36}+\sum_{1}^{5}\frac{0.5}{36}\log_{1.01}\frac{0.5}{36}]=356$$

In other words, in this toy example, the observational and experimental studies would have similar effects on reducing the entropy of the total system *even though they made very different statements about the underlying*

3. We ignore for the time being the difficulty in enacting these research designs.

causal structure. Intuitively, we can learn a lot from establishing a specific causal link in a specific direction with a high degree of certainty, but we can also learn a lot from examining associations between variables, even if we cannot arrive at conclusive predictions. The point of this exercise is not to suggest that observational methods are better than experimental methods, but rather that the value of each depends on the nature of the causal problem, and it is *not* always the case that experiments produce more causal knowledge than observational studies.

We can further extend this discussion by considering the mechanisms implied by the edges. Instead of treating the edges as solely representing causal relationships, we could imagine a distribution over mechanisms for each edge. We could similarly decrease our entropy over mechanisms by learning which mechanisms are the most likely and least likely for a given node. Uncertainty over mechanisms could then be weighted with the overall association probability as entropy is additive on the log scale.

The intention of this exercise is to show that the concept of statistical entropy sheds light on the difficult decisions that must be made when considering research designs. Rather than propose a single causal paradigm as the most important, the criterion of reducing entropy suggests that we aim for maximizing the amount we can learn from an application of any of the paradigms. To illustrate the principle, I apply the theory to actual areas of research.

This entropy principle generalizes the common intuition that researchers should look for "low-hanging fruit" when examining a given research field. Practically, if a certain method has been employed exhaustively, it would probably be prudent to switch to another type of analysis. First, consider the outcome of the massive number of experiments performed on elections, canvassing and voter behavior in American politics. Kalla and Broockman (2018) recently documented how 49 field experiments produced an average null effect of campaign advertisements and mailings on voter behavior. This finding is of course puzzling as campaigns spend a large amount of money to do what this study says is on average unlikely to occur: affect voter choices in an election.

This study is fascinating as it represents a field where the application of experiments has reached a far greater depth than other areas of political science. In this case, it would seem that we could reduce entropy to a greater degree in the study of voter behavior by looking at other methods than experiments, either observational large-N analysis or qualitative research aimed at identifying subsets of voters for interviews. Part of the problem, as the authors of the study note, is that there is widespread treatment heterogeneity, which is shorthand for a large number of background factors that also affect the success of the treatment. Observational and qualitative analysis could help uncover what these background factors are, such as broad

geographical, economic or cultural factors, or very specific group or individual-level mechanisms that the treatment is interacting with. In other words, observational analysis could help situate the experimental findings within a broader theory as some have called for in psychology (Muthukrishna and Henrich 2019).

Conversely, we can consider a situation where observational and qualitative analysis has been remarkably widespread: the study of democratization. Coppedge (2012) documents the wide variety of statistical models fit to ever-increasing datasets, including most recently the Varieties of Democracy of Project (Coppedge et al. 2017). In addition, countless country-level case studies of democratic processes exist in the scholarly literature dating back to the origins of comparative political science (B. Moore 1966). Yet there are relatively few if any experiments on building democratic institutions and on encouraging support for democracy as opposed to authoritarian politics, which suggests that entropy could be further reduced in this field by considering more experimental approaches.

6 Conclusion

Ongoing debates about causal inference threaten to create divides that impede research progress. Part of the problem is the growing assumption that causal inference requires an RCT or at minimum a model expressed in terms of potential outcomes. This divide separates research into causal and "mere" association, with the former preferred over the latter without reference to the relative amount of information to be gained.

Rather than point to problems with RCTs as a reason to distrust them, I aver that the underlying issue is that we lack a non-tautological definition of causality. As a result, it is not that RCTs have more issues than are commonly acknowledged, but rather that the definition of RCT as the gold standard means that we artificially deflate the value of other modes of causal analysis. Without a crystal clear yardstick for deciding precisely what causality means, we should take a charitable approach by admitting that various theories and practices of inference contain some, though not all, information about causal relations.

I identify the three *silver* standards of causal analysis as correlational, counterfactual and mechanistic inference. This typology, while being remarkably simplistic, is intended as a heuristic for researchers looking for the "right way" to conduct a research design. The principle of entropy provides one helpful framework by imagining the benefit of a study from the relative reduction in entropy it achieves. While in the future it may be possible that we resolve all remaining uncertainty in causal thinking, the more ambiguous present demands an ecumenical approach to permit free-ranging scientific inquiry even if it also permits a degree of cognitive dissonance.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

———. 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.

Arkhangelsky, Dmitry, and Guido W. Imbens. 2018. "The Role of the Propensity Score in Fixed Effects Models." *ArXiv*. https://arxiv.org/pdf/1807.02099.pdf.

Beck, Nathaniel. 2006. "Is Causal-Process Tracing an Oxymoron?" *Political Analysis* 14 (3): 347–52.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2017. "Redefine Statistical Significance." *Nature Human Behavior* 2: 6–10.

Bennett, Andrew. 2014. "Disciplining Our Conjectures: Systematizing Process Tracing with Bayesian Analysis." In, edited by Andrew Bennet and Jeffrey T. Checkel, 276–98. Cambridge Univ Press: Cambridge, UK.

Bethe, H. A. 1990. "Supernova Mechanisms." *Reviews of Modern Physics* 62 (4): 801–66.

Boyko, Edward J., and Beth W. Alderman. 1990. "The Use of Risk Factors in Medical Diagnosis: Opportunities and Cautions." *Journal of Clinical Epidemiology* 43 (9): 851–58.

Campbell, James E. 2016. "Forecasting the 2016 American National Elections." *PS: Political Science & Politics* 49 (4): 649–54.

Caughey, Devin, and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008." *Political Analysis* 19 (4): 385–408.

Clarke, Kevin A., and David M. Primo. 2007. "Modernizing Political Science: A Model-Based Approach." *Perspectives on Politics* 5 (4): 741–53.

Collier, David, Henry E. Brady, and Jason Seawright. 2010. "Toward an Alternative View of Methodology: Sources of Leverage in Causal Inference." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited by Henry E. Brady and David Collier. Lanham, MD: Rowman & Littlefield.

Coppedge, Michael. 2012. *Democratization and Research Methods*. Cambridge University Press.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, et al. 2017. "V-Dem Dataset V7." Working Paper. Social Science Research Network.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2968289.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Control Trials." *Social Science & Medicine* 210: 2–21.

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

Egami, Naoki. 2018. "Identification of Causal Diffusion Effects Using Stationary Causal Directed Acyclic Graphs." *Archiv*. https://arxiv.org/abs/1810.07858.

Elster, Jon. 1994. "The Nature and Scope of Rational-Choice Explanation." In *Readings in the Philosophy of Social Science*, edited by Michael Martin and Lee C. McIntyre, 311–22. Massachusetts Institute of Technology: Boston, MA.

Evera, Stephen Van. 1997. *Guide to Method for Students of Political Science*. Ithaca: Cornell University Press.

Fisher, R. A. 1935. *The Design of Experiments*. Oxford, England: Oliver; Boyd.

Flake, Jessica, and Eiko Fried. 2019. "Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them." *PsyArXiv*. doi:10.31234/osf.io/hs7wm.

Frank, Steven A. 2009. "The Common Patterns of Nature." *Journal of Evolutionary Biology* 22 (8): 1563–85.

Gelman, Andrew. 2011. "Causality and Statistical Learning." *American Journal of Statistics* 117 (3): 955–66.

———. 2018. "Benefits and Limitations of Randomized Control Trials: A Commentary on Deaton and Cartwright." *Social Science & Medicine* 210: 48–49.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "P-Hacking" and the Research Hypothesis Was Posited Ahead of Time." http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. CRC Press.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2014. "The Illusion of Learning from Observational Research." In *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in*

the Social Sciences, edited by Dawn Langan Teele. Yale University.

Gerring, John. 2017. "Qualitative Methods." *Annual Review of Political Science* 20: 15–36.

Gershman, Boris, David P. Guo, and Issa J. Dahabreh. 2018. "Using Observational Data for Personalized Medicine When Clinical Trial Evidence Is Limited." *Fertility and Sterility* 109 (6): 946–51.

Gerstein, Hertzel G., John McMurray, and Rury R. Holman. 2019. "Real-World Studies No Substitute for Rcts in Establishing Efficacy." *The Lancet* 393 (10168): 210–11.

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. 2017. "Broken or Fixed Effects?" http://www.jcsuarez.com/Files/Suarez_Serrato-BFE.pdf.

Goodman, Steven N., Daniele Fanneli, and John P. A. Ioannidis. 2016. "What Does Research Reproducibility Really Mean?" *Science Translational Medecine* 8 (341): 341–47.

Green, Donald P., and Alan S. Gerber. 2003. "Reclaiming the Experimental Tradition in Political Science." In *Political Science: The State of the Discipline*, edited by Ira Katznelson and Helen V. Milner, 805–33.

Grimmer, Justin, Eitan Hersh, Brian Feinstein, and Daniel Carpenter. 2011. "Are Close Elections Random?" Stanford University: Working Paper.

Hedström, Peter, and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36: 49–67.

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

Humphreys, Macartan, and Alan Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109 (4): 653–73.

Imai, Kosuke, and In Song Kim. 2016. "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Online.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15 (4).

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society Series A* 171 (2): 481–502.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical*

Sciences. Cambridge University Press.

Ioannidis, John P. A. 2018. "Randomized Control Trials: Often Flawed, Mostly Useless, Clearly Indispensable: A Commentary on Deaton and Cartwright." *Social Science & Medicine* 210: 53–56.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Kahn-Lang, Ariella, and Kevin Lang. 2018. "The Promise and Pitfalls of Differences-in-Differences: Reflections on '16 and Pregnant' and Other Applications." NBER Working Paper Series. National Bureau of Economic Research. https://www.nber.org/papers/w24857.pdf.

Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112 (1): 148–66.

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

Kropko, Jonathan, and Robert Kubinec. 2018. "Why the Two-Way Fixed Effects Model Is Difficult to Interpret, and What to Do About It." *Social Science Research Network*, February.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142: 675–97.

Levitt, Steven D., and John A. List. 2008. "Field Experiments in Economics: The Past, the Present, and the Future." Working Paper No. 14356. National Bureau of Economic Research. https://www.nber.org/papers/w14356.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods & Research* 41 (4): 570–97.

Moore, Barrington. 1966. *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Beacon Press: Boston.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Muthukrishna, Michael, and Joseph Henrich. 2019. "A Problem in Theory." *Nature Human Behavior*.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115 (11):

2600–2606.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science*.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

———. 2018. *The Book of Why: The New Science of Cause and Effect*. Penguin RandomHouse UK.

Pearlman, Wendy. 2013. "Emotions and the Microfoundations of the Arab Uprisings." *Perspectives on Politics* 11 (02). Cambridge Univ Press: 387–409.

Plümper, Thomas, and Vera Troeger. 2019. "Not so Harmless After All: The Fixed-Effects Model." *Political Analysis* 27 (1): 21–45.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66: 688–701.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *The Journal of Politics* 78 (3): 941–55.

Sampson, Robert J. 2018. "After the Experimental Turn: A Commentary on Deaton and Cartwright." *Social Science & Medicine* 210: 67–69.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423.

Sloman, Steven A., and David Lagnado. 2015. "Causality in Thought." *Annual Review of Psychology* 66: 223–47.

Smith, Robert W. 1989. "The Cambridge Network in Action: The Discovery of Neptune." *Isis* 80 (3): 395–422.

Waldner, David. 2014. "What Makes Process Tracing Good? Causal Mechanisms, Causal Inference, and the Completeness Standard in Comparative Politics." In *Process Tracing: From Metaphor to Analytic Tool*, edited by Andrew Bennet and Jeffrey T. Checkel, 126–52.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.