What Good is a Regression?

Inference to the Best Explanation and the Practice of Political Science Research*

Arthur Spirling† Brandon M. Stewart‡

Abstract

We consider the claim that political science research has become too focused on causal inference at the expense of substantive concerns. We contend that debates on this question are unproductive because they fail to recognize that almost all empirical social science research uses "Inference to the Best Explanation" (IBE) and must be assessed in that light. Noting its wide acceptance and use elsewhere, we review the basic principles of IBE, and examine its ubiquity in our discipline. We show that disputes regarding the merits of different approaches can be reduced to debates over the relative weight scholars place on quality of inference versus quality of explanation. We argue that many pathologies of current practice can be both explained and potentially ameliorated this way. These include incentives to p-hack, the undervaluing of description, and the confusion over the value of non-plausibly causally identified studies. Explicitly embracing IBE helps structure discussions in the discipline.

*First version: March 25, 2022. This version: April 6, 2022. Keith Dowding and Kevin Munger provided very helpful comments on an earlier draft.

†Professor of Politics and Data Science, New York University (arthur.spirling@nyu.edu)

‡Assistant Professor of Sociology and Office of Population Research at Princeton University (bms4@princeton.edu)

1 Introduction

The use of regression is ubiquitous in social science research. Despite this dominant status—or perhaps because of it—what one can learn from a given regression has become increasingly contested. This is partly a question of whether a given association can be plausibly claimed to be causal (e.g. Samii, 2016), but it is much broader than that. It includes the appropriate role that regression might play in suggesting explanations (e.g. Huber, 2013), and in testing them (e.g. Ashworth, Berry and de Mesquita, 2021). This paper is about providing a framework for understanding, and potentially resolving the differences implied by, these seemingly opposed positions. That framework, known as "abductive inference" in general, and "Inference to the Best Explanation" (IBE) in particular, is by no means novel to us (it has been known in some form since at least Peirce, 1878). But we believe it is almost entirely unknown to scholars in the discipline (though see Dowding and Miller 2019 in political science and Tavory and Timmermans 2014 in sociology). This is surprising because, as we will argue, almost all published work is doing some version of the IBE process—including qualitative studies. This matters: it means that much of the controversy around the use of regression, and indeed quantitative work more generally, is missing an interrogation of the underlying inferential framework on which the work is built. In particular, our contention is that in making their claims and counter-claims, researchers are in fact advocating for the importance of different elements of the *same* mode of inference. With this is mind, they have much more in common than they initially realize. In short then, the answer to "what good is a regression?" depends on what element of that inferential framework you value most. And while there may be good reasons to value some particular part more than others, researchers should make that case explicitly when critiquing other approaches.

Inference to the Best Explanation (IBE)—a phrase coined in Harman (1965)—is a mode of inference that describes everyday reasoning. Given a set of facts, we infer that out of a set

of possible explanations, the one that 'best' explains the evidence is the one most likely to be true. The best explanation might be one that is simpler, more complete, or possessing some other desirable property. As a form of abductive inference, IBE yields plausible conclusions—not ones that are logical consequences of a set of premises as in deductive inference. We are most confident in IBE when the candidate set of explanations are *ex ante* plausible and the evidence discriminates well between them. Because we cannot do deduction except in special cases such as formal theory, and inductive reasoning (strictly defined) is rare, IBE is essentially how arguments always proceed. Social scientists will recognize it most clearly in articles and books where the authors lay out multiple theories (explanations) and adjudicate between with them with a variety of regression-based tests (facts/evidence). In practice this is not just a difference of terminology; it emphasizes that the *candidate explanations* have as important a role in credibility of the inference as the evidence; credible estimation of a descriptive fact or causal query is only one piece of the larger inferential framework. Any one fact is consistent with countless possible stories, so we are always—even if only implicitly—adjudicating between competing explanations.

Our claims are general and broad, and in keeping with that, we use 'regression' as something of a catch-all. Concretely, we have in mind a parametric or non-parametric model in which there is an outcome (dependent variable, Y) that is a function of at least one predictor (independent variable, X). That is, we are interested in results from a very wide range of techniques which approximate a *conditional expectation function* of the form $\mathbb{E}(Y|X)$ where $\mathbb{E}$ is the expectations operator. This includes linear regression as a special case, but also subsumes generalized linear models, and various techniques traditionally deemed part of "machine learning" (e.g. random forests).[1]

In Section 3 we will provide a succinct definition of IBE, explaining what it is in abstract

[1]See e.g. Aronow and Miller (2019) for a textbook account of standard linear regression as an approximation of an unknown conditional expectation function.

terms before showing how it is used (knowingly or not) in the theory and practice of political science research. In Section 4 we explain why this matters. In particular, we show that many concerns about the practice of quantitative research can be understood as either differences in opinion about what part of IBE is most important, or unfortunate consequences of a general lack of understanding about what role regression can and does play in the abductive inference framework. This includes the danger of HARKing (Kerr, 1998) and p-hacking (Franco, Malhotra and Simonovits, 2014), claims of "big" vs "small" research questions, and the potentially marginalized role of description (Gerring, 2012; Munger, Guess and Hargittai, 2021). Of course, we make no claims that acknowledging the role of IBE will solve all problems, and we are explicit in that section about currently open issues that would benefit from more considered attention in this framework. We conclude in Section 5 with thoughts about what the future of IBE in political science might look like. Before all of this, we first review current debates in the discipline over what quantitative methods can and ought to be used for.

2 The State We are In: Current Methodological Debates in Political Science

To fix ideas for what follows, consider a common scenario. An author presents a regression table and discusses the "effects" of a given (independent) variable on an outcome of interest. The data is observational, meaning that assignment to treatment and control is not within the power of the researcher. From a causal inference perspective, one of the variables X_j is a treatment (in the sense of e.g. Gelman and Hill, 2006), but the author does not necessarily use that term. In addition, the author controls for (conditions on) various other variables in the regression (X_{-j}), but does not provide particularly clear assertions about the nature of the assumed confounding, any potential post-treatment bias, the plausibility of conditional

ignorability in this case, or the implied Directed Acyclic Graph (in the sense of Keele, Stevenson and Elwert, 2020). Over various specifications, the relationship between X_j and Y is consistent, insofar as the sign and direction of a coefficient (or risk ratio, first difference etc.) remains similar. Put crudely: what is this regression good for, and how might we assess its merits? In particular, how might we judge the worth of this regression (or these regressions) relative to one where, conditioned on confounders, the treatment is plausibly as-if randomized to units? Suppose in addition, that in this latter case only one potential explanation for the data-generating process is offered. Assume that in neither case is prediction (in the sense of Cranmer and Desmarais, 2017) the goal. These abstractions are perhaps extreme, but capture the spirit of many examples in modern published works.

In a thin sense, there is no dispute about what a regression does in either case: it approximates a conditional expectation function. But when, and in what ways, that conditional expectation function is useful is the subject of considerable debate. This is not simply a matter of disagreement over what assumptions are plausible when; it is, in practice, a question of competing visions about what the work of the discipline should be and where effort is most profitably applied.

Although a relatively recent development in the history of social science, the "credibility revolution" (Angrist and Pischke, 2010) is a natural place to begin this discussion. The central idea here is that making causal statements is difficult with observational data, and can only be done in a more limited set of circumstances than may be initially realized. Indeed, per Samii (2016) (see also e.g. Gelman and Hill, 2006; Gerber et al., 2014; Keele, Stevenson and Elwert, 2020), regressions without thought about these issues may be actively misleading. Consequently, scholars must search for a "strong design" in order to make "persuasive" causal claims (Sekhon, 2009, 503). This is hard to achieve even in seemingly propitious circumstances where, for example, treatment and control may be randomized but the groups thus created are not comparable (e.g. Sekhon and Titiunik, 2012).

While the technical claims of these scholars are not in doubt, there has been much disagreement about what the credibility revolution means, or should mean, for the focus of political science research in general. For some, including Huber (2013; 2017), the consequences are potentially baleful. They argue that in discouraging the running of non-causally identified regressions, the revolution rules out the study of certain types of questions not amenable to those techniques and assumptions (see also Clark and Golder, 2015; Binder, 2020). A related but distinct concern comes from those who contend that causal empiricists and formal theorists are "pulling apart", when they should be cooperating and learning from each other (Ashworth, Berry and de Mesquita, 2021).

For others, the priority is not producing causal claims (of whatever plausibility), but description. Thus we see work by Gerring (2012) that emphasizes the importance of the descriptive task *per se* and independent of theory-testing. Indeed, scholars have proposed entire journals to help counter the fact that "[c]ausal research that asks the question *why* has largely taken the place of descriptive research that asks the question *what*" (Munger, Guess and Hargittai, 2021, 3, emphasis as original). Here then, regressions are informative about the state of the world in terms of associations—nothing more and nothing less. Partly in an attempt to connect this associational logic to the goal of inference, researchers have recently argued that flexible machine learning approaches—capable of including non-linear interactions—ought to be more broadly deployed for political science tasks (e.g. Montgomery and Olivella, 2018). Whatever the model, the fundamental challenge is that the associations are all highly conditional (that is, conditional on many things). But this can make interpreting them—in terms of an all-else equal logic—difficult and counter-intuitive (Ashworth, Berry and de Mesquita, 2021).

Whatever one's priority for the purpose of regressions, there has been increasing agreement on what their properties ought to be. In particular, the importance of replication and robustness in results. At one level, the concerns regard the potentially malign motivations

of researchers to "p-hack" or else leave insignificant results in the "file-drawer" (e.g. Franco, Malhotra and Simonovits, 2014); for others, there are broader issues of "forking paths" wherein researchers make ad-hoc but crucial decisions about data and estimation (Gelman and Loken, 2014, e.g.). Scholars have proposed various solutions, from "multiverse analysis" of all possible choices (Steegen et al., 2016) to more focused efforts at assessing sensitivity (e.g. Imai and Yamamoto, 2010; Blackwell, 2014). Related in spirit, but different in practice, other authors have suggested methods for incorporating distributional assumptions about bias (as in, the difference between an estimated coefficient and the 'true' causal effect) into more nuanced interpretations of regression results (Little and Pepinsky, 2021).

Of course, the logic thus far assumes that authors are sufficiently clear about what they are estimating to effectively connect the results to explanations or broader theory in a credible way. Lundberg, Johnson and Stewart (2021) point out that this is often not the case; in fact, many authors define their quantity of interest entirely in the context of a parametric model and fail to provide a sense of what it approximates should the model not hold exactly. This leads to situations where debates can be entirely about disconnects over the target estimand (see e.g. the literature in Lundberg, Johnson and Stewart, 2021). An older literature, (e.g. Lieberson and Horwich, 2008) worries that the link between theory and evidence has so substantially frayed that social science is merely 'mimicking' science. A necessary first step in understanding how evidence informs theory is understanding what the author believes the evidence to represent.

The sheer diversity of inferences about and from regression might suggest that the discipline has no unified way to communicate about these issues. At the very least, it would seem that authors have instinctual understandings of what makes their arguments compelling but that those visions have no common framework in which to be placed and assessed "scientifically". Our contention is more positive, however. We claim that a framework already exists for accommodating all of these positions, and that it is Inference to the Best Explanation.

We now define IBE in more detail, before making this point with reference to studies in the field.

3 What is *Inference to the Best Explanation*?

Abductive inference, in the form of IBE, is ubiquitous in scientific enquiry (e.g. Harman, 1965; Boyd, 1984; Douven, 1999; Lipton, 2003). The study of politics is no exception (Dowding and Miller, 2019). Despite the fact that it is believed to capture 'everyday' reasoning, or perhaps because of this, it is not always easy to find precise statements as to what IBE is. Introductory accounts (e.g. Psillos, 2002; Douven, 2021) will typically give a definition along the following lines:

> Given some data D (some observations, or facts about the world), and some candidate explanations or hypotheses $E_1, \ldots, E_n$ that potentially explain D, the one that is most compatible with D is most likely to be true.

Unsurprisingly, there is considerable philosophical debate as to what constitutes an "explanation" *per se* (e.g. Achinstein, 1983). But at this level of abstraction, it is straightforward to give examples of how such a process might work in principle. Suppose we observe that as countries become more developed (say, in per capita income terms) they are generally more likely to become democracies. One explanation might be that of *Modernization Theory* in the sense of Lipset (1959): essentially, that changing social conditions make middle class citizens (in particular) more likely to embrace democratic ideals. An alternative explanation might follow this basic logic, but specify that middle class sympathy is a product of particular social relations some centuries before (Moore et al., 1993). Other explanations might focus on the role of income inequality (e.g. Boix and Stokes, 2003) or elite responses to the threat of revolution (Acemoglu and Robinson, 2001). Which of these explanations is most

plausible is not our central interest here.[2] Our point instead is that this is a near-universal undertaking: scholars observe (essentially) the same data, and attempt to provide a 'best' explanation for this data. And when they do this, they are doing IBE. We can push this point further. When scholars gather *new* data and suggest a 'best' explanation for those observations—relative to other explanations or even simply a null hypothesis—they are also doing IBE. Consequently in empirical social science, almost everyone, all of the time, is doing IBE. We will expand on this idea below, but before doing so we clarify the position and nature of IBE more broadly.

First, abductive inference stands in contrast to both *deduction* and *induction*. If we observe data D, then deduction requires that our inference is a logical consequence of that D. In social science, such reasoning is rare outside of formal theory work where we specify predicates (say, assumptions of a theoretical model) and generally agree on what operations one can undertake on those predicates (say, what makes for an 'equilibrium' in a given game). Logic of this kind extends to the "deductive approach" to causal inference—which requires careful specification of reality and the representation of that reality (Pearl, 2014).[3] But this does not mean that a given *empirical* study using such tools is itself "deductive" in terms of its inferential approach: it cannot be if its conclusions do not follow *with certainty* from its data. In contrast, *induction* does not involve guaranteed conclusions. This is the case even though we may accept the truth of the premise D. In line with our comparative politics example above, an inductive inference might be that a randomly chosen rich country is very likely to be democracy. Unlike for deduction though, we make no claim that this *must* follow from D—we acknowledge that there are a small number of rich non-democracies. In addition, and crucially unlike in abductive inference, induction does not require we offer a

[2]Though obviously political science as a whole has invested considerable effort in this question (see, e.g., Przeworski et al., 2000)

[3]That is, that a causal quantity can be identified as a particular operation on a structural causal model is a deductive conclusion that follows from the assumptions of that structural causal model.

causal 'story' as to why we expect a rich country to be democratic. IBE does require that extra step.

Second, it is important to clarify that however common it is in practice, IBE is not (claimed to be) a perfect strategy for inference. For one thing, IBE contains an obvious logical fallacy: "affirming the consequent". This is straightforward to see from our case above; for example, if *Modernization Theory* is correct, then it follows that we would see a particular pattern of democratization. But seeing that particular pattern cannot be conclusive evidence that *Modernization Theory* is correct (Clarke and Primo, 2012, make a similar point in their discussion of 'models'). For another thing, we have no guarantees that the set of explanations from which we are purportedly selecting the 'best' one contains the truth. Indeed, the fact that political scientists continue to propose new explanations for the development data we observe suggests that the field as a whole is not yet convinced we have at the end point of this search. Finally—and a point that belies the simplicity of abstract definitions of IBE—it is unclear exactly what constitutes a 'best' explanation. It seems reasonable to prefer explanations that have better predictive accuracy (even within sample), but how that is traded off against the qualities of the explanation itself (e.g. its parsimony) may be ambiguous (see, e.g, Barnes, 1995).

Of course, one does not need to believe that a given inference method is perfect (or even coherent—see, e.g., Van Fraassen, 1989; cf Douven, 1999) for it to be popular in practice. And if it is popular, it is important to understand its characteristics and implications.

3.1 IBE in Social Science: Theory

To be more specific about the machinery of IBE in social science, it is helpful to think about how it works in a "classic" case, such as medicine (the discussion of the Semmelweis case in Lipton, 2003, is in this vein). There, a patient presents with symptoms. The doctor conducts a differential diagnosis: that is, the doctor attempts to enumerate several possible

explanations for those symptoms. Typically, tests are performed (including, say, simply examining some aspect of the patient's body) in an effort to rule out some of the possible explanations; in particular, the doctor is showing that the data (say, the blood sample) is inconsistent with the predictions of at least some of those explanations. At the end of this process there is one best explanation: it need not be the only one compatible with the data, but it is the most preferred in some sense.[4]

Generally, the moving parts here are

1. characterization of the symptoms (the data)
2. the enumeration of possible explanations
3. tests that discriminate between explanations that are and are not consistent with the data (or that are more or less consistent with the data)

The order is important: to avoid overfitting, one cannot use the *same* data to both create explanations and test those explanations. Nonetheless, there can be iteration: doctors may update their diagnosis after positing a particular condition, but then testing for it, and finding it absent. But there are limits, as we note below.

In the context of social science, steps (2) and (3) take specific forms. First, our explanations are almost always *causal* in nature (King, Keohane and Verba, 1994; Clarke and Primo, 2012) though exactly how precise they need to be is debatable (Dowding, 2015, Ch3). Second, in quantitative work, it is often the case that only one explanation is offered, and that it is tested against *only* a null hypothesis (Gross, 2015). Of course, the null hypothesis is not itself an explanation for the data (a point that has been made in the framing of the null hypothesis as implausible, as in McShane et al. 2019). So failing to reject the null is potentially awkward in the world of IBE. That is, we have not found evidence consistent with our preferred E_i, but nor have we made an inference to another, 'best', explanation.

[4]See, e.g. Bird (2007) for a discussion of some subtleties here in the Semmelweis case.

This does not mean that the study is of no value. Obviously, as with medicine, it can be important in an individual case to show that a particular explanation is unlikely to be correct (e.g. we ascertain that patient does not have an allergy to shellfish, and can eat shrimp). More optimistically, at a *field* level of surveying many studies regarding similar data, we can presumably claim that a process consistent with IBE is taking place. That is, when we look at a literature at a high level, we are collectively generating new explanations, and making an inference to the best one by rejecting those that are not supported. But some care is required in this weaker, aggregated understanding of IBE. Put crudely, simply generating a large number of bad (unsupported) explanations is not especially helpful.

3.2 IBE in Social Science: Practice

One part of our argument is that everyone is doing inference to the best explanation, whether they know it or not. IBE is ubiquitous because it is often the only thing we can do—no single exercise is going to unambiguously show that we are right (in the way that a deduction would). This in turn makes the three components—the description of the world, the enumeration of possible explanations, and the tests that discriminate those explanations—all essential to the quality of the argument.

Prior work has devoted considerable attention to the specifics of the tests themselves and their credibility. Lundberg, Johnson and Stewart (2021) outline a research framework which connects the *theoretical estimand* (the possibly unobservable quantity of interest), the *empirical estimand* (the observable analog connected through identification assumptions) and the *estimation strategy*. This focuses on the credibility of the chain, but does not specifically tackle the way in which the evidence informs the broader theory or goal. Yet the specific choice of the estimand is a way to adjudicate between different explanations of the world and the IBE framework provides the philosophical basis for making those choices in practice. This focus on the credibility of connecting observed data to specific causal or

non-causal quantities is shared more generally across social science (Samii, 2016; Ashworth, Berry and de Mesquita, 2021).

In this section, we show how different types of social science work, estimating specific estimands within the context of a way to evaluate different explanations. While the credibility of the facts is important, our point here is that the way those facts are marshalled as evidence for or against some explanation is also essential and not always straightforward.

3.2.1 IBE in Observational Studies

In one of the most highly cited articles ever published in the *American Political Science Review*, Fearon and Laitin (2003) seek to explain why the 20th Century saw a notable rise in civil conflict. They open their second paragraph by posing the question: "What explains the recent prevalence of violent civil conflict around the world?" (75). They then spend the next several paragraphs enumerating a set of candidate explanations from conventional wisdom of what makes countries susceptible to civil war: the end of the Cold War and associated changes in the international system, ethnic or religious diversity, and ethnic or broad political grievances. They then place their fourth preferred explanation on the table—conditions that favor insurgency, including weak central governments, positive shocks to insurgent capabilities, and rough terrain. Much of the article is devoted to detailing 10 different empirical facts (framed as hypothesis tests) that would be implied by different explanations of why civil wars happen. While the explanations themselves are essentially causal in nature, the tests are often framed in either descriptive or predictive terms. Like the doctor examining symptoms, Fearon and Laitin (2003) describe the conditions that they would expect to see in the world were each explanation true. They then test these conditions using five different logistic regression models containing thirteen predictor variables run on cross-country data.

Fearon and Laitin (2003) has been the subject of critiques on methodological grounds, many of which have emphasized the concerns around post-treatment bias and thus challenged the empirical credibility of the tests (e.g. Samii, 2016). Focusing on claims about the relationship between economic shocks and violence, Ashworth, Berry and de Mesquita (2021, Ch 9.2) note that the entangled nature of the mechanisms concerned make it hard to know what association we would expect to see between economic performance and civil conflict even if the economy was driving the conflict. This is one of the challenges of IBE: even when trying to apply descriptive or predictive tests, causal reasoning is generally necessary to determine what we would expect to see given that the explanation were true. Of course, some explanations can be more easily removed from consideration: Fearon and Laitin (2003) quickly dismiss the first common wisdom explanation—the increase is due to the changes at the end of the Cold War—by showing in their first figure that civil war had been steadily rising since at least 1950.

Regardless of whether one finds the tests convincing, the strategy here is one of inference to the best explanation. Thus we can evaluate the contribution of the work not just on the quality of empirical evidence, but also on the development of the candidate explanations. Even if we believe the tests of the 10 hypotheses are convincing, this does not rule out other explanations they do not consider, it simply suggests that their preferred explanation is the best of the four they offered. This makes the credibility of their conclusion—and the policy implications they draw from it at the end of the article—turn as heavily on this candidate set of explanations as the empirical credibility of the tests. We think this fact that is under-appreciated in the methodological literature.

3.2.2 IBE in Experiments

While IBE helps to reinterpret what is happening in observational studies with large regression tables, it is also the mode of inference used in randomized experiments targeting

very specific quantities. This is most obvious in the context of laboratory experiments that capture some abstract form of behavior that is intended to generalize to a broader class of real-world settings. In their paper "Winners don't punish", Dreber et al. (2008) aim to assess a discipline-crossing literature in social choice about the role of self-sacrificing punishment in maintaining cooperation. They bring subjects into the lab to play a form of repeated Prisoner's Dilemma games where—in addition to the usual cooperate and defect—there was a third option to "punish" by paying 1 unit of money to have the other player lose 4 units. They find that those who perform well don't use the ability to exact the punishment. They write, "this suggests that costly punishment behaviour is maladaptive in cooperation games and might have evolved for other reasons" (Dreber et al., 2008, 1). That is, their implicit understanding is that a common set of behaviors governs the way humans behave in many settings. Here the data is the result of this simplified experiment, but their inference is that this is reflective of behaviors in a much wider array of settings.

The logic of IBE also pertains to settings where we have a relatively focused field experiment. Consider for example, the famous audit study run by Bertrand and Mullainathan (2004) to study discrimination. The motivating observation of the study is that in the United States, Black prospective workers are twice as likely to be unemployed as white prospective workers. This is consistent with a range of possible explanations including racial discrimination at the point of application or differences in the supply of applicants (presumptively due to prior history of racial discrimination). The authors design a method of data collection that adjudicates these explanations: they submitted the same resumes to different jobs, swapping out the names with a set intended to convey that the applicant was white or Black. The data is that a list of names including Emily, Anne, Todd and Neil received 50% more callbacks than a list of names including Aisha, Keisha, Rasheed, Tremayne. Since the resumes were the same and we might reasonably believe that employers don't have especially strong preference towards names *per se*, they infer that the best explanation is a discrimination on

race at the point of application.

Even this relatively clean comparison involves an inference that the driving feature is signaling race and not socio-economic status or some other property of the applicant. Bertrand and Mullainathan (2004) attempt to adjudicate among these explanations by showing that the gap between the names is no higher for jobs that rely on soft skills or interpersonal interactions. They also show that their evidence is not well explained by either of the two major economic theories usually used to explain discrimination: taste-based and statistical discrimination (Becker, 1957; Arrow, 2015). They suggest then that an alternative model based on a heuristic screening might be a better explanation for the behavior they observe.

Bertrand and Mullainathan (2004) quite explicitly consider a host of explanations for observed discrepancies that their experiment cannot directly adjudicate. However, the combination of the data they do observe allows them to make inferences about the most plausible of the remaining explanations. Even so, numerous studies have explored alternative explanations for the patterns observed in the original study.[5]

Experiments are attractive because the randomization makes the correspondence between the observed association and the causal estimand particularly plausible. However, this 'merely' establishes a particularly reliable fact about the world, but does not directly evaluate a theory. IBE is the framework that let's us move between the particular facts we observed to the broader explanation. Even in the case of specifically-designed experiments, the set of possible candidate explanations plays a crucial role in our understanding of what is a reasonable inference.

[5]See e.g., Fryer and Levitt (2004) on distinctively black names and the summary of critiques in Pager (2007).

4 How IBE helps and how it doesn't

If IBE is at the center of all science, then it seems reasonable to assert that it is—or should be—at the center of social science too. Indeed, we would argue that placing it there helps to understand why certain "debates" in the literature seem interminable. To fix ideas for what follows, consider two complete studies. In the first, the analyst conducts a series of observational data regressions and provides reasonable evidence for E_1 over E_2 and E_3. In the second study, the author conducts an experiment and concludes not E_2 (i.e. there is no evidence for explanation number 2). Which of these should the field, as a whole, prefer? In what follows, we will not give a definitive answer, but we contend that IBE provides structure to assess this question. That is, our claim is that doing IBE involves implicit tradeoffs, and recognizing those tradeoffs improves how we think about the relative merits of approaches.

4.1 HARKing and related replication problems

In Kerr's (1998, 197) original words, Hypothesizing After the Results are Known (HARKing) is defined as "presenting a post hoc hypothesis in the introduction of a research report *as if* it were an a priori hypothesis" (emphasis added). Put simply, the idea is that authors run several (perhaps many) different analyses involving statistical tests, and then write hypotheses consistent with the (strongest) results they find. Given that hypotheses should be first derived from theory, and thus offer a way to test the implications of that theory, the fact that HARKing reverses this basic process is immediately concerning. And indeed, the problems with this kind of interpretative overfitting have become well known. They include specific issues, such as a general tendency for Type-I errors ("false positives") to be presented as findings (Ioannidis, 2005). When this is combined with author and journal selection effects such as the "file drawer problem" (in the sense of Rosenthal, 1979) publication bias may be inevitable (Dickersin, 1990). Of course, HARKing is not the only problem in modern

empirical research. More narrowly, researchers may overfit in a purely statistical sense by "p-hacking"—trying different model and data specifications in a deliberate attempt to reach a particular level of significance (Simmons, Nelson and Simonsohn, 2011). Or they might do such things less consciously, but ultimately resulting in the same problem of findings that cannot be replicated (Gelman and Loken, 2014).

Whatever the specific causes and consequences of such practices, a corollary is that two practices are undervalued. First, the reporting of null results (e.g. Franco, Malhotra and Simonovits, 2014); second, exploration and description (e.g. Gerring, 2012). An appreciation of IBE can improve matters here by refocussing scientific attention. To begin, recall that inference to the best explanation obviously requires candidate explanations. In that sense, generating new explanations is part of IBE. HARKing occurs because scholars explore the data and then pick a particular explanation to (superficially) "test" via a hypothesis. But under IBE, this subterfuge is pointless. That is, if analysts want to explore data to suggest new explanations, they should just do this without artificially stating hypotheses *as if* testing a pre-existing theory. And, for all the usual reasons, those explanations should not be tested with the data used to generate them. But that is a separate matter.

Second, IBE suggests that description is *per se* important. In the medical example we gave above, the physician needed symptom data to make a diagnosis. So, anything that devalues "mere" description—and the incentives behind HARKing certainly do—is anathematic to IBE and the scientific approach that embodies.

Finally, IBE helpfully broadens our understanding of what a useful finding might be. In HARKing, there is little value *per se* to a scientifically "correct" explanation: it exists simply as an author justification for a hypothesis test which has already been conducted. If it turns out to be right, that is fortunate, but not required. By extension, there is no motivation to show that, for example, an explanation previously thought to be "best" is in fact weaker than believed. Or that none of the usual explanations for this sort of case (say,

a rich country that is not a democracy) do not work well. But IBE more correctly aligns the benefits and costs: hypothesis tests and p-values help us rule things *out* as much as ruling things in.

The point here is not that HARKing is worse than we thought. The point is to focus on IBE as an alternative incentive structure that would severely limit or extinguish concerning empirical practices.

4.2 The Myth of "Big" vs "Small" Research

To reiterate our comments above, there is ongoing debate as to the appropriate role of causal identification in the discipline. In one corner, scholars like Huber (2013) (see also Huber 2017, Ch 6) argue that the turn to causal inference designs in social science is not an unalloyed good. The first reason for this is that many substantively interesting phenomena do not naturally lend themselves to such work (because e.g. the treatment cannot be plausibly randomized), and thus we see less effort to study such questions. The second reason is that focussing on identification opportunities crowds out theory development: the claim here is that traditional (not plausibly causal) regression designs help us refine our understanding of relationships in observational data. In the context of randomized controlled trials (RCT), Deaton and Cartwright (2018) make an allied argument. That is, the results of (necessarily) specific RCTs cannot be easily extrapolated to broader questions of interest in a field. In the other corner, scholars like Samii (2016) contend that these fears are either overblown or exactly wrong. More specifically, the contention is that traditional designs, for various technical reasons, generate "pseudo-general pseudo-facts" (Samii, 2016, 1). And such entities are a bad basis for either trying to understand phenomena or building theories about them. Thus, to the extent that the credibility revolution has changed practices, it has done so in a way that moves authors away from actively misleading themselves from their results.

Though neither author uses such crude terms, the debate over the relative importance

of causal inference designs can be summarized thusly: is it better to get potentially wrong answers to 'big' questions, or the right answers to 'small' ones? Attention to IBE cannot answer this question, but it can help structure our answers to it. If the analyst takes the position that what matters (most) is causal identification over generating explanations, they are focused on the *quality of inference* in IBE. That is, whatever the—potentially very narrow—set of explanations, they want to believe that the inference they make is the correct one. By contrast, scholars like Huber can be seen as emphasizing the importance of *quality of explanation* in IBE. That is, they want the data and models to tell us about the relative plausibility of many explanations, or actively develop new explanations. And they want to do this, even if that same machinery cannot do a particularly good job of determining which one of those explanations is 'best'. Notice that this is *not* a question of "external" versus "internal" validity (in the sense of, e.g. Morton and Williams, 2010; McDermott, 2011): the problem is not that one is sacrificing generalizability for (local) credible estimation of causal effects. The issue instead is the preferred tradeoff between inference and building explanations.

Put this way, there is more agreement than is initially obvious. If the primary concern is tests of extant theory, then quality of inference must be the priority. But this will not help with theoretical progress *per se*: it is trivially true that an infinite number of theories is consistent with a particular (well-identified) causal effect. Related, focusing on quality of explanation is not *a priori* wrong. But if the tests of that explanation are weak in the sense described by Samii (see, also, Aronow and Samii, 2016), IBE may not happen. A different and perhaps more positive way to express our sentiments here is to consider the value of a regression study for which the relevant treatment effect is either poorly specified, poorly identified, or both. We would contend that this regression—which would include most quantitative observational work done in the discipline until very recently—is not deleterious to our general efforts unless one of two conditions holds. First, that it leads the reader to

the *wrong* inference about which explanation is best. Second, that it suggests that a new *wrong* explanation should be included in the set of explanations from which a best one will hopefully emerge. This may be a low bar, and the empirical frequency with which these conditions are met may be very high. But IBE makes clear that explicit consideration of this frequency matters.

4.3 Revaluing Exploration and Description

As hinted at above in our comments on HARKing, IBE encourages us to value exploration and description *per se*. The idea that exploratory data analysis (EDA) is important is not new, and begins at least with Tukey (1977). Indeed, in an instructive analogy, Tukey (1977, 1) notes that the data analyst should be a "detective" attempting to fact find, before—in a separate stage–the jury system makes a ruling. This judicial role is where "Confirmatory Data Analysis" (CDA) steps in. Since CDA involves hypothesis *testing*, it is different to EDA, which involves

> a focus on tentative model building and *hypothesis generation* in an iterative process of model specification, residual analysis, and model respecification

in the words of Behrens (1997, 132, emphasis added). In line with our argument above, we think that in much modern political science work, the idea of EDA (hypothesis generation) is not kept separate to CDA (hypothesis testing). The result is that researchers often try to do both and end up doing neither. Thus we see statements of hypotheses in papers that do not comport with either Fisher or Neyman-Pearson decision theory. That is, there is no explicit null hypothesis, nor is there discussion of a test statistic; nor ultimately, is there a mutually exclusive decision to reject or fail to reject the null.[6] Instead, scholars write of

[6]See Mayo (1996, Ch 12) for a nuanced discussion on the historical relationship between Peirce's work and Neyman-Pearson statistics.

statistical evidence "partially consistent" with a hypothesis (Thompson, 1975, 474); or they talk of hypotheses which are "strongly supported" by the data (Lau, 1985, 130).[7]

Studies occupy this awkward middle ground between EDA and CDA in part because incentives are misaligned. IBE is then helpful here precisely because generating potential new explanations is valuable in this framework. Indeed, we are not the first to explicitly link the idea of EDA to Peirce's original (1878) work on abductive inference. Behrens and Yu (2003, 40), for example, note that

> In EDA, after observing some surprising facts, we exploit them and check the predicted values against the observed values and residuals. Although there may be more than one convincing pattern, we "abduct" only those that are more plausible for subsequent confirmatory experimentation.

If the generation of "plausible" hypothesis were itself valued, this would allow subsequent (confirmatory) hypothesis-testing language to more closely resemble the inference frameworks (e.g. Neyman-Pearson) for which it was designed. Though not motivated by IBE itself, one tool that may encourage such practices is pre-analysis plans, which are becoming more accepted and more popular (see, e.g. Ofosu and Posner, 2021). They do this by limiting the extent to which researchers can add explanations (Es) after performing the (confirmatory) analysis stage (which itself is after the EDA).

A special case of these problems is presented by much "text as data" work. There, scholars will use unsupervised techniques—including topic models (Quinn et al., 2010; Roberts et al., 2014, e.g.)—which are designed for summarizing and organizing a corpus. Yet researchers routinely use the output of these models to make statements about the plausibility of various theories. The issue here is not simply about the availability of forking paths arising from decisions one must make in fitting the models (see, e.g., Denny and Spirling, 2018). It is

[7]We use older studies here, because our goal is not to critique specific scholars, but rather to give examples of the types of language that can be found in many modern works.

that, fundamentally, unsupervised techniques are primarily methods of discovery which can only be repurposed as tools of measurement with substantial care and validation (Grimmer and Stewart, 2013; Grimmer, Roberts and Stewart, 2022). The danger is what we term "PEACHing": [P]resenting [E]xplorations [A]s [C]onfirmable [H]ypotheses. This is, in a sense, the opposite of HARKing. In HARKing, techniques (like regression) that allow for hypothesis tests are used to assess a large number of possible hypotheses, and those with suitable p-values presented as if *a priori* theorized. In PEACHing, techniques that do *not* naturally facilitate hypotheses tests are used to suggest hypotheses that cannot be "tested" at all, at least on that data and with that approach (though see, e.g., Egami et al., 2018, on testing with an explicit heldout sample). What IBE makes clear is that these approaches can be used to help suggest new explanations, but the role of explanation generation must be carefully separated from the testing of those explanations.

An additional interesting consequence of valuing exploration is that it recasts some supposed contrasts between quantitative and qualitative methods. In that literature, there have been considerable recent efforts to use mixed methods for *inference* (e.g. Glynn and Quinn, 2011; Humphreys and Jacobs, 2015). Such attempts are not uncontroversial, in the sense that some scholars of quantitative (e.g. Beck, 2006) and qualitative (e.g. Collier, Brady and Seawright, 2004) approaches argue those techniques can and should be used for fundamentally different purposes. But from an IBE perspective, where the goal might be to generate new explanations, it is not clear that a distinction between "within-case" rather than "between-case" variation makes much difference for the problem at hand. That is, even if one asserts that the (causal) *inferences* possible with different methods vary, it is not obvious that one is obviously preferable to the other for the initial component of abductive inference.

4.4 Open Problems in IBE

As we hinted at above, exactly how one *does* IBE is often unclear (see, e.g., Van Fraassen, 1989). Indeed, in the words of Lipton (2003, 2) it "is more a slogan than an articulated philosophy." Unsurprisingly, there have been considerable efforts to make IBE *more* of an articulated philosophy, with special attention to its relationship—and compatibility—with various of forms of Bayesianism (Douven, 2013; Schupbach, 2017; Henderson, 2020). But ambiguity persists, not least for social scientists. To see why, consider two separate studies that examine the same phenomenon, but come to different conclusions as to that phenomenon's ("best") explanation. This could include, say, culture versus institutions as the cause of some political outcome like the effective number of parties in a system (see e.g. Neto and Cox, 1997). But it would also extend to the special case of situations in which some scholars assert a given causal effect of a treatment from data (e.g. Kroenig, 2013), while others find the opposite and deny it exists (e.g. Sechser and Fuhrmann, 2013).

Put crudely, what does IBE have to say about which of these (hypothetical) studies is "better"? To answer this, recall that abductive inference has only two places in which one can make a judgement about which study is preferred. First, studies can differ in terms of the nature of the *explanations* themselves. But exactly what properties are desirable, and how they should be traded off is unclear. For example, *parsimony* might be preferable for some scholars but not others (relative to, say, *generality*). Second, studies can differ in the way that they move from data to explanation: that is, for a fixed set of explanations, the purported inference to the "best" one is of higher or lower quality. In practice here, there is presumably some notion of "plausibility" or "credibility" (in the sense of e.g. Angrist and Pischke, 2010). But, to be clear, the requisite standards are not obvious and have not been constant over time.

5 Discussion

The credibility revolution has been a "catastrophic success"; it has so profoundly changed the practice of social science, that it has created new problems in its wake. In particular, scholars have claimed that it is crowding out questions, and approaches to those questions, that also deserve attention. Consequently, the argument goes, the discipline is all the poorer. It is true that we see more attention to ensuring claims of causal effects are plausible, and this presumably reduces the production and/or value of "pseudo-general pseudo-facts" (in the words of Samii, 2016, 941). But this isn't the full story. Our argument above was that both improving inference and improving explanations have a natural role in political science. A given researcher or research agenda may value one aspect over another, but they are not necessarily in direct competition. Furthermore, the abductive inference mode in which they must both simultaneously exist as crucial elements—known as *Inference to the Best Explanation*—is both ubiquitous and of long-standing. We argued that this was true of essentially all empirical set ups: from observational regressions where conditional ignorability may or may not be plausible, to randomized lab experiments where it is generally agreed that a treatment effect has been identified.

Understanding that almost all such research is some variant of IBE allows for a common framework of discussion across the discipline. But it does more than that. By clarifying the importance of explanations—their development and their plausibility—IBE ameliorates a series of debates in political science. In particular, we argued that if explanation matters, then exploring and describing data becomes *per se* valuable. And if that is true, there is no reason to misrepresent ('causal') hypotheses as if they occurred to the reader after running the relevant regression: the incentive to "p-hack", or to hide null results away, is reduced. More generally, we argued that recognizing the centrality of IBE means avoiding unhelpful and uncharitable comments about whether a given research question is "big" or "small"—

and thus worthy of answering (or not), and with different degrees of precision. Seen this way, these arguments are in fact mostly displays of researchers' preferences—over what they value in terms of the components of the IBE framework. We contend that if these are made explicit, progress is more likely.

In short then, the answer to "what good is a regression?" depends on what you want from it. But our point was that there are reasonable answers that arise from the same philosophical framework being used (implicitly) by those that particularly promote causal inferences. This does not mean anything goes. Just as an example, IBE for a given problem is not served by including controls that make most sense if the goal was the estimation of a causal effect yet asserting (caveating) that the regression results should not be interpreted causally.[8] Similarly, there is little IBE value in writing out *ad hoc* "hypotheses" that connect neither to the proposed (best) explanation, nor to how one might make an inference to it.

We sought to make the basic point that the fact that a regression coefficient does not have a plausibly causal interpretation—or that it does, but not to the standard required by a given reader—is not, and cannot be, a reason to say that regression is *per se* unhelpful. Given this main task, we have not given a deep philosophical treatment or resolved all issues. This is a problem, because IBE has no immediate and obvious solution for many concerns that are required for working across the discipline. For instance, as others have noted, it is not obvious what makes for a "best" explanation: certainly, generality and parsimony might be properties of interest, but they are often in tension. Nor is clear exactly where explanations ought to come from, or how different they need to be from one another in practice. We leave such tasks for future work.

[8]The current authors have heard this practice referred to as the "Causal Two Step" but have no citation for this term.

References

Acemoglu, Daron and James A Robinson. 2001. "A theory of political transitions." *American Economic Review* 91(4):938–963.

Achinstein, Peter. 1983. *The nature of explanation*. Oxford University Press on Demand.

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

Aronow, Peter M and Benjamin T Miller. 2019. *Foundations of agnostic statistics*. Cambridge University Press.

Aronow, Peter M and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1):250–267.

Arrow, Kenneth J. 2015. *The theory of discrimination*. Princeton University Press.

Ashworth, Scott, Christopher R Berry and Ethan Bueno de Mesquita. 2021. *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton University Press.

Barnes, Eric. 1995. "Inference to the loveliest explanation." *Synthese* pp. 251–277.

Beck, Nathaniel. 2006. "Is causal-process observation an oxymoron?" *Political Analysis* 14(3):347–352.

Becker, Gary S. 1957. *The economics of discrimination*. University of Chicago press.

Behrens, John T. 1997. "Principles and procedures of exploratory data analysis." *Psychological Methods* 2(2):131.

Behrens, John T and Chong-ho Yu. 2003. "Exploratory data analysis." *Handbook of psychology* 2:33–64.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review* 94(4):991–1013.

Binder, Sarah. 2020. "How we (should?) study Congress and history." *Public Choice* 185(3):415–427.

Bird, Alexander. 2007. "Inference to the Only Explanation." *Philosophy and Phenomenological Research* 74(2):424–432.

Blackwell, Matthew. 2014. "A selection bias approach to sensitivity analysis for causal effects." *Political Analysis* 22(2):169–182.

Boix, Carles and Susan C Stokes. 2003. "Endogenous democratization." *World politics* 55(4):517–549.

Boyd, Richard N. 1984. 3. The Current Status of Scientific Realism. In *Scientific realism*. University of California Press pp. 41–82.

Clark, William Roberts and Matt Golder. 2015. "Big data, causal inference, and formal theory: Contradictory trends in political science?: Introduction." *PS: Political Science & Politics* 48(1):65–70.

Clarke, Kevin A and David M Primo. 2012. *A model discipline: Political science and the logic of representations*. Oxford University Press.

Collier, David, Henry E Brady and Jason Seawright. 2004. Sources of leverage in causal inference: Toward an alternative view of methodology. In *Rethinking social inquiry: Diverse tools, shared standards*. Rowman and Littlefield pp. 229–266.

Cranmer, Skyler J and Bruce A Desmarais. 2017. "What can we learn from predictive modeling?" *Political Analysis* 25(2):145–166.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210:2–21.

Denny, Matthew J and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." *Political Analysis* 26(2):168–189.

Dickersin, Kay. 1990. "The existence of publication bias and risk factors for its occurrence." *Jama* 263(10):1385–1389.

Douven, Igor. 1999. "Inference to the best explanation made coherent." *Philosophy of Science* 66:S424–S435.

Douven, Igor. 2013. "Inference to the best explanation, Dutch books, and inaccuracy minimisation." *The Philosophical Quarterly* 63(252):428–444.

Douven, Igor. 2021. Abduction. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Summer 2021 ed. Metaphysics Research Lab, Stanford University.

Dowding, Keith. 2015. *The philosophy and methods of political science*. Macmillan International Higher Education.

Dowding, Keith and Charles Miller. 2019. "On prediction in political science." *European Journal of Political Research* 58(3):1001–1018.

Dreber, Anna, David G Rand, Drew Fudenberg and Martin A Nowak. 2008. "Winners don't punish." *Nature* 452(7185):348–351.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts and Brandon M Stewart. 2018. "How to make causal inferences using texts." *arXiv preprint arXiv:1802.02163* .

Fearon, James D and David D Laitin. 2003. "Ethnicity, insurgency, and civil war." *American political science review* 97(1):75–90.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Fryer, Jr., Roland G and Steven D Levitt. 2004. "The causes and consequences of distinctively black names." *The Quarterly Journal of Economics* 119(3):767–805.

Gelman, Andrew and Eric Loken. 2014. "The statistical crisis in science: data-dependent analysis–a" garden of forking paths"–explains why many statistically significant comparisons don't hold up." *American scientist* 102(6):460–466.

Gelman, Andrew and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gerber, Alan S, Donald P Green, Edward H Kaplan, Ian Shapiro, Rogers M Smith and Tarek Massoud. 2014. "The illusion of learning from observational research." *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* pp. 9–32.

Gerring, John. 2012. "Mere description." *British Journal of Political Science* 42(4):721–746.

Glynn, Adam N and Kevin M Quinn. 2011. "Why process matters for causal inference." *Political Analysis* 19(3):273–286.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

Grimmer, Justin, Margaret E Roberts and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Gross, Justin H. 2015. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." *American Journal of Political Science* 59(3):775–788.

Harman, Gilbert H. 1965. "The inference to the best explanation." *The philosophical review* 74(1):88–95.

Henderson, Leah. 2020. "Bayesianism and inference to the best explanation." *The British Journal for the Philosophy of Science* .

Huber, John. 2013. "Is Theory Getting Lost in the 'Identification Revolution'?" *The Money Cage* .

Huber, John D. 2017. *Exclusion by elections: inequality, ethnic identity, and democracy*. Cambridge University Press.

Humphreys, Macartan and Alan M Jacobs. 2015. "Mixing methods: A Bayesian approach." *American Political Science Review* 109(4):653–673.

Imai, Kosuke and Teppei Yamamoto. 2010. "Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis." *American Journal of Political Science* 54(2):543–560.

Ioannidis, John PA. 2005. "Why most published research findings are false." *PLoS medicine* 2(8):e124.

Keele, Luke, Randolph T Stevenson and Felix Elwert. 2020. "The causal interpretation of estimated associations in regression models." *Political Science Research and Methods* 8(1):1–13.

Kerr, Norbert L. 1998. "HARKing: Hypothesizing after the results are known." *Personality and social psychology review* 2(3):196–217.

King, Gary, Robert O Keohane and Sidney Verba. 1994. *Designing social inquiry*. Princeton university press.

Kroenig, Matthew. 2013. "Nuclear superiority and the balance of resolve: Explaining nuclear crisis outcomes." *International Organization* 67(1):141–171.

Lau, Richard R. 1985. "Two explanations for negativity effects in political behavior." *American journal of political science* pp. 119–138.

Lieberson, Stanley and Joel Horwich. 2008. "Implication analysis: a pragmatic proposal for linking theory and data in the social sciences." *Sociological Methodology* 38(1):1–50.

Lipset, Seymour Martin. 1959. "Some social requisites of democracy: Economic development and political legitimacy1." *American political science review* 53(1):69–105.

Lipton, Peter. 2003. *Inference to the best explanation*. Routledge.

Little, Andrew T and Thomas B Pepinsky. 2021. "Learning from biased research designs." *The Journal of Politics* 83(2):602–616.

Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. "What is your estimand? Defining the target quantity connects statistical evidence to theory." *American Sociological Review* 86(3):532–565.

Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. University of Chicago Press.

McDermott, Rose. 2011. "Internal and external validity." *Cambridge handbook of experimental political science* pp. 27–40.

McShane, Blakeley B, David Gal, Andrew Gelman, Christian Robert and Jennifer L Tackett. 2019. "Abandon statistical significance." *The American Statistician* 73(sup1):235–245.

Montgomery, Jacob M and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Moore, Barrington et al. 1993. *Social origins of dictatorship and democracy: Lord and peasant in the making of the modern world*. Vol. 268 Beacon Press.

Morton, Rebecca B and Kenneth C Williams. 2010. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press.

Munger, Kevin, Andrew M Guess and Eszter Hargittai. 2021. "Quantitative description of digital media: a modest proposal to disrupt academic publishing." *Journal of Quantitative Description* 1(1):1–13.

Neto, Octavio Amorim and Gary W Cox. 1997. "Electoral institutions, cleavage structures, and the number of parties." *American Journal of Political Science* pp. 149–174.

Ofosu, George K and Daniel N Posner. 2021. "Pre-analysis plans: An early stocktaking." *Perspectives on Politics* pp. 1–17.

Pager, Devah. 2007. "The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future." *The Annals of the American Academy of Political and Social Science* 609(1):104–133.

Pearl, Judea. 2014. "The deductive approach to causal inference." *Journal of Causal Inference* 2(2):115–129.

Peirce, Charles. 1878. "How to make our ideas clear." *Popular Science Monthly* 12(Jan):286–302.

Przeworski, Adam, R Michael Alvarez, Michael E Alvarez, Jose Antonio Cheibub, Fernando Limongi, Fernando Papaterra Limongi Neto et al. 2000. *Democracy and development: Political institutions and well-being in the world, 1950-1990*. Cambridge University Press.

Psillos, Stathis. 2002. Simply the best: A case for abduction. In *Computational logic: Logic programming and beyond*. Springer pp. 605–625.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural topic models for open-ended survey responses." *American journal of political science* 58(4):1064–1082.

Rosenthal, Robert. 1979. "The file drawer problem and tolerance for null results." *Psychological bulletin* 86(3):638.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.

Schupbach, Jonah N. 2017. "Inference to the best explanation, cleaned up and made respectable." *Best explanations: New essays on inference to the best explanation* pp. 39–61.

Sechser, Todd S and Matthew Fuhrmann. 2013. "Crisis bargaining and nuclear blackmail." *International organization* 67(1):173–195.

Sekhon, Jasjeet S. 2009. "Opiates for the matches: Matching methods for causal inference." *Annual Review of Political Science* 12:487–508.

Sekhon, Jasjeet S and Rocio Titiunik. 2012. "When natural experiments are neither natural nor experiments." *American Political Science Review* 106(1):35–57.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological science* 22(11):1359–1366.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman and Wolf Vanpaemel. 2016. "Increasing transparency through a multiverse analysis." *Perspectives on Psychological Science* 11(5):702–712.

Tavory, Iddo and Stefan Timmermans. 2014. *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Thompson, William R. 1975. "Regime vulnerability and the military coup." *Comparative Politics* 7(4):459–487.

Tukey, John W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford University Press.