Causes, Theories, and the Past in Political Science[1]

Sanford C. Gordon

Wilf Family Department of Politics

New York University

19 West 4th Street, 2nd Floor

New York, NY 10012

sanford.gordon@nyu.edu

Hannah K. Simpson

Department of Political Science

Texas A&M University

4220 George Bush Drive West

College Station, TX 77840

hannah.simpson@tamu.edu

[1]We thank Pat Egan, Catherine Hafer, Carlo Horz, Kris Kanthak, Dimitri Landa, Cyrus Samii, and participants at the USC Causal Inference and American Political Development Conference for valuable comments and clarifying discussions.

Abstract

A theoretically-grounded approach to causal questions illuminates both the utility and limitations of the potential outcomes (PO) framework as a model for historically-focused, quantitative empirical research. While some causal questions are immediately reconcilable with the PO framework, for others, theoretical guidance is valuable in ascertaining relevant comparisons or characterizing the generalizability of findings to different contexts. A third category of important causal relationships feature strategic or information-based interactions, or multiple or unobservable mechanisms, many of which cannot be directly tested using the PO framework. Here, theory is critical in elucidating additional, observable implications that may be tested empirically. In all three categories, historical research holds special benefits: it expands the set of cases on which to test causal claims, may provide counterfactuals not available in contemporary contexts, and can feature institutional transformations that function as plausibly exogenous modifier variables. We clarify this classification of causal questions using examples from our own historical research.

1 Introduction

Over the past several decades, the potential outcomes (PO) framework of Neyman and Rubin has emerged as a unifying, paradigmatic approach for empirically establishing causal relationships, with the randomized, controlled trial (RCT) serving as the "gold standard" for causal inference. Even when actual experiments are impossible for logistical or ethical reasons, the PO framework encourages us to think about causal inference in terms of hypothetical interventions that we would want to run if we could; to be attentive to the problem of unobserved confounders; and to look for situations in which observational data most closely approximate the hypothetical experimental ideal. The adoption of the PO framework has brought welcome attention to fundamental challenges of causal inference, and afforded questions of design a renewed place of prominence in the research process. But some challenges stand in the way of a wholesale adoption of the PO framework by scholars of a historical bent, including those interested in American political development (APD).

On the one hand, some social scientists focused on the study of history may perceive the PO framework as providing an insufficiently rich toolkit to address many of the causal questions that animate their field of inquiry. Historical institutionalists, for example, may be interested primarily in mapping the vast tapestry of interconnected events that together produced some macropolitical outcome of interest (e.g., the modern administrative state). Such scholars may reject the PO framework altogether, along with the claim that the only path to causal knowledge is the one provided by research reconcilable with it. On the other hand, some of the more zealous advocates of the PO framework may view historical research as beside the point. For example, several prominent advocates of experimental political science have argued that given uncertainty about potential biases (e.g., from measurement error or omitted variables), "the only possibility for further learning comes from experiments, particularly experiments with strong external validity" (Gerber, Green, and Kaplan, 2014). Historical research – or, potentially, any observational research – is, in this view, of limited utility in producing general causal knowledge of political phenomena and processes: it's

impossible to go back in time to run RCTs; there are only so many natural experiments out there; and even an unimpeachable research design is of limited value if we believe that the study's historical context renders its results inapplicable in contemporary settings. Historical scholars who accept this view can apply the PO framework to historical causal questions – but must narrow the scope of their inquiry to apply only to the specific context under scrutiny.

In this article, we argue that once we think more theoretically about the substantive comparisons we choose to make and the contexts in which we choose to make them, it becomes clear that historical research holds special value for the project of rigorously identifying causal relationships of general interest, and that the PO framework can be of substantial use to this objective. Most obviously, a causal relationship that is theorized to be general (i.e., that is theorized to exist under a range of different background conditions) ought to be demonstrable — and therefore testable — across a range of times as well as places. There is nothing representative about the present as compared to the past, and as has been noted by others, the quintessential feature of observational studies is the *choice* of appropriate units to study that provide the maximum degree of control and comparability (Rosenbaum, 1999). In the most immediate sense, then, looking at the past expands the choice set of where to look for comparable observational units.

This suggests that historical applications in political science research may be comfortably reconciled with the PO framework. Theory, which can identify substantively relevant counterfactuals and anticipate the background conditions under which causal relationships will and will not hold (cf., Ashworth, Berry, and Bueno de Mesquita, 2021), facilitates this reconciliation: historical applications can furnish us with meaningful counterfactuals not available in contemporary contexts; historical institutional transformations, when carefully selected, can serve as plausibly exogenous modifier variables. Indeed, it is never the historical nature of a research endeavor that limits the usefulness of the PO framework. Rather, some causal questions, such as those requiring researchers to identify or adjudicate between specific causal

mechanisms, isolate the role of unobservable mediating variables, consider off-the-path causal outcomes, or estimate the effect of informational treatments, require sustained theoretical attention before the analyst can determine if, and how, the PO framework may be deployed effectively. When answering such questions requires researchers to evaluate constellations of causal relationships, non-causal correlations, and point or interval predictions, historical research is again useful in expanding the choice set of where to look. Following an extended discussion of these issues, we turn to examples from our own research on the emergence of legal institutions and the development of local appropriations politics.

Causal Questions in Political Science Research

The question "What explains Y?" lies at the heart of nearly every social science inquiry. Supposing we are interested in understanding what *causes* Y, the potential outcomes framework can provide a useful way to understand causal relationships and to think about the assumptions that underlie our ability to recover causal estimates from data: by conceiving observable attributes of units (individuals; states; roll call votes) in terms of the potential values they *would* take under different treatment conditions. In other words, for a treatment $T \in \{0, 1\}$, Y_{1i} denotes the value Y_i would take given $T_i = 1$, Y_{0i} the value it would take given $T_i = 0$, and the unit-level causal effect of the treatment is $Y_{1i} - Y_{0i}$. While we of course cannot actually simultaneously observe Y_{1i} and Y_{0i} — instead observing the realized value $Y_i = T_i Y_{i1} + (1 - T_i)Y_{i0}$ — under some assumptions we can identify the average treatment effect, $E[Y_{i1} - Y_{i0}]$ from the comparison of the sample analogs of $E[Y_i|T_i = 1]$ (the expected value of Y given treatment $T_i = 1$) and its counterpart $E[Y_i|T_i = 0]$.

An especially vexing assumption for scholars working with observational data is ignorability: if treatment assignment is correlated with the potential outcomes, then $E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$ will yield a biased estimate of $E[Y_{1i} - Y_{0i}]$. The great virtue of a randomized experiment is that in expectation, it entails treatment ignorability by design. It is for this reason that the randomized controlled trial (RCT) is considered the "gold standard" of evaluating causal relationships. While the laws of physics rule out using RCTs in histori-

cal studies, in their absence we may still be able to recover conditional average treatment effects if treatment assignment is ignorable conditional on some set of covariates X. Our most powerful non-experimental tools for inference — differences-in-differences, regression discontinuity, instrumental variables — are all interpretable through this lens.

PO-Irreconcilable Causal Claims

How should we think about the connection between the PO framework and political science applications that rely on historical data? As a first cut, whether or not the framework applies seems to depend very heavily on the nature of the inquiry. Consider the following causal claims:

- The piecemeal regulatory response to the economic upheavals of the late 19th century was driven in large part by the decentralized administrative apparatus of the American state (Skowronek, 1982).
- Dramatic increases in the punitiveness of the criminal justice system in the United States resulted from the mobilization of opponents of racial equality following the 1964 Civil Rights Act (Weaver, 2007).
- An agency problem between the crown and colonial governors caused the emergence of separation of powers in Britain's North American colonies in the 17th and 18th centuries (Gailmard, 2017)

Each of these claims is causal, in the sense that it defines an outcome with respect to an implicit counterfactual: had the U.S. government been more centralized, the regulatory response to industrialization would have been more robust; had the Civil Rights Act not mobilized opponents of racial equality, the criminal justice system would be less punitive. But the methods used to substantiate these claims empirically seem irreconcilable with the PO framework. For one thing, given our inability to re-run history with slight modifications, it is difficult to envision the hypothetical randomized experiments that one could run in order to test these causal claims in a potential outcomes framework. For another, the focus

on individual cases affords no variation on the outcome of interest. The Gailmard paper may be the exception that proves the rule: as there were multiple British colonies, it's conceivable that they varied in the magnitude of the agency problem, and thus, potentially, the nature of the institutional solution to it. Variation may also be found by comparing, e.g., British and Spanish holdings in the New World (but note the obvious confounding problem, as British and Spanish colonies differed on innumerable dimensions other than the stated agency problem).

The mode of inquiry associated with answering this kind of question generally proceeds not by making the kinds of comparisons envisioned in the PO framework but by extensively mapping the context in which Y emerged in order to understand how Y came about. Labeled "historical institutionalism" by some of its practitioners, it is typically qualitative, and most closely associated in APD with, inter alia, Bensel, Katznelson, Orren, Pierson, Schickler, Skocpol, and Skowronek. According to Pierson and Skocpol (2002), historical institutionalists address "big, substantive questions that are inherently of interest to broad publics as well as to fellow scholars" by 'tak[ing] time seriously, specifying sequences and tracing transformations and processes of varying scale and temporality;" and by "analyz[ing] macro contexts and hypothesiz[ing] about the combined effects of institutions and processes rather than examining just one institution or process at a time."

Typically, a quintessential feature of such approaches is the perceived irreducible importance of context—meaning the causal knowledge they impart ought not, as a general rule, extend beyond the particular context under study. This feature of historical institutionalism make the method fundamentally irreconcilable with the PO approach. If contexts, institutions, and processes are inextricably intertwined and case-specific, then treatment ignorability cannot in principle be satisfied, and the existence of comparable units is ruled out a priori – each case is *sui generis*.

Immediately PO-Reconcilable Causal Claims

Now consider a different variety of historical claim:

- The adoption of direct primaries exacerbated (or did not exacerbate) partisan polarization (Hirano et al., 2010).

- Coming from a family that owned slaves made white southerners more likely to fight in the U.S. Civil War (Hall, Huff, and Kuriwaki, 2018).

These are claims that both invoke historical evidence and whose adjudication fits comfortably in the PO framework. We can observe states with and without direct primaries, and critically, we can observe the dates on which states shifted to direct primaries. With measures of legislator extremism in hand, we can assess whether the adoption of the direct primary contributed to polarization via differences-in-differences. This is the approach taken by Hirano et al. (2010) (who, incidentally, find little evidence to support the hypothesis). Similarly, Georgia's 1832 land lottery, the winners of which were more likely to hold slaves, provides an exogenous source of variation in slave ownership on the eve of the Civil War. We call instances of this second variety of hypothesis, to which the PO framework can be readily applied, *immediate* causal claims. They are immediate because (1) there is a clear potential outcomes interpretation – we can envision the hypothetical experiment that one would adopt to test it; and (2) the variables of interest, or good proxies thereof, are observable.

Why should scholars test immediate causal claims using *historical* data? There are three reasons. First, we may have strong reasons to believe that a causal effect should hold in a variety of contexts, and there may be no contemporary contexts in which one can obtain plausible exogenous variation on the treatment variable. As Rosenbaum (1999) succinctly puts it, compelling evidence of causal effects using observational data requires "an active search for those rare circumstances in which tangible evidence may be obtained to distinguish treatment effects from the most plausible biases." It would be foolish to restrict *when* one looks for exogenous variation unless there is something specifically temporally contingent about the posited causal relationship. According to this view, criticizing historical studies as lacking external validity because they are not representative of the present risks missing the point: such criticisms erroneously apply a standard relevant for descriptive inference –

how representative is a sample of a broader population – to questions of causal inference.

Second, even if relevant contemporary contexts exist, if our theory about a causal relationship anticipates that the relationship should exist across a broad range of situations, validating the theory requires testing it in numerous situations — including different temporal contexts. There is certainly no reason to believe that moments in the past are *less* representative of the human experience or human behavior (and the theories that pertain to these things) than moments in the present. Perhaps even more fundamentally, testing general claims using data from the past provides a critical test of the robustness of a theory's simplifying assumptions. Theorists are fallible, and may possess a natural tendency to base their assumptions on a host of unstated premises drawn from their lived engagement with the political world. Testing one's theory using data from the past can be an efficient way to unmask those premises, and, potentially, their violation.[1]

Finally, an historical research question may be intrinsically interesting even if the analyst makes no claims about the broader applicability of the causal effect. Hall et al. is valuable not only because it contributes to the large body of scholarship concerned with the antecedents of participation in war, but also because it speaks directly to an ongoing debate about the composition of the Confederate army. Similarly, Feigenbaum, Lee, and Mezzanotti (2018) document persistent economic effects of Sherman's march to the sea in 1864-65, adding to our understanding of the long-term implications of this monumentally important historical event.

What is the role of theory in articulating and assessing direct causal claims like these? In principle, it need not have contributed to the derivation of the primary hypothesis at all, beyond a simple hunch or gut feeling. We could swap out any institution and outcome and apply the same rigorous methodological approach to test the effect of legislator term limits on delegation to the bureaucracy, or public financing of campaigns on the electoral performance of working class candidates, for example. As it happens, theory does motivate

[1]Below, we consider an example of research that exploits historical data for precisely this purpose.

us to test *these* primary hypotheses and not, for example, the hypothesis that the typeface on a legislator's website affects her attendance in committee meetings. That is because in each, we can envision a particular mechanism or set of mechanisms that leads from X to Y: in the case of primaries, that primary electorates are more concerned with ideological purity than winning (Jacobson, 2004), and are therefore more willing to gamble on extremist candidates than party elites; in the case of slave ownership and combatant status, we anticipate a significant relationship if the stakes of fighting outweigh the incentives to free ride. But the potential outcomes framework will work just as well for poorly-motivated hypotheses. And if the assumptions underlying the research design are satisfied, the estimated causal relationship will be valid, if substantively vacuous.[2]

What's Missing?

One critical methodological concern, when working with historical data with the aim of estimating causal effects, is adequately accounting for observable and unobservable variables that may operate as potential confounders. The PO framework shows us how to address that concern. But with respect to a number of other, equally important concerns, the PO framework may be insufficient. Among them:

1. *Specifying appropriate counterfactuals.* If we are interested in the effect of X on Y, what is "Not X"?

2. *Generalizability* or *external validity.* Should a causal relationship (or causal mechanism) established in context Z be expected to hold in context Z'? And how broad is the set of contexts in which the relationship can be anticipated to hold? ?

3. *Mechanism(s).* What is the process or processes responsible for the causal relationship between X and Y, and how can we identify (and/or adjudicate among) them?

[2]That practitioners in the PO tradition often do not explicitly state their assumptions concerning operative mechanisms lies at the heart Pearl's 2009 critique of the framework. Pearl argues that such mechanisms ought to be formalized, and that directed acyclic graphs (DAGs) are an appropriate technology of formalization. The response to Pearl by statisticians working in this area is extensive, but beyond the scope of this article.

4. *Observability.*[3] Why might some variables and/or implicit causal relationships be observable or unobservable to begin with? Are there ways to empirically substantiate fundamentally unobservable causal relationships?

In our view, research programs facing the first two of these concerns have no trouble working within the PO framework once the concerns have been addressed, while the framework may be of limited usefulness for those programs facing the last two. What all four share, however, is that their systematic examination requires theory, and addressing the empirical implications of theory often requires examining historical contexts.

Counterfactuals and Generalizability

Theories intended to be taken to data are often most useful when they can specify clear conditions (that naturally point to different real-world contexts) in which a relationship should hold (or not). In a game-theoretic model, explicit assumptions specify the contexts in which a particular equilibrium or comparative static prediction holds. (At the opposite extreme, a theory might involve so many contingencies that it can only apply in one high specific context: historical institutionalist accounts are often limiting cases of this sort of theorizing.) A formal model may also help us understand the most substantively meaningful counterfactual for evaluating a causal effect,

The issue of appropriate counterfactuals is critical when evaluating the causal effects of institutions. Several useful examples of the value of theory in pinning down substantively meaningful counterfactuals can be found in tests of accountability models. Alt, Bueno de Mesquita, and Rose (2011) seek to disentangle selection and incentive effects of elections by exploiting variation in gubernatorial term limits. Their formal model yields the comparisons of greatest substantive interest: between term-limited governors in states with two-term term limits with those in states with three-term term limits to isolate selection effects, and between term-limited and non-term-limited governors with the same tenure of office to

[3]We use the term "observability" rather than "selection" owing to substantial variation across scholarly communities in the use of the term "selection bias."

isolate incentive effects. Another example is Lee, Moretti, and Butler (2004), whose simple theoretical model permits them to derive appropriate comparisons necessary to decompose legislative incumbency advantage estimated from a regression discontinuity design.

Research by Gailmard and Jenkins (2009) highlights the complementarities between historical and theoretical analysis by illuminating the extent to which causal relationships are modified by different institutional contexts. The focus of their study is the consequences of the transition to popular election of US Senators brought about by the adoption of the 17th amendment in 1913. Prior to the adoption of the amendment, senators were selected by state legislators, who were themselves elected by (enfranchised) citizens. The indirect relationship between voters and senators is anticipated to attenuate the responsiveness of senators to the preferences of the mass electorate. Following the adoption of the amendment, responsiveness to voter preferences is anticipated to increase. At the same time, an increase in the information asymmetry between senators and their immediate principals is projected to produce greater discretion for individual senators, resulting in a larger gap between the preferences of a state's two senators.

Note that our discussion of the relevance of theory in identifying appropriate counterfactuals is related, though distinct, from issues raised in Sekhon and Titiunik (2012) in their discussion of natural experiments. The authors note that even under conditions in which some naturally occurring intervention is randomly assigned, assignment is not ignorable because treatment and control units are assigned post hoc and may not be comparable. Sekhon and Titiunik are, in effect, arguing that given a comparison of interest, there may be fundamental limits to using a natural experiment to make it. The sources of these limits are readily explicated using the PO framework. Our point, by contrast, is that we cannot take the comparison of interest as given: it must come from somewhere, and that somewhere is theory, formal or otherwise.

The role of theory in clarifying appropriate counterfactuals or the background conditions in which a causal relationship is more or less likely to hold is readily reconciled with the PO

framework. Once one knows what comparison is substantively appropriate, the task of the researcher is to look for exogenous variation that would permit one to credibly estimate the causal effect of the suitably defined treatment variable. Generalizability issues are similarly reconcilable: given a hypothesis concerning modifier variable Z, our task is to look for conditions Z and $\sim Z$ in which there is exogenous variation in a treatment variable of interest, and determine whether differences in the estimated causal effect under different conditions accords with theoretical expectations.

The extent to which theory can clarify these questions has substantial implications for the importance of quantitative historical analysis. First, as we discussed in the previous section, unless a theory argues that a particular relationship or set of relationships between variables is specific to the present (or to some other moment in time), to be validated the theory not only should but *must* be tested across as many different temporal contexts under which the conditions for the theory are met (or not) as possible. Second, theories that permit historical analysis are (arguably) more useful than those that do not: a theory that is specific to a single moment in time may be helpful in better understanding that particular moment, but it cannot contribute either to the search for broader truths about human political behavior that makes up a substantial part of the discipline of political science or to any governmental policy that will be applied in the future – which is necessarily a new context.

Likewise, if a theory illuminates a particular counterfactual as substantively relevant to the question at hand, under many conditions we may only be able to observe that counterfactual in the past. The question of the effect of direct primaries on polarization is a case in point. (Hirano and Snyder, Forthcoming in 2019, p. 217) write,

> ...we must ask the question, *Primaries compared to what?* Caucuses? Conventions? Party committee meetings? A top-two system? National party meetings? Lotteries? [Evidence suggests that the] main alternatives used in the past – that is, caucuses, party committee meetings, or conventions – might systematically produce nominees who are more extreme than those selected in primaries. In

addition, there are theoretical reasons to believe that these alternatives would be even more polarizing than primaries.

If the only contemporary example of a state party convention system is in Utah, which combines a convention with a primary, then observers contemplating counterfactuals involving selection exclusively by party elites will have no option other than to look backward in time.

Historical research can also be invaluable in addressing the question of generalizability. This is true in two senses. First and most obviously, if a theory anticipates a causal effect holding or not holding under modifier condition Z, then locating a context in which Z obtains may require reaching into the past. Second, perhaps even more critically, comparisons of modifiers that exploit temporal variation within a single context – particularly given a major institutional transition – can enhance the credibility of conditional causal claims in a way that cross-sectional variation cannot. In other words, it is generally more straightforward, as a methodological matter, to demonstrate that an institutional shift in country i from institutional environment $Z_{i,t}$ to $Z_{i,t+1}$ is not confounded by contemporaneous unobservable differences than a comparison of causal effects in contexts $Z_{i,t}$ and $Z_{j,t}$. In the former case, the within-country (or within-state, or within-organization) comparison permits one to hold constant innumerable other features of the environment correlated with the modifier variable of interest, whereas a cross-sectional comparison between two different countries or states at the same time leaves the observer with far more confounders. This is the methodological approach taken in (Gailmard and Jenkins, 2009) discussed above. Similarly, Eggers and Spirling (2014) document changes in the relationship between electoral marginality and legislative roll call and speech making behavior associated with successive historical reforms that brought about greater party cohesion in the British House of Commons. Finally, numerous historical studies provide the temporal variation necessary to exploit one of our most compelling tools for causal identification: differences-in-differences. Diff-in-diff studies permit us to control for all time-invariant features of cross-sectional units, as well time-specific effects, in recovering causal effects of interest.

Mechanisms, Observability, and Information

Suppose, finally, that the scholarly objective at hand is not simply to confirm a causal relationship between X and Y, but rather to document or explicate the operation of a specific causal mechanism or class of mechanisms; or alternatively, to adjudicate among competing mechanisms that might be separately or jointly responsible for an observed political phenomenon. Scholarship of this sort runs into three fundamental challenges for which the PO framework provides at best only a partial solution.

1. Unobservable mediator variables

2. Strategic anticipation of unrealized causal effects

3. Informational treatments

Unobservable mediator variables. Political methodologists working in the PO tradition have grappled seriously with the challenges of assessing mechanisms through causal mediation analysis (for example, Imai et al. (2011), but see Green, Ha, and Bullock (2010)). Mediation analysis – whether properly identified or not – is premised on the observability of the mediating variables. In many political science applications, however, the mediators are fundamentally unobservable. One general class of questions in which unobservable mediation problems are rife concerns discrimination. Generically, suppose individuals may be described by some underlying but unobservable type, and can be partitioned into two groups: A and B. An agent makes some payoff-relevant determination with respect to those individuals – for example, to employ them, to provide some benefit, or to impose some cost. We observe different distributions of outcomes by group: A receives more on average than B, or is punished more on average than B, for example. A fundamental question is: are those differences attributable to differences in the underlying group-specific type distributions – so called "statistical discrimination" – or due to some favoritism or animus on the part of the agent toward one group relative to the other. The analyst's problem is that she is incapable

of looking into the souls of the individuals to ascertain their types, or into the soul of the agent to ascertain their level of favoritism or animus. The mediators between treatment (group status, in this case) and outcome are unobservable.

Suppose, for example, the Department of Justice, when under the direction of a Republican administration, prosecutes Democratic public officials for corruption more frequently than Republican ones. Absent additional information, we cannot say whether the disparity is driven by unobservable partisan bias on the part of the Department's attorneys, or unobservable differences in the levels of corruption of Democratic and Republican public officials (Gordon, 2009). Or suppose women are less likely to hold public office than men. Is this association driven by discriminatory attitudes held by voters; by female potential candidates underestimating their abilities, or something else (Anzia and Berry, 2011)?

Strategic anticipation of unrealized causal effects. When theories describe causal mechanisms arising from strategic, sequential interaction, some causally relevant variables may be "off the path of play:" rationally anticipated and avoided, and thus never observed in equilibrium. Causal relationships driven by such anticipation generate *strategic selection bias* (Schultz, 2001) in the data: we cannot test them directly because the consequences undesirable treatments would induce are anticipated—and avoided—by the relevant actors.[4]

A notable example in political science is the difficulty of estimating the causal effect of the threat of challenger entry on incumbent behavior (Gordon and Huber, 2007): not only is direct causal inference in the PO framework hampered by non-random assignment of challengers; but if incumbents are sufficiently attentive to the threat of electoral competition, they will take pains in office to minimize the viability of a competitor. In fact, if the fear is sufficiently great, then we need observe no variation on the treatment variable of interest, because incumbents will do whatever is necessary to deter a challenger. But it would be erroneous to conclude that the threat is irrelevant from the absence of variation. Another example comes from International Relations: the causal effect of sanctions on behavior. If

[4]For a recent treatment of this issue, see Slough (2023).

target nations take the threat of sanctions seriously, only ineffectual sanctions will observed in equilibrium. But this need not, of course, imply the inefficacy of sanctions.

Informational Treatments. Efforts to estimate the effects of the behavior of an individual or group on the behavior of a different individual or group is complicated by the fact that the causal mechanism of interest concerns *information* conveyed by the behavior of the former (Bueno de Mesquita and Tyson, 2020). In such cases, finding an instrument or source of exogenous variation in the behavior of the former and applying the PO framework uncritically risks shutting down the informational channel: a behavioral treatment, if randomly assigned, conveys no information at all.

For example, consider the electoral consequences of judicial sentencing. We might suppose that voters are more punitive than judges, and that lenient sentences inform voters that judges are ideologically out of step. Supposing that, in line with the orthodox PO approach, we could locate an exogenous source of judicial leniency (thus circumventing the strategic anticipation problem). Under exogenous assignment of judicial leniency, sentences would no longer convey information to the voter about whether the judge was or wasn't out of step.[5] Another example, cited in Bueno de Mesquita and Tyson (2020), concerns estimating the effect of variation in protest size on state repression by instrumenting for size using variation in rainfall (the logic being that rain renders protests more costly and thus smaller). Since the state observes the rain and incorporates rainfall into its assessment of protest size, the use of this instrument does not capture the potential informational effect on the state of observing larger or smaller protests.

These three challenges – unobservable mediator variables; strategic anticipation of unrealized causal effects; and informational treatments – need not imply defeat for the empirical researcher if theory can generate empirical implications *besides* the primary effect of interest, and these implications can in turn be used to substantiate that effect. Theoretical elaboration of an (empirically unidentifiable) effect of interest can give rise to three kinds of such

[5]This variation could, of course, be useful as a placebo test for the mechanism.

implications. Two of these are ultimately amenable to assessment using the toolkit provided by the PO framework.

First, theory may imply additional causal relationships. Recall the judicial sentencing example. Suppose judges value both being reelected and deciding cases fairly. Then they might balance the electoral threat of a lenient sentence against fairness concerns, and we might anticipate that a judge would sentence the same defendant differently in different competitive reelection environments. If in addition to this, judges discount the future (or put differently, voters discount the past), we might anticipate that an elected judge would also sentence the same defendant differently at different points in her term. These implications can be tested using the PO framework. Thus, Huber and Gordon (2004) rely on random assignment of judges with staggered terms in the Pennsylvania court of common pleas to test the first claim. Gordon and Huber (2007) compare sentencing by judges in the same state (Kansas) and assigned close-to-identical cases, who stand for reelection under different systems (partisan competitive and retention).

In a similar vein, if only individuals who expected some return on investment make campaign contributions, then we will not be able to reliably recover the causal effect of contributions on the behavior of elected officials. If, however, sophisticated contributors have rational expectations, then we will expect those whose compensation is more sensitive to potential changes in government policy will make larger contributions, ceteris paribus. This is the approach taken by Gordon, Hafer, and Landa (2007), who document a significant relationship between the campaign contribution behavior of corporate executives and their pay-performance sensitivity.

Second, theory may reveal that the causal relationship of interest will induce other *non-causal* correlations that are observable and, thus, testable. Recall the question of whether disproportionate prosecution of Democratic public officials by the Republican Department of Justice is due to Democratic corruption or Republican bias. Suppose in the general population of public officials, corruption and membership in the Democratic party are uncorrelated,

but the DOJ indicts individuals on the basis of both traits. Then the two traits will be (negatively) correlated in the selected sample. This form of selection bias is sometimes called Berkson's paradox, and in the more recent nonparametric causal identification literature is cited as an example of "conditioning on a collider" (Elwert and Winship, 2014). Here, deliberately conditioning on a collider generates a *causally-relevant non-causal prediction* (Ashworth, Berry, and Mesquita, 2015): Democrats will be less corrupt, and thus have better case outcomes, than Republicans, conditional on being prosecuted for corruption by a Republican Department of Justice. The implication, of course, is not that membership in the Democratic party *causes* lower corruption, but that unobservable bias can generate a correlation between these traits. Since if Democrats actually *were* more corrupt this correlation would not emerge, by looking for this correlation we can determine whether partisan bias played a role in prosecution decisions by the Republican-led DOJ.

Finally, theory may generate *implied point or interval predictions* about the value or distribution of an observable variable. Thus, the Downsian model of electoral competition anticipates candidate convergence to the position of the median voter; accountability models a large incumbency advantage; and full-information veto bargaining models no vetoes in equilibrium. As Morton (1999, ch. 6) notes, since theoretical models are rarely exhaustive accounts of data generating processes, empirical rejection of categorical predictions should be taken with a grain of salt. Nonetheless, verification or rejection of these predictions can still be informative with respect to the operation of a particular mechanism.

In many empirical papers, of course, both relatively straightforward and unobservable causal relationships are investigated, and understanding the mechanisms behind a particular relationship, or the degree to which the relationship is generalizable, is as important as establishing the primary causal effect of interest. Here, researchers must toggle back and forth between direct application of the PO framework, theory to elucidate other observable implications where direct application of the framework is not feasible, and empirical testing for these implications either via the PO framework or via other empirical analytical tools.

We provide several examples of this iterative process in historical empirical papers in the applications below.

Applications

The Effect on Legal Emergence of Political/Fiscal Shocks in 12th Century Britain

Our first example illustrates how general causal theories might be tested with the PO framework by exploiting historical natural experiments — and how the mechanisms underpinning them may be validated by identifying and testing for additional, observable predictions that are consistent or inconsistent with these mechanisms. Consider research on the emergence of legal institutions. A prominent theory of legal development holds that for property protections to develop in a nascent state, the state's ruler must have a long time horizon and be willing (and able) to collect taxes. The intuition is as follows: a self-interested ruler will only bother to provide costly, consistent protection from outside predators (and limit his own predation) if he expects to benefit in the long run (via taxes) from the economic growth this strategy promotes.

While this theory is highly plausible, one of arguably the most influential institutions in the world—the English common law system—began its development during a time of substantial political instability, where rulers' time horizons were short. How did a robust system of property protection emerge in such an environment? One potential explanation is that legal systems are valuable in their own right: they can be used to generate state revenue via fees and fines, or to build regime support by strategically assigning rights.[6] These benefits may explain legal institutions' development during at least one type of political insecurity: when a ruler faces internal threats to his survival. Although political threats in general are thought to motivate fiscal capacity building, rulers facing domestic threats may fear the political consequences of raising taxes and perceive legal fees and fines as less risky because

[6] Anecdotally, an example of the former effect might be the use in the U.S. by local governments of court fee and fine increases, rather than tax increases, to meet revenue shortfalls (Goldstein, Sances, and You, Forthcoming). An example of the latter effect might be the selective assignment of property rights to regime supporters (Onoma, 2010).

they target isolated individuals or are rendered in exchange for a service. Similarly, under conditions of internal political threat, the ability to provide legal privileges and rights may be especially important to maintaining a supportive coalition.

If this theorized causal relationship holds, increases in a ruler's domestic political insecurity ought to cause an increase in his investment in, and use of, the state legal system to collect revenue and/or to generate support. Since we might think that increased use of the state legal system might also affect, or covary with, political security, testing this relationship requires finding an exogenous source of variation to domestic political security. In addition to finding a case where we can measure the causal effect we have hypothesized, we might also care about finding a case that is important in its own right. Since there is nothing to suggest that this relationship should not be observed across a number of times and places, we may wish to look to the past to test it, both to broaden the parameters of our search for exogenous variation in domestic political security and since, substantively, the development of many legal institutions occurred in the past.

In Simpson (2021), one of us attempts just that, focusing on the canonical case of the emergence of the English common-law system in the 12th century and leveraging an interesting natural experiment that occurred during this period: the kidnapping for ransom of King Richard I of England in 1192. Because the kidnapping resulted from a series of mishaps, beginning with King Richard's shipwreck at sea, and precipitated an attempted rebellion by Richard's brother John, it produced a completely unanticipated exogenous shock to the internal political stability of England as well as to the state's expenditure needs. Moreover, the event permits an adjudication of the degree to which revenue generation and support building relatively motivated post-kidnap deployment of the Royal Court through an exploration of the local effect of different "doses" of political insecurity, since English counties varied in their proximity to military strongholds of John's and thus to imminent political threat.

A basic examination of Royal Court activity post-kidnap indeed reveals variation by

county threat level: immediate post-kidnap increases in court activity occurred only in high-threat counties, suggesting that differential political insecurity did play some role in court deployment. More is needed, however, to assess whether these changes were due to attempts by the Court to build support for the regime, to generate revenue, to do both, or to do neither. With respect to this last possibility, other kidnap-related mechanisms that might generate the same outcome include a post-kidnap breakdown in law and order in high-threat counties, or post-kidnap migration away from John's military strongholds into high-threat counties. These mechanisms each generate observable, testable implications: a breakdown in law and order that motivated court activity would have generated an increase in court cases involving law and order offenses (it did not); or mass migration would have been recorded in the English administrative data (it is not).

To adjudicate between the revenue-generation and support-generation hypotheses, Simpson next looks to test observable implications specific to each. A legal system deployed to raise revenue might raise fees and fines, change its focus to more profitable types of cases and/or to assisting richer litigants, and might increase pressure for repayment on individuals who owed money to the Court. By contrast, a legal system deployed to build support might lower fees and fines, change its focus to more in-demand or socially useful types of cases, expand access to poorer litigants, and decrease pressure for repayment on debtors. Testing these implications yields the finding that in high-threat counties, the Royal Court did indeed aim to build support for the regime: it changed its focus to cases involving property rights protection and property litigation, decreased its fees and fines, and generally appears to have expanded litigant access. In other counties, however, there is some evidence that the Royal Court may have been attempting to generate additional revenue. These findings suggest a primary focus on building political support in highly threatened areas, with a secondary focus on generating revenue in safer areas.

After leveraging the PO framework to derive evidence that legal systems are indeed used to build political support and generate revenue in response to domestic political insecurity,

Simpson expands her focus to a twenty-year period around the kidnap in order to understand whether the kidnap had long-term effects on the Royal Court's evolution in England. She finds that beginning after the kidnap and continuing for at least the next ten years, the Royal Court's level of activity, caseload, efficiency, and accessibility all substantially increased, as did the English state's investment in the Court. Although the incompleteness of tax records in the years immediately around the kidnap prevents a causal analysis of the relative effect of the kidnap on taxes and court revenue, non-causal evidence does support the theory that domestically insecure rulers might prefer to generate revenue via court fees and fines: throughout the twenty-year period surrounding the kidnap, state income from court fees and fines was substantially higher than income from feudal dues and taxes.

Local Appropriations in the 19th Century U.S.

Our second example illustrates how scholars may attempt to adjudicate between different causal mechanisms even when the PO framework is of only partial utility. An extensive literature in U.S. politics concerns the geographic allocation of federal appropriations, and there are two possible causal mechanisms underlying the political antecedents of this allocation. The first mechanism emerges from a broad class of distributive politics models whose essential elements are as follows:

1. Voters are divided into N districts, indexed by i.
2. The benefits of spending on district i are linear in spending in that district and enjoyed exclusively by those residing within it.
3. The costs of spending are linear and shared by all N districts.
4. Some agenda setter proposes a vector of allocations, subject to a budget constraint. The legislature votes on this vector, via majority rule or qualified majority rule.

There are many variants of this model. But rather than pin down the specifics (e.g., who is the agenda setter), we can already think of some observable implications of this family of mechanisms.

1. Let q denote the fraction of geographic constituencies required to pass a local spending bill. Then the fraction of geographic constituencies with net positive benefits from the legislation must be weakly greater than q.

2. A weakly dominant strategy for a legislator is to vote for a bill if and only if her constituents receive net positive benefits from the bill.

3. Suppose we model agenda power via a vector of recognition probabilities. Then legislators with more agenda power (i.e., higher recognition probability) can anticipate more spending in their districts than those with less agenda power, ceteris paribus.

A few things to note about these predictions. First, even though none of the three we enumerate constitutes a direct causal claim, validation of these predictions would provide evidence for the operation of a specific family of causal mechanisms. Second, only the third predicts a correlation, and it is with respect to this prediction that the PO framework can be most easily applied: because the prediction holds only 'ceteris paribus' it is important, in testing it, to account for potential confounders. Thus we may, for example, apply legislator fixed effects to establish the extent to which members of a relevant appropriations subcommittee enjoy disproportionate largesse at the federal trough. (Note that our inability to detect disproportionate benefits would not by itself invalidate the family of mechanisms, because the family is consistent with uniformly distributed recognition rights. It would, however, count as evidence of a subset of mechanisms in which certain committees exercise dominance in the appropriations process.) Finally, the first and second predictions clearly also play a role in either substantiating or falsifying the family of mechanisms. This is so even though the PO framework does not easily apply to point/interval predictions, and even though the predictions generated by different models associated with the family may vary in important ways. For example, the fraction of constituencies directly benefiting from local appropriations may be equal to or larger than q (corresponding to predictions of minimum-winning, oversized, or universal coalitions) depending on the conditions of uncertainty included in the model.

But if an examination of votes and benefit allocations in some context reveals that *fewer* than q districts routinely benefit from spending – in a bill or session or Congress – we've uncovered an empirical fact that is inconsistent with the point predictions generated by *all* the models consistent with the family of mechanisms. Likewise, if we see a legislator voting *against* a bill that provides net benefits to his/her district, this too would present evidence against the operation of the mechanism.

Now consider a second mechanism, which we label the partisan-ideological model, whose elements are as follows:

1. Legislators have ideological or partisan commitments for or against a particular vision of what the national government ought to be doing to spur economic development.

2. Spending on individual development projects are a manifestation of that vision.

3. Geographic location matters to the return on investment from specific types of local spending.

4. Spillovers from projects are substantial, but their magnitude need not be reducible to geographic proximity.

What observable implications are consistent with this mechanism?

1. Appropriations bills in which spending is directed to a minority of districts.

2. Legislators voting for bills that don't directly benefit their districts, or voting against bills that do.

3. Support for those bills correlated with partisanship or ideology, *independent of majority control*.

4. District fixed effects accounting for a high proportion of variation in spending.

The last two of these implications constitute non-causal (or causal) correlations to which the PO framework may be applied, while the first two constitute point/interval predictions. Moreover, as discussed above, findings inconsistent (or consistent) with the first two point/interval predictions can sometimes provide definitive evidence for or against the operation of the mechanism, even without application of the PO framework. For example, while spending bills benefiting a majority of districts would be consistent with both mechanisms, bills benefiting a minority would be consistent with only the ideological mechanism. Likewise, observing legislators voting for bills that benefit their districts is consistent with both mechanisms, whereas legislators voting for bills that don't, or against bills that do, is consistent only with the partisan/ideological account. Note that the knowledge produced in adjudicating between these mechanisms need not be definitive: the exercise will not conclusively demonstrate that one is correct as much as that the other needs to be rethought or the theory revised. But our beliefs concerning the set of potentially operative causal relationships will almost certainly have shifted by the conclusion of the empirical exercise.

In Gordon and Simpson (2018), we adjudicate between these competing accounts of local appropriations using data on over 9,000 specific federal appropriations on local projects in the 19th Century. Why is the 19th century the appropriate venue to test these accounts? There are several reasons, beyond intrinsic interest in the period. First, many influential models of distributive politics were formulated with the 19th century Congress in mind, so assessing the accuracy of the model amounts to testing the original conjecture. There is, moreover, a sense in which this context presents an easy case for distributive models and a hard case for partisan models. Second, there are a number of complicating factors that arise in later periods that are absent in this one: specifically, Congress tended during this period to appropriate for specific projects (albeit often in consultation with the Army Corps of Engineers), rather than to appropriating for general purposes and delegating responsibility for project allocation to the bureaucracy.

Our findings favor the operation of the partisan/ideological mechanism (as currently

conceived) until the early 1870s, and the distributive mechanism (as currently conceived) afterward. In the first three-quarters of the 19th century: (1) nearly all congressional sessions were characterized by local appropriations to a minority of districts; (2) legislators frequently voted for bills that didn't directly benefit their districts, and some voted against bills that did; (3) support for local appropriations bills correlated with ideology; (4) local appropriations were uncorrelated with legislative committee status of members, majority status, or electoral marginality; and (5) geography explained a large proportion of spending variance. Starting in the 1870s, however, (1) district supermajorities began receiving local appropriations; (2) legislative supermajorities began voting in favor of local appropriations bills; and (3) support and opposition in non-unanimous votes ceased to be correlated with ideology (as captured by first dimension DW-Nominate scores), despite the fact that ideology was an increasingly strong predictor of positions on other contemporaneous legislation.

In the last part of the paper, we consider possible reasons for the shift. This exercise is speculative, but points in the direction of future tests. We identify two specific changes in the background conditions that may have suppressed one mechanism and amplified another. The first was Southern Democratic support for local appropriations brought about by the destruction wrought by the Civil War. The second was a process of accretion over the preceding century: even if only a small minority of districts benefited from local appropriations in a specific Congress, given enough time a supermajority of districts would ultimately have some local project, and thus potentially an organized constituency for its continued upkeep. This latter factor is consistent with a variant of the electoral connection (legislator responsiveness to organized consistuencies) but would require a revision to the existing theory.

Conclusion

We have endeavored to place an extremely valuable set of empirical tools embodied by the potential outcomes framework into a broader conceptual framework that admits a variety of causally relevant questions beyond the simple question of "Does X cause Y," and described how this broader framework often points in the direction of historical applications as a

route to more general political knowledge. In doing so, we have articulated a number of considerations for which theoretical guidance is essential: specifying correct counterfactuals; specifying contexts in which causal relationship will and will not be expected to operate; specifying mechanisms of action and adjudicating among them; and considering how to assess causal relationships that are in principle unobservable. Each of these questions varies in the way in which it interfaces with the potential outcomes framework. With this broader conception of causal inquiry in hand, we explain the essential value of historical research in establishing substantively relevant causal relationship of interest to the broad community of political scientists. Applications from our own work display the different varieties of causal inquiry described in the paper.

Much of the discussion here is preliminary. What is needed, and what would be an appropriate next step, is a formal language to describe the interface of theory, the potential outcomes framework, and the generalizability of historical findings more precisely. One promising avenue along these lines is the non-parametric approach described in Pearl (2009), under which some of the issues presented here have received formal elaboration: generalizability (Pearl and Bareinboim, 2014) and endogenous selection (Elwert and Winship, 2014), for example. We leave the specifics of such a formalization as a direction for future research.

References

Alt, James, Ethan Bueno de Mesquita, and Shanna Rose. 2011. "Disentangling Accountability and Competence in Elections: Evidence from U.S. Term Limits." *The Journal of Politics* 73 (1): 171–186.

Anzia, Sarah F., and Christopher R. Berry. 2011. "The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?" *American Journal of Political Science* 55 (3): 478–493.

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

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

Bueno de Mesquita, Ethan, and Scott A. Tyson. 2020. "The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior." *American Political Science Review* 114 (2): 375–391.

Eggers, Andrew C., and Arthur Spirling. 2014. "Electoral Security as a Determinant of Legislator Activity, 1832-1918: New Data and Methods for Analyzing British Political Development." *Legislative Studies Quarterly* 39 (4): 593–620.

Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40 (1): 31–53.

Feigenbaum, James J, James Lee, and Filippo Mezzanotti. 2018. "Capital Destruction and Economic Growth: The Effects of Sherman's March, 1850-1920." NBER Working Paper No. 25392.

Gailmard, Sean. 2017. "Building a New Imperial State: The Strategic Foundations of Separation of Powers in America." *American Political Science Review* 111 (4): 668–685.

Gailmard, Sean, and Jeffery A. Jenkins. 2009. "Agency Problems, the 17th Amendment, and Representation in the Senate." *American Journal of Political Science* 53 (2): 324–342.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2014. "The Illusion of Learning from Observational Research." In *Field Experiments and their Critics*, ed. Dawn Langan Teele. New Haven, CT: Yale University Press pp. 9–32.

Goldstein, Rebecca, Michael W. Sances, and Hye Young You. Forthcoming. "Exploitative Revenues, Law Enforcement, and the Quality of Government Service." *Urban Affairs Review* .

Gordon, Sanford C. 2009. "Assessing Partisan Bias in Federal Public Corruption Prosecutions." *American Political Science Review* 103 (4): 534–554.

Gordon, Sanford C., Catherine Hafer, and Dimitri Landa. 2007. "Consumption or Investment? On Motivations for Political Giving." *Journal of Politics* 69 (4): 1057–1072.

Gordon, Sanford C., and Gregory A. Huber. 2007. "The Effect of Electoral Competitiveness on Incumbent Behavior." *Quarterly Journal of Political Science* 2 (2): 107–138.

Gordon, Sanford C., and Hannah K. Simpson. 2018. "The Birth of Pork: Local Appropriations in Americas First Century." *American Political Science Review* 112 (3): 564–579.

Green, Donald P., Shang E. Ha, and John G. Bullock. 2010. "Enough Already about Black Box Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose." *The ANNALS of the American Academy of Political and Social Science* 628 (1): 200–208.

Hall, Andrew, Connor Huff, and Shiro Kuriwaki. 2018. "Wealth, Slaveownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War." *American Political Science Review* 113 (3): 658–673.

Hirano, Shigeo, and James M. Snyder. Forthcoming in 2019. *Primary Elections in the United States*. Cambridge University Press.

Hirano, Shigeo, James M. Snyder Jr., Stephen Ansolabehere, and John Mark Hansen. 2010. "Primary Elections and Partisan Polarization in the U.S. Congress." *Quarterly Journal of Political Science* 5 (2): 169–191.

Huber, Gregory A., and Sanford C. Gordon. 2004. "Accountability and Coercion: Is Justice Blind When It Runs for Office?" *American Journal of Political Science* 48 (2): 247–263.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105 (4): 765–789.

Jacobson, Gary. 2004. *The Politics of Congressional Elections*. 6th ed. New York, NY: Pearson.

Lee, David S., Enrico Moretti, and Matthew J. Butler. 2004. "Do Voters Affect or Elect Policies? Evidence from the U. S. House*." *The Quarterly Journal of Economics* 119 (3): 807–859.

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

Onoma, Ato Kwamena. 2010. *The Politics of Property Rights Institutions in Africa*. Cambridge: Cambridge University Press.

Pearl, Judea. 2009. *Causality*. New York, NY: Cambridge University Press.

Pearl, Judea, and Elias Bareinboim. 2014. "External Validity: From Do-Calculus to Transportability Across Populations." *Statistical Science* 29 (4): 579–595.

Pierson, Paul, and Theda Skocpol. 2002. "Historical institutionalism in contemporary political science." *Political science: The state of the discipline* 3: 693–721.

Rosenbaum, Paul R. 1999. "Choice as an alternative to control in observational studies." *Statistical Science* pp. 259–278.

Schultz, Kenneth A. 2001. "Looking for Audience Costs." *Journal of Conflict Resolution* 45 (1): 32–60.

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

Simpson, Hannah K. 2021. "Justice for sale: political crises and legal development." *Political Science Research and Methods* 9 (4): 779–799.

Skowronek, Stephen. 1982. *Building a new American state: The expansion of national administrative capacities, 1877-1920.* New York, NY: Cambridge University Press.

Slough, Tara. 2023. "Phantom Counterfactuals." *American Journal of Political Science* 67 (1): 137–153.

Weaver, Vesla M. 2007. "Frontlash: Race and the Development of Punitive Crime Policy." *Studies in American Political Development* 21 (2): 230–265.