Causal Inference and Knowledge Accumulation in Historical Political Economy[1]

Anna Callis, Thad Dunning, and Guadalupe Tuñón

Word count: 7,511 (without abstract/references)
9,330 (with abstract/references)

ABSTRACT: Empirical scholarship on historical political economy (HPE) has been greatly influenced by the so-called credibility revolution. Critics rightly worry, however, about the revolution's capacity to aid explanation, as opposed to the estimation of treatment effects. We describe three empirical strategies that we argue can foster stronger accumulation of knowledge about generalizability and causal mechanisms, and we survey their use in HPE, especially in research on the consequences of European colonial expansion. Grounded on design-based approaches, these strategies offer leverage for explanation and provide avenues for the development and testing of generalizable theory. We see no necessary tension between the credibility revolution and the construction of a comprehensive, cross-case evidence base. However, it would be useful if strategies for cumulative learning—such as those we describe—were leveraged more explicitly and purposively across studies.

---

[1] Prepared for inclusion in the The *Oxford Handbook of Historical Political Economy*, edited by Jeff Jenkins and Jared Rubin. Author affiliations: University of California, Berkeley (Anna Callis and Thad Dunning) and Princeton University (Guadalupe Tuñón).

Empirical research on historical political economy (HPE) has been transformed over recent decades by what is sometimes called the "design-based" turn. Historical natural experiments, instrumental-variables and regression-discontinuity designs, and especially different varieties of difference-in-differences methods are the workhorses of much contemporary research on politics and economics in historical settings.[2] For many quantitatively oriented scholars, these methods have replaced a previous reliance on the selection-on-observables assumptions made in applications of multivariate regression and related approaches. These days, claims to valid causal inference in quantitative work often rely on some plausible as-if random assignment to a treatment variable or some means of testing key identifying assumptions, for example, through placebo tests.[3] The so-called "credibility revolution" appears firmly established in HPE: the revolutionaries are practically retirees.[4]

Yet many scholars voice unease—we think rightly—about what design-based empirical research can contribute to our general knowledge of historical political economy. Sometimes, this discomfort is cast in terms not only of a paucity of theory but also in terms of the kinds of theory that a focus on as-if random assignment can help to test. Rozenas (2021), for example, suggests that prioritizing the role of as-if random for empirical assessment may give pride of place to contingent factors as key causes in quantitative HPE. Yet the treatments that may really matter for outcomes about which we care may be sticky, very non-random, and quite difficult to manipulate. Other times, the worry is more about the difficulty of extrapolating findings beyond the particular design features that allow for credible design-based causal inference in specific, sometimes apparently idiosyncratic, settings. Gailmard (2021a,b), for instance, argues that achieving the important goal of extrapolation from empirical findings in a particular case depends, first, on theory and second, often on "weaker" methods relying mainly on selection-on-observables assumptions. Yet, if theory and those weaker methods are sufficient for extrapolation to a new case, he argues, then they may be sufficient for inference in the original setting in which the design-based method was used—casting some doubt on the utility or inferential priority granted to the tools of the credibility revolution in the first place.

These concerns about the contributions of design-based methods are multifaceted.[5] As a whole, though, we think they often relate to the capacity of the credibility revolution to generate *cumulative learning*, also known as the accumulation of knowledge. While internal validity is a key part of learning, two other dimensions of causal assessment seem especially critical: (i) the *generalizability* of causal claims and findings; and (ii) evidence on the *mechanisms* that connect causes to effects. These topics are not identical, but they are certainly related. As many philosophers of science have underscored, mechanisms can be critical for generalizability. For example, knowing the mechanism that led cause X to produce effect Y in context A shapes expectations about whether we would expect a similar effect in context B, where the key

---

[2] See, e.g. Angrist and Krueger (2001); Diamond and Robinson (2010); Cantoni and Yuchtman (2021); or Valencia Caicedo (2021).
[3] On placebo tests, see e.g. Eggers et al. (2022).
[4] On the revolution and the revolutionaries, see Angrist and Pischke (2010).
[5] The utility of these methods is the subject of longstanding debates in both political science and economics; see e.g. Deaton (2010), Imbens (2010), and Angrist and Pischke (2010).

mechanism may or may not be operative.[6] We do not presume that cumulative learning is always the key goal of HPE; sometimes, simply assessing treatment effects in a particular case or historical event, by means of a single discrete study, is highly valuable. For political scientists and economists, though, the accumulation of generalizable knowledge about mechanisms does often appear critical, so we take this as a central if sometimes implicit goal.

To the extent that cumulative learning is the–or an important—goal for HPE, key issues seem to revolve around (a) whether or to what extent any empirical method can aid the accumulation of knowledge about generalizability and mechanism; and (b) how well design-based methods for causal inference can do this, relative to other tools. Our goal in this chapter is to engage these questions. For (a), we highlight the general difficulties of assessing generalizability and mechanism, whatever the method. The challenge of cumulative learning about generalizability or mechanism is also not an obvious function of method, however. Sometimes, the seeming advantages of "selection-on-observables" types of methods can be only apparent.

This takes us to (b). While there are important challenges to cumulative learning from design-based approaches, we highlight the useful, if necessarily limited, routes to generalizability and mechanism that can be—and increasingly have been—achieved using methods associated with the credibility revolution. We lay out several strategies for knowledge accumulation about generalizability and mechanism and show how they have or can be leveraged, implicitly or explicitly, in historical political economy. However, we also argue that greater explicit attention to these goals in the design and interpretation of studies would aid cumulative learning. We consider especially research on European colonialism and the global expansion of the West, including work on the slave trade, extractive institutions of forced labor, and the mode of colonial rule (e.g. indirect vs. direct), as well as related questions about the impacts of missionary activity, and we discuss work on internal migration and technological innovation. Across these different areas, scholars have addressed questions of generalizability by using related designs to examine similar questions in different contexts. The comparison of findings from related studies with similar treatments and different outcomes, or with variation in classes of a treatment variable, can also shed light on operative mechanisms. Moreover, the credible evidence base is critical for developing, disciplining and validating generalizable theory and thus can provide a route to knowledge accumulation.

To clarify our earlier "we think rightly" in this chapter's second paragraph: we think that critics are correct to raise concerns. Yet we also think the challenge of assessing generalizability and mechanism is substantial, whatever the empirical method and whatever the theoretical focus. Theory can narrow the range of plausible mechanisms consistent with an empirical finding; but we think it is critical for an empirical field such as HPE to consider and advance empirical routes to generalizability and mechanism. Our point is not to suggest that the modes of cumulative learning we highlight represent the only or even always best route to cumulative learning. Our goal, however, is to highlight ways in which the apparently stark conflict between achieving internal validity in particular contexts using design-based methods and understanding of generalizability and mechanism can be overstated. Greater attention to generalizability and the

---

[6] Cartwright and Hardie (2010).

ways in which different studies may replicate and build on one another can help to soften the tradeoffs between internal and external validity, or between the estimation of treatment effects and the understanding of mechanisms. This can foster more useful knowledge accumulation.

THE CREDIBILITY REVOLUTION, GENERALIZABILITY AND MECHANISMS

Cumulative learning about generalizability and mechanism may not always be the explicit or even the implicit goal of historical political economy. Even where scholars care about generalizable lessons, assessing the validity of inferences about causal relations in a particular case or event may rightly take priority. If we are not even right about the historical consequences of the Glorious Revolution—to take just one example—then extrapolating lessons about the effects of parliamentary checks on royal authority for investment, government borrowing, or economic development in general might seem besides the point.[7]

Yet, there are many reasons that historical political economy as a field focuses or should focus, either implicitly or explicitly, on generalizability and mechanism. Whether empirical work is seen as an avenue for developing new theoretical insights or as a means for testing theory, the question of how findings apply beyond the setting at hand is often implicit, if not explicit. Substantive work in HPE is also often–though of course not always—especially concerned with empirical phenomena that extend beyond a particular case, in the sense that many cases may be shaped in similar ways by similar phenomena. Indeed, claims to generalizable knowledge may help to define HPE as a field and in some ways set it off from cognate work in the discipline of history. How, then, can we best make progress in this area?

The assessment of generalizability and of mechanism relies critically on theory, a point made well for example in the excellent recent article of Gailmard (2021a); see also, inter alia, Ashworth et al. (2021). A mechanism itself can be seen as a theoretical construct of a social process that causes an event rather than as an intervening variable (Waldner 2016). This does not gainsay the role of observables, however. We can call a "mediating variable" a mechanism, an indicator of a mechanism, or a variable that would take on a particular value if a specific theory of mechanism were true without doing undue violence to interpretation. What seems critical is (at least) to assess whether particular theories are consistent or inconsistent with specific patterns of evidence. While many potential theories may help to explain a particular observed set of facts, not all theories can; and multiplication of the relevant facts sharpens the theoretical challenges but also the potential insights. Theories can also generate new observable implications that can be tested against additional evidence from a case, a perspective echoed in diverse work on within-case process tracing (e.g. Collier, Brady, and Seawright 2010; Collier 2011) as well as other advice on theoretical development in relation to qualitative evidence (e.g. King, Keohane and Verba 1994).

---

[7] North and Weingast (1989); for critiques, see Stasavage (2002; 2007) or Pincus and Robinson (2014), among others.

Empirical routes to the evaluation of generalizability and mechanism are thus essential in any field. They are perhaps all the more so in HPE, focusing as it does definitionally on concrete temporal processes that have unfolded in "the past." Without empirical means of validating claims to generalizability, HPE could risk getting not only the political economy but also the history wrong. Thus, here we are explicitly concerned with empirical routes to the identification of mechanisms and generalizabilability—what could be called an inductive approach—even though theory can and should play a critical potential role in identifying classes of important mechanisms. We ask how empirical methods can best work together with theory to aid the accumulation of valid knowledge about mechanisms and generalizable causal relations in HPE. In other words, we do not take the central task to be the identification of the relevance of findings in one case for likely findings in a case that has not yet been or will not be studied empirically (Gailmard 2021a). Rather, we aim to explore how empirical research can best aid—and help to validate—theory and inference about generalizability and mechanism across a variety of cases.

From this perspective, one role of strong design is to provide believable "stubborn facts" that theory can seek to reconcile—while sometimes also generating new lines of testable inquiry. The "believable" part is perhaps the key focus of the credibility revolutionaries, dating at least from Leamer (1983). However, this point has most often been made with respect to the internal validity of causal claims in a particular case or event.[8] In this context—notwithstanding debate about the virtues of particular designs in specific studies—the potential usefulness of design-based methods is by now fairly widely extolled. We largely agree with the positive view, even as we also think that different inferential tools and methods are appropriate for different kinds of problems; and even as we recognize that there are many difficulties with natural experiments, discontinuities, and the like.

What seems more controversial is whether and to what extent such methods can contribute to the accumulation of knowledge about the generalizability of causal findings and the mechanisms that undergird them. By the accumulation of knowledge—which can also be called cumulative learning—we mean valid, useful, or correct insights, e.g. about the working of the social and political world, that build on or depend on one another, often (though not always) across distinct studies conducted by different researchers.[9]

A simple—but we think essential—point that we make here is that for purposes of assessing generalizability and mechanism, the same reciprocal relationship between credible evidence and theoretical development that we sometimes find within cases should apply across cases. Yet, what are good methods for advancing this reciprocal relationship?

---

[8] We use "internal validity" as Campbell and Stanley (1963: 5) did with respect to quasi-experiments: "the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance?"
[9] For discussion and definitions, see e.g. Mahoney (2003); Dunning et al. (2019); or Callis et al. (2022).

a. Challenges of assessing generalizability and mechanism

It is useful to start with the recognition that assessing both mechanisms and the generalizability of findings empirically can be extraordinarily hard—whatever the method. With respect to mechanisms, the difficulties are underscored by the critical literature on path models, post-treatment bias, and mediation analysis. Even in a randomized controlled experiment where researchers manipulate the treatment, controlling for post-treatment variables—a.k.a. mediators—destroys the expected symmetry between treatment and control groups (Angrist and Pischke 2009: 64-68). Unobservables that influence a mediator are also likely to influence the outcome, leading to bias in Baron and Kenny-type analyses in a regression framework (Glynn 2012; Bullock and Green 2021); and the response schedules implied by path models often involve other untenable assumptions (Freedman 2009: 94-101). In a more flexible potential outcomes framework, even the definition of direct and indirect effects involves imagining impossibilities that are inherently unobservable (as distinct from counterfactual), such as "complex potential outcomes" (Gerber and Green 2012: 329) Identifying assumptions such as sequential ignorability may be so strong as to be untenable in applications (Imai, Keele and Yamamoto 2010). Moreover, direct manipulation of a mediator does not necessarily identify the indirect effect of a treatment working *through* a mediator.

Against these kinds of arguments with respect to randomized experiments—where the possibility of successful mediation analysis might initially appear plausible—the empirical assessment of mechanisms in HPE applications could seem like hubris. In the domain of observational studies, neither treatments nor mediators are manipulated; and many other threats to inference arise. We agree with Gailmard (2021a) and others: empirical studies can rule out some mechanisms and provide support that a given channel is "a" mechanism, yet they are often unlikely to identify "causal uniqueness" when it comes to mechanisms or successfully to parse a treatment's direct and indirect effects.

Assessing generalizability is also highly challenging in HPE, as in other areas. Researchers sometimes advance the idea that we get more bang for the buck from selection-on-observables assumptions. It is common, for example, to contrast the "local" average treatment effect estimated in, say, an instrumental-variables or regression-discontinuity design with some presumably more general or non-local causal effect estimated under a selection-on-observables assumption—such as a research design using a cross-national regression. However, this advantage is often more apparent than real. One familiar but important point is that if confounding or selection effects lead to misleading estimates in the more comprehensive dataset, the findings may simply distort—rather than generalize—our understanding of causal relationships. The apparent generality can also be undercut by the mechanics of modeling selection effects. A country fixed-effects analysis regression upweights units with greater within-country (over time) variation in treatment status. As Aronow and Samii (2016) nicely show, this results in an "effective sample" that can differ sharply from—and that is substantially less "general" than—the full study group, undercutting the apparent virtues for generalizability of a more comprehensive dataset.

b. Empirical strategies for cumulative learning

Despite these difficulties, we think that real—if limited and modest—progress can be made towards understanding mechanisms and generalizability through a combination of mixed methods, qualitative information, theory, and research design. Our goal is to lay out empirical strategies that we think are plausible; to show how elements of those strategies are in practice today; and to propose some potential modifications to current practice that may bolster generalizable learning about mechanisms.

As we have explored elsewhere (Callis et al. 2022), at least three strategies appear broadly important as empirical (or "inductive") routes to cumulative learning about causal relations:

(1) replicating a similar study design across different contexts or with different populations;

(2) varying a component of a bundled treatment while measuring the same outcome of interest; and

(3) maintaining the same treatments while examining different outcomes, or the same outcome at different points over time.

These strategies are relatively "design-based" in that they do—or potentially can—rely on variation akin to that which could—hypothetically—be introduced by an experimental researcher.[10] Response schedules and modeling assumptions play a role, but to a much less constricting degree than the more heavily "model-based" assumptions associated with path models and other modes of formal mediation analysis. Of course, in the observational studies that characterize HPE, there is no experimental intervention; yet an analysis may take advantage of such potential manipulations in a way that is relatively simple and that avoids many of the challenges to inference associated with, e.g., path models and formal mediation.

Here, (1) is clearly most central to the assessment of generalizability, while (depending on the aim), (2) and (3) could be used to assess either generalizability or mechanism (or both). With respect to strategy (1), it is very substantially an empirical question whether an effect found in one context replicates in another. If one finds an effect of treatment X on outcome Y in context W, one may be able to assess empirically whether such an effect also holds in context Z. In principle, one can do this without any insight into the operative mechanism that engendered the effect in context W or any theoretical understanding of whether we would expect a similar mechanism to be active in context Z. We would argue that even if potentially atheoretical, such an effort contributes to a base of evidence that is important for cumulative learning—and certainly, for a historically and empirically grounded route to the accumulation of knowledge.

However, the replication effort implied by strategy (1) may result in different kinds of conclusions. One possibility is that a treatment effect found in context W is broadly similar or goes in the same direction in context Z (and perhaps contexts N, O, P, and Q as well, if we are lucky enough to have multiple studies of a similar phenomenon in different contexts). We give examples of this kind of finding from diverse HPE literatures below. Even if such an empirical

---

[10] Bullock and Green (2021); Gerber and Green 2012: 322-25.

finding does not advance our theoretical understanding of how X produced Y across diverse contexts, the set of findings is valuable on its own: the results may speak directly to the external validity of a historically important treatment. A different possible conclusion, however, is that we observe distinct effects of treatment X on outcome Y in contexts W and Z (and perhaps in N, O, P, and Q as well)--for example, the direction of the effect differs, such that we estimate positive, negative, and null effects in distinct contexts. This especially amplifies the onus on *explanation*.

Understanding why a cause has an effect—an important aspect of explanation—is a multifaceted problem, but one facet relates to understanding the "active" element of a treatment. In natural experiments, treatments are often "bundled," complicating inferences.[11] Strategy (2) can be seen as a form of "implicit mediation analysis" (Gerber and Green 2012: 333-36)—implicit because it does not seek precisely to parse direct and indirect effects but rather to take advantage of variation across versions of a treatment to assess active "channels." The strategy is also "design-based" in its reliance on variation that could be induced by a manipulation, as in an experiment.[12] Such analyses sometimes shed light empirically on mechanisms in an admittedly limited and blunt—but nonetheless potentially very useful—way.

Strategy (3) can address other key questions for mechanisms. Finding an impact of a treatment on some outcomes but not others can shed light on why the effect sometimes occurs: for example, a given treatment might have an effect on behavioral but not attitudinal outcomes, possibly suggesting that it changes incentives without changing perceptions or beliefs. In addition, researchers may assess impacts on "intermediate outcomes," which are themselves conceptualized as mechanisms. To be sure, leveraging such variation is also an aspect of formal mediation analysis but the goals here are more modest—and the inferences, if more limited, are also perhaps more credible. It is possible in principle for an intermediate outcome to be a mediator even if a treatment has no average effect on it, for instance, in a potential outcome framework in which an effect in one direction for some subjects and an effect in a different direction for others averages to zero across subjects. These are fairly knife-edged cases, though. From a pragmatic perspective, knowing whether a treatment moves—or fails to move—such a variable can be substantially helpful for evaluating claims about mechanisms.

In sum, these three broad empirical strategies can provide, we think, a useful empirical route to cumulative learning about generalizability and/or mechanism. For explanation, we highlight especially the potential utility of strategies (2) and (3). These strategies may be most effective when they leverage a wide source of qualitative and quantitative information and when they engage substantially with theoretical insights. For example, they might be used to distinguish between two theoretical explanations that are both consistent with some observed event, fact, or treatment effect.

---

[11] Dunning (2012, Chapter 10).

[12] Bullock and Green (2021). Gerber et al. (2008), for example, devise varied experimental prompts to distinguish between (a) a sense of civic duty, (b) Hawthorne (or "observer") effects, and (c) social pressure as the mechanisms that explain why a mobilization message may spark voter turnout.

These strategies could in principle be leveraged by the same set of authors and sometimes even within a single study. More often, however, new knowledge generated by one researcher or set of researchers may be linked to or depend on the knowledge previously obtained *by other researchers*; and the use of the three strategies emerges through such a collective endeavor. It is for this reason that we consider the assessment of generalizability and mechanism as a facet of knowledge accumulation (Mahoney 2003, Callis et al. 2022). Moreover, empirical knowledge of mechanism often does build cumulatively. Indeed, across different areas of scientific inquiry, an impact of X on Y may be well established long before the mechanism is understood.[13] We note also that replication in the form of strategy (1) can have ancillary benefits, for example, as a check on routines and procedures in an original study—and sometimes more generally as a way to generate critical scholarly dialogue that can lead to further verification of the first study's findings in a given context. Such replication may not always arise, of course: cumulative learning about generalizability and mechanism may occur (or fail to occur) as a function, for example, of the nature of the professional production of knowledge in an academic discipline.

Reflecting on the three strategies, we see little tension in principle between the goal of constructing a credible evidence base against which generalizable theories can be assessed and the emergence of the design-based turn in HPE. (Practice may be another question; we turn to that next). Several caveats are useful, however. One potential price of weaker modeling assumptions is more limited conclusions. For instance, strategies (2) or (3) compromise on the ability to parse direct and indirect effects exactly, relative to the apparent precision provided by a path model. Because of the omnipresent and distorting role of unobservables in applications of those models, however, we think "apparent" is often right: the precision is illusory. Where some might see a bug, we therefore see a feature and even a virtue of these simpler methods. The aims for assessing mechanisms are more modest but, we think, more credibly achievable.

We also note that the ability to assess mechanisms and infer generalizability may vary substantially across different modes of HPE, especially studies with different goals. Much historical work emphasizes the importance of temporality and sequence, and also the ways in which historical treatments or critical junctures may generate aftermaths and long-run legacies that consolidate their impact. A valid natural experiment can in principle (and sometimes in practice) identify long-run impacts, for example, by using treatment-control comparisons measured across a longue durée. But a long-run effect need not imply a legacy, conceptualized as the consequence of a series of reactions and counter-reactions that follow a critical juncture.[14] Some historical treatments are only found to be impactful under certain later, often contingent historical conditions; in the argument of Wilfahrt (2021) in her excellent book, for example, the legacies of pre-colonial kingdoms in Senegal only become effective for generating shared public goods across villages after post-colonial democratic decentralization. The ability of the tools of the credibility revolution to identify such complex modes of moderation and mediation implied by arguments that emphasize temporality, sequence, and legacies can be

---

[13] See e.g. Freedman (2009) on how knowledge that infected waste and water causes cholera transmission preceded the theory of germs.

[14] Collier and Munck (2022) make the presence of a legacy definitional to a critical juncture: no legacy, no critical juncture. This does not gainsay the possibility of long-run effects that are not legacies, however.

limited (though other kinds of methods face substantial challenges in validating such claims empirically as well). As for generalizability, although arguments involving sequence and temporality in one context may heighten our attention to the possibility of similar patterns in others, generalizable inference may be neither the goal nor a primary metric by which we should evaluate the success of such studies.

CUMULATIVE LEARNING IN HISTORICAL POLITICAL ECONOMY

Research on historical political economy since the "design-based turn" has led, in our view, to important forms of cumulative learning. We turn in this section to bodies of recent research in historical political economy that collectively leverage empirical strategies we discussed in the previous section (or that can be seen, from the perspective of a systematic review, as doing so). Our goal is to survey this progress and to highlight the attention that some scholars have brought to issues both of generalizability and mechanism, while also suggesting possible improvements that could foster stronger accumulation.

One clarifying note is important before turning to our survey. We describe work that has used natural experiments (with either randomized or as-if random assignment to treatment conditions), discontinuities and/or instrumental-variables methods. Our focus, however, is not on internal validity of the causal claims in particular studies nor on the strengths of the different designs from that perspective. We recognize that in some of the studies we discuss, as-if random or other key identifying assumptions—such as exclusion restrictions, non-interference, and so on—might fail. Our goal is not to judge such elements of the designs but to assess how the combination of such studies may contribute to assessment of generalizability or mechanism, using the strategies we have outlined above.

European expansion

A broad array of scholarship examines the long-term effect of European colonial expansion on contemporary political and economic outcomes. There is obviously no question that many of the institutions associated with European colonialism reaped devastating consequences on colonized societies. Yet different forms of colonialism may be more or less destructive, and some may even exhibit some positive externalities on outcomes ranging from economic growth to state capacity.[15] Understanding this variation—in terms of both the presence of different types of effects across settings and the causal mechanisms that drive these effects—has been a major focus of recent work in historical political economy.

Direct vs. Indirect Colonial Rule. A first set of studies explores the impact of direct, versus indirect, colonial rule on economic development. In colonial India, Iyer (2010) examines the effect of being governed directly by the British Crown, versus indirectly through Indian princely states, on economic development. Between 1848 and 1856, the British sought to assume direct

---

[15] Some studies, which are not discussed here directly, have pointed to the potential for divergent effects based on the national identity of colonizers. See, e.g., Feyrer and Sacerdote (2009). See also Mattingly (2017) for a study of the effects of non-European colonial rule in China.

control over any Indian states whose king died without a natural heir. Iyer leverages possibly as-if random variation in the absence of an heir at the time of a ruler's death—introduced by this "Doctrine of Lapse"—to instrument for the effects of direct rule. She finds that states that experienced direct rule as a result had fewer public goods and lower levels of economic development. This finding contrasts markedly with a (likely confounded) cross-sectional comparison of directly and indirectly colonized areas: the British tended to prefer direct rule in more fertile, wetter regions, creating the appearance of a direct-rule advantage for some development outcomes. Lee and Schultz (2010) report similar findings in Cameroon, where regions colonized by the British were more likely to experience indirect rule than regions under French control. Leveraging a geographic discontinuity design, they find that areas under British control have higher levels of wealth and public goods provision in the contemporary period.

In contrast to this work, studies in Namibia find a combination of null and positive effects of direct colonial rule on economic development. In the 1890s, Germany divided colonial Namibia into two regions based on the prevalence of an infectious cattle disease—one region ruled directly and the other ruled indirectly through indigenous elites. The border is possibly exogenous, cross-cutting existing ethnic boundaries and other attributes that might be likely to affect outcomes such as economic development. Using a geographic discontinuity design, Lechler and McNamee (2018) leverage this colonial division to estimate the effect of direct rule on economic outcomes.[16] They find no effect of direct rule on economic development, as measured by educational attainment, density of night lights, poverty indices, and infrastructure quality. In related work, Chlouba and He (2020) leverage this same colonial division to estimate a *positive* effect of direct, versus indirect, rule on a different set of development outcomes—the commercialization of agriculture, household living standards, and paved roads.

The divergent findings in India and Cameroon on the one hand, and Namibia on the other, raise important questions about the causal mechanisms driving the effect of (in)direct colonial rule on economic development. One potential mechanism is the role of land tenure regimes in areas with indirect rule. Lechler and McNamee (2018) describe the importance of communal land tenure in areas of Namibia under indirect colonial rule.[17] Similarly, Chlouba and He (2020) point to this land tenure arrangement in reducing economic development in areas under indirect rule.[18]

Might different forms of land tenure help to explain the distinct effects of direct rule on economic development? For land tenure to be a compelling explanation of the variation in effects, we would want to establish both that i) more secure land tenure systems have a positive effect on economic development; and that ii) the link between direct rule and more secure land tenure regimes documented in Namibia is not present in contexts where scholars have found a

---

[16] Their main outcome of interest is support for democracy. However, they also analyze measures of economic development and find no difference across areas with direct, versus indirect, colonial rule.

[17] In related work, McNamee (2019) finds suggestive evidence of the enduring role of communal land tenure arrangements in areas of Namibia that were under indirect colonial rule.

[18] While relying on the same discontinuity as Lechler and McNamee (2018), the empirical strategy in Chlouba and He (2020) limits the sample to a subset of households, employing a matching strategy on either side of the colonial border.

negative effect of direct rule on economic development.[19] These are not trivial inferential targets, but design-based variation in land tenure systems (as an independent variable) and assessing the impact of direct rule on security of land tenure (as a dependent variable) is useful.

Studies in India provide suggestive evidence on both these counts. First, using an instrumental-variables design, Banerjee and Iyer (2005) find that areas in which peasants' property rights were less secure experienced lower levels of economic development over the long term. Second, to explain the negative effect of direct rule on economic development, Iyer (2010) examines whether land tenure arrangements vary across regions assigned to different types of colonial rule. She finds no evidence of a difference. We can think of Banerjee and Iyer (2005) as directly "manipulating" the mediator of interest (land tenure arrangements) and Iyer (2010) as probing the impact of the treatment variable on an intermediate outcome (i.e. the mediator). Combined, these studies suggest that land tenure arrangements play an important role in shaping the economic effects of direct, versus indirect, colonial rule.

In sum, here we see a combination of strategy (1)—related study designs across India, Cameroon, and Namibia—and strategies (2) and/or (3). In Namibia, the treatment bundles the form of rule and the security of property rights, while there is no such bundling in India. Alternatively, one can think of land tenure regimes as an intermediate outcome—and a possible mechanism—and assess impacts of direct vs. indirect rule on it. Implicit mediation analysis and/or variation in effects on distinct outcomes thus give useful insights that may help to explain the differences in the effects across context. To be sure, other mechanisms are likely consistent with the divergent effects, and the evidence is therefore perhaps only suggestive. Yet this is also an area where theory could play an important role in developing further testable implications that discriminate between rival explanations. A key point is that a body of evidence from a *set* of related studies is important for explanation and for illuminating generalizability and mechanism.

Missionary activity. Another recent literature explores the long-term effects of missionary activity during colonialism on contemporary educational outcomes. This research builds on Woodberry (2004, 2012), who documents a positive association between the historical presence of missionaries and current per capita income and democracy across former non-settler colonies due to religion's role in fomenting education. Subsequently, scholars have further explored and fleshed out the role of missionary activity during colonial rule in shaping educational attainment today.

One notable feature of this literature is its breadth; scholars have probed the effect of missionary activity on educational outcomes across vastly different geographic contexts. A first set of studies examines the role of missionaries in different areas in Africa, leveraging the possibly as-if random location of missions' location historically to study their effects on contemporary

---

[19] Finding evidence in support of (i) and (ii) may be neither necessary nor sufficient to establish that variation in land tenure regimes explains the differences in the effects of direct versus indirect rule across contexts. In this sense, testing these claims is comparable to a "straw in the wind" test in process tracing (Van Evera 1997, Bennett 2010, Collier 2011).

outcomes.[20] Gallego and Woodberry (2010) provide evidence that regions of Africa in which Protestant missionaries predominated in the colonial period have higher literacy rates today than those regions where Catholic missionaries predominated. Nunn (2014) finds that both Protestant and Catholic missionary activity in Africa exerted long-term, positive impacts on education. Studying colonial Benin, Wantchekon, Novta, and Klašnja (2015) show that families who live in close proximity to historical missions assign greater value to education.

More recently, scholars have sought to extend these findings beyond Africa. Valencia Caicedo (2019) finds that Jesuits had a positive effect on education in South America, where a key component of their missionary activities was teaching the indigenous Guaraní to read and write. He shows that educational attainment today is higher in places where the Jesuits had a mission prior to their expulsion from the Americas in 1767. Waldinger (2017) examines the long-run effects of different Catholic missionary orders in colonial Mexico on educational outcomes. She finds that mendicant orders—which were committed to reducing poverty and social inequality by educating the native population—improved educational attainment in the regions of Mexico's countryside where their historical presence was greater.[21] In China, Bai and Kung (2015) find that regions with a greater presence of Protestant missionaries during the 19th century had higher urbanization rates at the beginning of the 20th century. Though the primary outcome is not education, the causal mechanism the authors identify is knowledge diffusion associated with the schools and hospitals that missionaries built.

Through strategy (1), these studies probe the external validity of the causal relationship between missionary activity and educational outcomes. Across a wide array of settings around the world, they find that the presence of missionaries increased literacy, educational attainment, and related outcomes. Unlike studies that pursue strategies (2) and (3), these do not make the leap to explanation through design-based, within- and cross-case comparisons that allow us to assess effects on intermediate outcomes or use implicit mediation analysis. Yet they do help inform our understanding of the generalizability of this causal relationship.

Forced Labor. A prominent body of scholarship in historical political economy examines the role of institutions of forced labor in shaping contemporary economic development. Building on Acemoglu et al. (2001, 2002), who argue that broadly extractive institutions hamper economic growth over the long run, these studies have explored such effects for a particular type of extraction: forced labor under colonialism.[22] For example, Dell (2010) employs a geographic discontinuity design to document the negative effect of exposure to forced labor in the silver mines of Peru and Bolivia during Spanish colonization on a range of social and development outcomes. Lowes and Montero (2016) leverage a similar empirical strategy to examine

---

[20] Jedwab et al. (2022) examine the economics of missionary expansion in Ghana and find that missionaries went to healthier, safer, more accessible, and richer areas, prioritizing investments in these better locations first. They argue that these factors might spuriously explain why locations with past missions, and better missions in particular, are more developed today.

[21] In contrast, the presence of Jesuit missionaries, who focused educational efforts on the colony's elite in the city centers, had no lasting effects.

[22] The alternative to these extractive institutions are those that protect private property and establish checks against government overreach (Acemoglu et al. 2001, 1370).

extractive rubber concessions in the Democratic Republic of the Congo, which entailed the forced cultivation of rubber. They find similarly negative effects.

Not all studies reach the same conclusion, however. Dell and Olken (2020) employ a geographic regression discontinuity design to identify the effect of forced labor in colonial Java's sugar industry. Notably, they find that this institution had a *positive* effect on economic development. They tentatively suggest that this divergent finding is due to the role of the institutions and economic industries that accompanied forced labor in Java, which facilitated economic growth over the long term, perhaps outweighing the negative effect of forced labor itself.

This literature engages in strategy (1) to explore the effects of forced labor across different contexts (here Peru, the Democratic Republic of Congo, and Java). While the differences in observed effects may be due to the complementary institutions that accompany the forced labor institutions, as suggested by the Java study, there is no empirical work to assess the validity of this claim. Indeed, Gailmard (2021a) highlights Dell and Olken's (2020) research in Java as indicative of a body of empirical work that does little to inform our understanding of theory and generalizability. Yet this is partly an empirical issue that reflects the bundled "treatment" in colonial Java, which includes both an extractive institution and a set of complementary institutions that sprung up in tandem with forced labor in this case. Future research might seek to pursue strategy (2) to begin to parse the effects of these distinct institutions, and thus gain purchase on important theoretical questions about the causal relationship between forced labor and economic development.

One potentially fruitful direction may also be to explore the possibly contrasting impacts of forced labor in a single context on different outcomes, per strategy (3). In recent work in Peru, for example, Carter (2021, 2022) suggests that forced labor conscription to rebuild the Inca Road in the 1920s increased the ability of indigenous leaders and their communities to secure institutions that protected their communal land and other traditional institutions. Such institutions, however, can have mixed welfare effects, and under some conditions they may limit the ability of communities to secure government benefits or broader economic development. Such a strategy may not resolve many questions related to the generalizable effects of forced labor in this context—but it can help shed important light on variation in effects across different domains, by leveraging contrasting impacts on different outcomes.

The Slave Trade. Other research studies an alternative form of extraction, that of the slave trade in Africa. Nunn (2008) explores the effect of slave trading across Africa on contemporary levels of economic development. Leveraging the distance of each African country to the markets where slaves were received to instrument for the intensity of the slave trade, he finds a significant, negative causal effect of the intensity of slave trading on economic development today. While few would question slavery's devastating impact on societies across Africa, Nunn's findings are important to quantify the dimensions of that harm, and may help to pinpoint origins of underdevelopment across the region.

Subsequent scholarship has sought to shed light on the mechanism linking the slave trade with contemporary economic development. Nunn and Wantchekon (2011) argue that slave trading engendered a culture of mistrust among those individuals who were most affected, which has persisted to the present day.[23] In complementary work, Whatley and Gillezeau (2011) examine the effect of the slave trade on ethnic fragmentation along Africa's western coast.[24] They find that slave trading increased the number of ethnic groups in Africa, an important source of social conflict that is often thought to impede economic growth.[25]

These studies certainly contribute to the cumulation of knowledge by exploring the causal mechanisms that may link the slave trade with lower levels of contemporary development. They rely on strategy (3), employing similar designs across overlapping regions of Africa to explore the effects of slave trading on distinct, but related outcomes. Collectively, they suggest the slave trade's pernicious effects are driven, at least in part, by changing social dynamics, both across and within African ethnic groups. The studies provide evidence on impacts on intermediate outcomes—like trust—that may be mechanisms linking the slave trade to development outcomes. They cannot, we think, establish that these are "the" mechanisms that account for the effect on development; other causal mechanisms may be at play as well. Nonetheless, they provide very useful empirical purchase on questions of mechanism and generalizability across Africa, and they use relatively design-based empirical strategies to investigate explanations for the effect. Future work might attempt to identify complementary mechanisms by further employing both strategies (2) and (3).

Cumulative learning in other areas of historical political economy

Outside of the large literature on European colonial expansion, research in other substantive areas of historical political economy appears to leverage strategies (1), (2), and (3). For example, a growing body of work sheds light on the causal effects of Black migration from the southern to the northern United States in the 1940s-1970s using similar versions of a shift-share instrument.[26] Calderon et al. (forthcoming) find that the Great Migration of African Americans northward led to increases in Democratic vote share and support for organizations associated with the civil rights movement among both whites and African Americans. Boustan (2010) argues that the influx of southern Blacks triggered white flight to the suburbs. In a similar vein, Derenoncourt (2019) finds that the influx of migrants reduced the upward mobility of African Americans in the North, in part due to greater local segregation and spending on police. Overall, these studies employ strategy (3) to explore the effects of the Great Migration on a range of

---

[23] Here, Nunn and Wantchekon (2011) instrument for the incidence of the slave trade using the distance of a given ethnic group from the African coast, where slaved were purchased before being sent to their final destination. The geographic focus of this paper is all countries in Subsaharan Africa except Cameroon, Gabon, the Democratic Republic of Congo, Angola, Sudan, Ethiopia, and Eritrea.

[24] Their study extends from Tunisia in the north to South Africa in the south.

[25] See, e.g., Easterly and Levine (1997).

[26] Beyond these studies, a vast literature has explored the effects of the Great Migration using other empirical methods. Similarly, scholars have employed versions of this shift-share strategy to explore earlier waves of African American migration, as well as the migration of other groups. See, e.g., Fouka and Tabellini (2021) and Fouka et al. (2021).

political and economic outcomes. Their findings suggest that greater local segregation does not preclude, and may facilitate, support for policies that promote racial equality on other domains.

In other work, scholars have explored the electoral consequences of technological innovations in agriculture by combining local-level variation in the ability to take advantage of new technologies with temporal variation in the timing of these technologies' introduction. Dasgupta (2018) examines the introduction of high-yield crops in India, demonstrating that it led to a change in the political preferences of rural voters and played a pivotal role in the rise of agrarian opposition parties and the decline of single-party Congress dominance. In a working paper on the United States, Dasgupta and Ruiz Ramirez (2022) examine the effects of the post-war introduction of improved irrigation systems, which enabled farmers to expand their cultivation into arid regions of the Great Plains. They find that the technological shift triggered higher demand for conservative economic policies and deregulation. Both studies document the rise of an upwardly mobile agricultural class in driving these changes. These studies employ strategy (1) to identify the effects of technological innovations in agriculture on electoral realignment across widely distinct contexts. Doing so sheds light on the generalizability of the causal relationship between such innovations and electoral outcomes. Moreover, these studies separately explore the causal mechanisms that drive these observed effects through strategy (3); in both cases, they find that the rise of an upwardly mobile agricultural class appears important in driving these changes. They thus contribute to both dimensions of cumulative learning by informing our understanding of generalizability and mechanism.

CONCLUSION

Empirical routes to the evaluation of generalizability and mechanism are essential in most areas of social science and perhaps especially in historical political economy—a field focusing definitionally on the empirical terrain of "history." Social scientists have laid out a range of strategies, ranging from qualitative process tracing to path models. We have proposed a different alternative in this chapter, describing how design-based replication, implicit mediation analysis, and assessment of impacts on varied outcomes may help us learn cumulatively about generalizability and mechanism. A credible evidence base is critical both for testing and developing theory. For generalizable theory, it is important that the evidence spans contexts, helps to unbundle treatments, and/or allows assessment of varied effects on distinct outcomes.

Our partial review of work in several substantive areas of HPE suggests that some elements of the strategies we describe are at work in practice, either implicitly or explicitly. However, we found in our survey several important practical issues, with implications for the construction of the evidence base. In related work (Callis et al. 2022), we have studied knowledge accumulation via natural experiments focused on "contemporary" politics. Our impression is that in that context examples were somewhat easier to find than in HPE. Part of the explanation might indeed be that these strategies are harder to implement in historical work, due perhaps to data constraints and the fact that relatively "off-the-shelf" designs—such as close race regression discontinuity designs that rely on the presence of democratic elections—are less prevalent in historical settings. However, we also believe that more systematic focus on

replication and extension of previous work, with an eye to assessing generalizability and mechanism, will bear fruit in HPE as well.

Another critical issue, raised by other scholars, relates to case selection. We think it is only partially true that the idiosyncratic nature of natural experiments and other design-based methods undermines the ability to select cases theoretically (as opposed to inductively, based on where as-if random variation arises). Moreover, issues of case selection also arise in the analysis of more ostensibly comprehensive (e.g. cross-national) datasets, where it is not however always readily apparent which cases contribute most to inferences. Formal meta-analysis may not be desirable with the strategies we have outlined here; there is usually not a large enough N of cases nor enough consistency or harmonization across cases for that to be possible. However, by taking advantage of design-based variation both across and within cases and by comparing findings across studies and contexts, we can enhance the credibility of inferences about generalizability of causal relationships and mechanisms that undergird them.

We see no necessary tension between the credibility revolution and the construction of a cross-case evidence base that can be used to discipline, advance, and validate generalizable theory. However, it seems important that the goal of assessing generalizability and mechanism be explicit. We think it would be useful if strategies for cumulative learning—such as those we have laid out here—were leveraged more purposively by different sets of scholars, using design-based variation both within and across studies.

References

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

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2002. "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution*." *The Quarterly Journal of Economics* 117(4):1231–94.

Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15(4):69–85.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

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

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

Bai, Ying, and James Kai-sing Kung. 2015. "Diffusing Knowledge While Spreading God's Message: Protestantism and Economic Prosperity in China, 1840–1920." *Journal of the European Economic Association* 13(4):669–98.

Banerjee, Abhijit, and Lakshmi Iyer. 2005. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review* 95(4):1190–1213.

Bennett, Andrew. 2010. "Process Tracing and Causal Inference." Pp. 207–20 in *Rethinking Social   Inquiry*, edited by D. Collier and H. E. Brady. New York, NY: Rowman and Littlefield.

Boustan, Leah Platt. 2010. "Was Postwar Suburbanization 'White Flight'? Evidence from the Black Migration." *Quarterly Journal of Economics* 125(1):417–43.

Bullock, John G., and Donald P. Green. 2021. "The Failings of Conventional Mediation Analysis and a Design-Based Alternative." *Advances in Methods and Practices in Psychological Science* 4(4):25152459211047228.

Calderon, Alvaro, Vasiliki Fouka, and Marco Tabellini. forthcoming. "Racial Diversity and Racial Policy Preferences: The Great Migration and Civil Rights." *Review of Economic Studies*.

Campbell, Donald T., and Julian Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. 1st edition. Belomt, CA: Cengage Learning.

Cantoni, Davide, and Noam Yuchtman. 2021. "Historical Natural Experiments: Bridging Economics and Economic History." Pp. 213–41 in *The Handbook of Historical Economics*, edited by A. Bisin and G. Federico. Academic Press.

Carter, Christopher L. 2021. "The Representational Effects of Communal Property: Evidence from Peru's Indigenous Groups." *Comparative Political Studies* 54(12):2191–2225.

Carter, Christopher L. 2022. "Extraction, Assimilation, and Accommodation: The Historical Foundations of Indigenous-State Relations in Latin America."

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press.

Chlouba, Vladimir, and Jianzi He. 2021. "Colonial Legacy, Private Property, and Rural Development: Evidence from Namibian Countryside." *Economic History of Developing Regions* 36(1):30–56.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44(4):823–30.

Collier, David, Henry E. Brady, and Jason Seawright. 2004. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology." Pp. 229–66 in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited by H. E. Brady and D. Collier. Lanham: Rowman and Littlefield.

Collier, David, and Gerardo L. Munck. 2022. *Critical Junctures and Historical Legacies: Insights and Methods for Comparative Social Science*. Rowman & Littlefield.

Dasgupta, Aditya. 2018. "Technological Change and Political Turnover: The Democratizing Effects of the Green Revolution in India." *American Political Science Review* 112(4):918–38.

Dasgupta, Aditya, and Elena Ruiz Ramirez. 2022. "Explaining Rural Conservatism: Political Consequences of Technological Change in the Great Plains."

Deaton, Angus. 2010. "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives* 24(3):3–16.

Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining Mita." *Econometrica* 78(6):1863–1903.

Dell, Melissa, and Benjamin Olken. 2020. "The Development Effects of the Extractive Colonial Economy: The Dutch Cultivation System in Java." *Review of Economic Studies* 87(1):164–203.

Derenoncourt, Ellora. 2022. "Can You Move to Opportunity? Evidence from the Great Migration." *American Economic Review* 112(2):369–408.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press.

Easterly, William, and Ross Levine. 1997. "Africa's Growth Tragedy: Policies and Ethnic Divisions." *The Quarterly Journal of Economics* 112(4):1203–50.

Eggers, Andrew C., Guadalupe Tuñón, and Alan Dafoe. 2022. "Placebo Tests for Causal Inference."

Evera, Stephen Van. 2015. *Guide to Methods for Students of Political Science*. Cornell University Press.

Feyrer, James, and Bruce Sacerdote. 2009. "Colonialism and Modern Income: Islands as Natural Experiments." *The Review of Economics and Statistics* 91(2):245–62.

Fouka, Vasiliki, Soumyajit Mazumder, and Marco Tabellini. 2022. "From Immigrants to Americans: Race and Assimilation during the Great Migration." *The Review of Economic Studies* 89(2):811–42.

Fouka, Vasiliki, and Marco Tabellini. 2021. "Changing In-Group Boundaries: The Effect of Immigration on Race Relations in the United States." *American Political Science Review* 1–17.

Freedman, David A. 2009. "Statistical Models and Shoe Leather." Pp. 45–62 in *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, edited by D. Collier, J. S. Sekhon, and P. B. Stark. Cambridge: Cambridge University Press.

Gailmard, Sean. 2021a. "Theory, History, and Political Economy." *Journal of Historical Political Economy* 1(1):69–104.

Gailmard, Sean. 2021b. "What Are We Doing in Historical Political Economy?" *Broadstreet*. Retrieved (https://broadstreet.blog/2021/01/08/what-are-we-doing-in-historical-political-economy/.).

Gallego, Francisco, and Robert Woodberry. 2010. "Christian Missionaries and Education in Former African Colonies: How Competition Mattered." *Journal of African Economies* 19(3):294–329.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. Illustrated edition. New York: W. W. Norton & Company.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102(1):33–48.

Glynn, Adam N. 2012. "The Product and Difference Fallacies for Indirect Effects." *American Journal of Political Science* 56(1):257–69.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1).

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48(2):399–423.

Iyer, Lakshmi. 2010. "Direct versus Indirect Colonial Rule in India: Long-Term Consequences." *The Review of Economics and Statistics* 92(4):693–713.

Jedwab, Remi, Felix Meier zu Selhausen, and Alexander Moradi. forthcoming. "The Economics of Missionary Expansion: Evidence from Africa and Implications for Development." *Journal of Economic Growth*.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Leamer, Edward. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73(1):31–43.

Lechler, Marie, and Lachlan McNamee. 2018. "Indirect Colonial Rule Undermines Support for Democracy: Evidence From a Natural Experiment in Namibia." *Comparative Political Studies* 51(14):1858–98.

Lee, Alexander, and Kenneth A. Schultz. 2012. "Comparing British and French Colonial Legacies: A Discontinuity Analysis of Cameroon." *Quarterly Journal of Political Science* 7(4):365–410.

Lowes, Sara, and Eduardo Montero. 2021. "Concessions, Violence, and Indirect Rule: Evidence from the Congo Free State*." *The Quarterly Journal of Economics* 136(4):2047–91.

Mahoney, James. 2003. "Knowledge Accumulation in Comparative Historical Research: The Case of Democracy and Authoritarianism." Pp. 131–74 in *Comparative Historical Analysis in the Social Sciences*, *Cambridge Studies in Comparative Politics*, edited by D. Rueschemeyer and J. Mahoney. Cambridge: Cambridge University Press.

Mattingly, Daniel C. 2017. "Colonial Legacies and State Institutions in China: Evidence From a Natural Experiment." *Comparative Political Studies* 50(4):434–63.

McNamee, Lachlan. 2019. "Indirect Colonial Rule and the Salience of Ethnicity." *World Development* 122:142–56.

North, Douglass C., and Barry R. Weingast. 1989. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England." *The Journal of Economic History* 49(4):803–32.

Nunn, Nathan. 2008. "The Long Term Effects of Africa's Slave Trades." *Quarterly Journal of Economics* 123(1):139–76.

Nunn, Nathan. 2014. "Gender and Missionary Influence in Colonial Africa." Pp. 489–512 in *Africa's Development in Historical Perspective*, edited by E. Akyeampong, R. H. Bates, N. Nunn, and J. A. Robinson. New York, NY: Cambridge University Press.

Nunn, Nathan, and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101(7):3221–52.

PIncus, Steve, and James Robinson. 2014. *Institutions, Property Rights, and Economic Growth: The Legacy of Douglass North*. New York: Cambridge University Press.

Robinson, James A. 2010. *Natural Experiments of History*. edited by J. Diamond. Cambridge, Massachusetts: Belknap Press of Harvard University Press.

Rozenas, Arturas. 2021. "History and Theory in Historical Political Economy." *Broadstreet*. Retrieved (https://broadstreet.blog/2021/11/12/history-and-theory-in-historical-political-economy/).

Stasavage, David. 2002. "Credible Commitment in Early Modern Europe: North and Weingast Revisited." *The Journal of Law, Economics, and Organization* 18(1):155–86.

Stasavage, David. 2007. "Partisan Politics and Public Debt: The Importance of the 'Whig Supremacy' for Britain's Financial Revolution." *European Review of Economic History* 11(1):123–53.

Valencia Caicedo, Felipe. 2019. "The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America." *The Quarterly Journal of Economics* 134(1):507–56.

Valencia Caicedo, Felipe. 2021. "Historical Econometrics: Instrumental Variables and Regression Discontinuity Designs." Pp. 179–211 in *The Handbook of Historical Economics*, edited by A. Bisin and G. Federico. Academic Press.

Waldinger, Maria. 2017. "The Long-Run Effects of Missionary Orders in Mexico." *Journal of Development Economics* 127:355–78.

Waldner, David. 2016. "Invariant Causal Mechanisms." *Qualitative & Multi-Method Research* 14(1):28–34.

Wantchekon, Leonard, Marko Klašnja, and Natalija Novta. 2015. "Education and Human Capital Externalities: Evidence from Colonial Benin." *The Quarterly Journal of Economics* 130(2):703–57.

Whatley, Warren, and Rob Gillezeau. 2011. "The Impact of the Transatlantic Slave Trade on Ethnic Stratification in Africa." *American Economic Review* 101(3):571–76.

Wilfahrt, Martha. 2021. *Precolonial Legacies in Postcolonial Politics: Representation and Redistribution in Decentralized West Africa*. Cambridge University Press.

Woodberry, Robert D. 2004. "The Shadow of Empire: Christian Missions, Colonial Policy, and Democracy in Postcolonial Societies." Ph.D. Dissertation, University of North Carolina, Chapel Hill.

Woodberry, Robert D. 2012. "The Missionary Roots of Liberal Democracy." *The American Political Science Review* 106(2):244–74.