Knowledge Accumulation Through Natural Experiments[1]

Anna Callis
Thad Dunning
Guadalupe Tuñón

Word count: 9,615

ABSTRACT:

Scholars often extoll the benefits of knowledge accumulation. Natural experiments, however, are often thought of as idiosyncratic and one-off studies that may not therefore contribute to cumulative learning. We explore this case against natural experiments in this chapter. We emphasize two key dimensions of knowledge about causal effects—generalizability and mechanisms—and underscore three empirical strategies for boosting accumulation: comparing studies in which (1) context varies but treatments and outcomes are similar; (2) different treatments are employed in the same context and with the same outcome measures; and (3) similar treatments are carried out in the same context with distinct, but related, outcomes. Surveying examples of natural experiments across different substantive areas, we find that scholars can, and have already, leveraged these strategies to foster cumulative learning. However, several features of these designs and of their use do pose barriers. We outline several ways in which knowledge accumulation using natural experiments can be further enhanced.

INTRODUCTION

Scholars often hail the value of knowledge accumulation. Probing the effects of a political program or social policy in different cases, for example, can improve our understanding of whether or not a causal relationship we observe in one context travels to other settings. Similarly, examining the impacts of different components of a policy or program, as well as its effects on related outcomes of interest, can shed light on why those impacts arise and whether we should expect them to generalize to different settings. Knowledge accumulation can give rise to scholarly consensus about the impact—positive, negative, or null—of key independent variables, as well as a better sense of contextual variation in relations of cause and effect.

Such consensus building can happen in many areas of social science, using varied research methods. Mahoney (2003), for example, discusses knowledge accumulation in comparative historical research, with a focus on democracy and authoritarianism (see also Gerring, this volume). The recent growth of randomized controlled trials (RCTs) may also bode well for cumulative learning. Researchers' control over the design of interventions that are tailored to specific research questions and that may be implemented in diverse contexts allows, at least in

[1] Prepared for inclusion in the *Oxford Handbook of Methodological Pluralism*, edited by Janet Box-Steffensmeier, Dino Christenson, and Valeria Sinclair-Chapman. Author affiliations: University of California, Berkeley (Anna Callis and Thad Dunning) and Princeton University (Guadalupe Tuñón).

principle, for investigation of the generalizability of key findings and assessment of how they may vary across contexts. Several recent research initiatives, moreover, have sought more effectively to coordinate experimental research across contexts and thereby mitigate limitations that may constrain knowledge accumulation in RCTs.[2] These efforts may aid assessment of the generalizability of discrete findings and improve our understanding of the mechanisms that account for specific effects.

In contrast, the use of natural experiments may not appear conducive to knowledge accumulation. Natural experiments––in which the researcher relies on random, or "as-if" random, assignment to treatment rather than directly manipulating treatment status—have many attractive features for evaluating the effects of social, economic, and political phenomena in particular cases (Dunning 2012). They can aid the exploration of the effects of causes that are difficult to manipulate experimentally, expanding the potential scope of scholarly learning. The heightened realism of natural experiments relative to some RCTs lends credibility to their results, since treatments that occur "naturally" are more likely to avoid social desirability bias and other dynamics that might mask true treatment effects in an experimental setting. Furthermore, natural experiments are often implemented at the same level at which scholars seek to make causal inferences, precluding the risk that estimated effects observed in a subset of the population diverge from the effects that would result were the treatment carried out "to scale."[3] Yet the serendipitous nature of natural experiments—and their apparent reliance on idiosyncratic features of particular empirical settings to generate exogenous variation in a treatment of interest—seems to complicate replication across cases. Furthermore, because social scientists do not intervene directly in the design or implementation of natural experimental treatments, it may be difficult to isolate the effect of the cause that researchers would—ideally—be most interested in studying, rather than the one "Nature" deigned to assign. This appears to hinder scholars' ability to extrapolate findings beyond specific empirical contexts or isolate the causal contributions of distinct components of a bundled treatment.

In this chapter, we assess this case against cumulative learning through natural experiments. We argue that, in fact, these barriers do not preclude the possibility of knowledge accumulation. We identify three empirical strategies that can boost learning about causal effects across two key dimensions—generalizability and mechanisms. By comparing studies in which (1) context varies but treatments and outcomes are similar; (2) different treatments are employed in the same context with the same outcome; and (3) similar treatments are carried out in the same context on distinct, but related, outcomes, analysts can make substantial progress. Our survey of existing examples suggests that this approach can be leveraged in a number of substantive areas, providing opportunities to learn about the effects of different classes of a treatment on a variety of related outcomes, across a range of cases. In sum, natural experiments often are suitable for knowledge building. However, their capacity to foster cumulation does suffer from a

[2] These include, among others, Banerjee et al. (2015); the Metaketa Initiative of the EGAP network (see Humphreys et al., this volume, Dunning et al. 2019a,b; Blair et al. 2021; and https://egap.org/our-work-0/the-metaketa-initiative/), and related initiatives at J-PAL (https://www.povertyactionlab.org/initiatives) and CEGA (https://cega.berkeley.edu).

[3] For a discussion of this concern in relation to RCTs, see Acemoglu (2010) and Banerjee et al. (2017).

range of factors, many of them not inherent to the designs but rather reflecting aspects of research communities and the nature of knowledge production.

In the next section, we define knowledge accumulation, with a special focus on causal relations; underscore the importance of assessing generalizability and mechanism; and outline three empirical strategies that may generally aid in such assessment. We then turn to natural experiments, highlighting the use of the three strategies across different substantive areas. Finally, we discuss constraints on cumulative learning and suggest possibilities for more effective knowledge accumulation through natural experiments.

THE ACCUMULATION OF KNOWLEDGE ABOUT CAUSATION

A central aim of the social sciences is knowledge accumulation. By “knowledge,” we mean valid, useful, or correct insights, e.g. about the working of the social and political world. By “accumulation,” we mean that the insights from discrete studies or research findings build on one another. Mahoney (2003: 133) similarly notes that “accumulation occurs when the generation of new knowledge *is dependent on* previously obtained knowledge” (italics in original).

Knowledge accumulation, which may also be called “cumulative learning” (Dunning et al. 2019a), thus goes beyond the existence of disparate studies on a particular topic: the designs or findings of the studies must be linked in some way. Whether “accumulation” leads to more valid or reliable knowledge—compared, say, to a set of studies that do not build on one another and are thus non-cumulative—may best be left as a hypothesis, rather than made definitional to the term. Yet the expectation is often that knowledge accumulation “implies progress in understanding and learning” (Mahoney 2003: 132).

Accumulation has another aspect, which, while not strictly definitional, often appears important: new knowledge generated by one researcher or set of researchers depends on the knowledge previously obtained *by other researchers*. Thus, cumulative learning often occurs (or fails to occur) as a function, for example, of the nature of the professional production of knowledge in an academic discipline. The resulting nature of accumulation may reflect career incentives or other motivations that researchers have for undertaking particular strands of research. In addition, studies may consciously build on the findings of previous research. Yet this is also not definitional: knowledge accumulation can occur when a third party or reviewer systematically or holistically combines the results of disparate but related studies.

In this chapter, we are especially concerned with the accumulation of knowledge about causal relationships. Cumulative learning can happen in many areas of social science, including in research related to concept formation and to descriptive inference. Yet, a key aim of much social science research is to assess causal relationships. The accumulation of knowledge about causation may raise distinctive challenges, which we consider in this section.

What dimensions of causal assessment are most relevant to knowledge accumulation? Consider as a starting point a single study where a causal effect of treatment X on outcome Y is estimated in a particular context or population Z. One dimension that can involve accumulation concerns the accuracy of the causal finding. In natural experimental research, scholarly exchanges sometimes focus on the validity of key causal assumptions; see e.g. Kocher and Monteiro (2016)'s critique of Ferwerda and Miller (2014). Such valuable interchanges allow cumulative learning about the validity of a causal claim in a single study. However, they do not readily allow assessment of causal claims that go beyond those made in the initial study. Here, we largely set aside such concerns about *internal validity*---defined by Campbell and Stanley (1963: 5) as "the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance?"

Two dimensions appear especially pertinent. One is *generalizability*, or the "external validity" of findings: "to what populations, settings, treatment variables, and measurement [outcome] variables can this effect be generalized" (Campbell and Stanley 1963: 5)?[4] The recent formal framework of Egami and Hartman (2022) similarly involves "four dimensions of external validity…units, treatments, outcomes, and contexts/settings." Questions such as the following thus relate to generalizability: Does the effect travel to other contexts or populations than that of the original study? Would related but different treatments have similar impacts? And, would a given independent variable have distinct effects on different kinds of outcomes?

A second important dimension of causation concerns *mechanisms*: that is, why did the effect arise? This concern is not unrelated to generalizability; as philosophers of science have pointed out, understanding the mechanism in one context can shed light on whether a cause will have a similar effect in a different setting, where the enabling mechanisms may or may not be operative.[5] However, understanding a mechanism is not identical to investigating generalizability. We could assess whether an effect of X on Y in context Z also holds in context W without having any understanding of the operative mechanism at work. Conversely, understanding the mechanism has value beyond generalizability; for example, it can help inform the design of interventions other than X that could conceivably produce a similar effect on Y.

Three empirical strategies for cumulative learning

How can researchers investigate these two key dimensions of causation? As we will argue, this often involves—and may perhaps require—cumulative learning.

Assessing generalizability, for instance, may involve replication of a study in a new context or population, or with a slightly varied treatment or outcome variable. For example, scholars may assess effects in context W only after one in context Z has previously been found. The subsequent investigation builds on or arises in reference to the prior one, often across studies conducted by different researchers, and the findings may be juxtaposed to consider questions of

[4] We add "outcome" in brackets here, since we interpret Campbell and Stanley's use of "measurement variables" to refer to the measured outcome.

[5] Cartwright and Hardie (2012); see also Deaton (2010).

external validity. Investigation of variation in effects across contexts, treatments, or outcomes can thus provide a clear example of cumulative learning.

Assessing mechanisms may also involve, or even require, cumulative learning. Investigating mechanisms raises well-known difficulties. In quantitative work, an important statistical literature underscores the strong assumptions needed for path models and for formal mediation analysis, which can both be seen as methods for learning about mechanisms.[6] Yet learning about mechanisms is far from impossible. Qualitative information on causal processes can lend substantial insights.[7] Understanding mechanisms is also often related to understanding the "active" element in a bundled treatment. Within a single study, planned variation in treatments can thus produce insights about mechanisms, using what Gerber and Green (2012) call "implicit mediation analysis." Gerber et al. (2008a), 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. This approach of leveraging variation in related treatments can also provide an opportunity for cumulative learning *across* disparate sets of studies—often conducted by different researchers. For this purpose, researchers might compare the effects of related or slightly varying treatments—a strategy similar to those aimed at assessing generalizability but that here has a different purpose. Such variation may help illuminate, for example, what component of a bundled treatment X is active or operative (i.e., responsible for an effect) in an original study.

A cumulative process may also be useful for addressing another key question for mechanisms: the distinct effects of a given cause X on related outcomes. Finding an impact of a treatment on some outcomes and 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. While such variation in effects may be present within a particular study, researchers may also leverage findings from disparate research studies to assess mechanisms. We note that knowledge of mechanism is often distinct from questions of the presence or absence of an effect; yet it is often crucial for assessment of causal relations. 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.[8]

Building on the discussion thus far, three empirical strategies appear broadly important for assessing the generalizability of causal effects and/or the mechanisms that account for them:

(1) replicating the same study across different contexts or with different populations;
(2) varying a broader treatment while keeping the same outcome of interest; and
(3) maintaining the same treatments while examining different outcomes, or the same outcome at different points over time.

[6] See e.g. Freedman (2005), Gerber and Green (2012), or Imai, Keele, and Yamamoto (2010).
[7] See Brady and Collier (2010); Mahoney (2010, 2012); Dunning (2012); or Seawright (2016).
[8] An analogy to the biological sciences may be useful; see e.g. Freedman (2009) on how knowledge that infected waste and water causes cholera transmission preceded the theory of germs.

Clearly, (1) is 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).

A range of recent work has used these three strategies in the context of randomized control trials (RCTs), where—at least in principle—researcher control of interventions makes these approaches appear quite plausible. Specifically, the ability to design and implement a treatment of interest in RCTs facilitates replication across contexts as well as exploration of the impacts of different components of a broad treatment, as discussed for example by contributions to the recent Metaketa Initiative of the Evidence in Governance and Politics (EGAP) group.[9] Given parallels between experimental and natural experimental research, the challenges and possibilities for knowledge accumulation in RCTs are an important point of reference. We thus briefly elaborate on the use of these three strategies in field experiments before turning to natural experiments.

Empirical strategy (1) replicates studies across different places to examine the external validity of a causal relationship. For example, Dunning et al. (2019a,b) implement RCTs across six different countries to probe the effects of information about politicians' performance on individuals' vote choice; see also Banerjee et al. (2015). In the first study, the researchers make efforts to establish a standardized intervention and outcome measure in each RCT, allowing for a direct comparison of the effect of information campaigns on voter behavior across studies. More generally, employing similar interventions and outcomes allows researchers to explore the role of contextual factors in shaping causal effects across different empirical cases. Studies that follow this approach may thus maximize external validity, leading to directly comparable findings about the same causal relationship in contexts that, more often than not, vary widely.

Strategy (2) varies components of a broader treatment while maintaining the same empirical setting and outcome variables. This strategy may shed light on generalizability, the mechanisms that drive a causal relationship, or both. For example, Bold et al. (2018) explore the effect of hiring additional teachers on educational outcomes in Kenya. A previous experimental study in Kenya, in which new teachers were hired outside of normal Ministry of Education civil-service channels, found positive impacts on test scores (Duflo et al., 2015). Yet, does this stem from organizational and structural features that are specific to NGOs?[10] Bold et al. (2018) replicate the Duflo et al. treatment of hiring contract teachers through an international NGO but also add a second arm in which the central government hired teachers directly. While they find that the NGO hiring does indeed boost educational outcomes, these positive effects do not hold when teacher recruitment is carried out by the central government, perhaps due to the political economy of teacher hiring and training (including possible patronage dynamics) under the status quo. This reinforces the possibility that the mechanism explaining the finding in Duflo et al. is not just the hiring of additional teachers but also the specific way in which they are hired. A second example of strategy (2) comes from the United States, where Gerber and Green (2000, 2001, 2004) find that phone calls are generally ineffective in mobilizing voters. Subsequent

[9] See footnote 2.

[10] NGOs, for example, may attract a specific population of teachers that is especially qualified. Alternatively, they may be uniquely well positioned to monitor the performance of newly hired teachers.

research, however, has sought to probe whether this applied to all forms of phone banking, or whether personalizing phone interactions could lead to a positive effect. For example, Nickerson (2006) finds that phone banking by campaign volunteers, who tend to personalize phone calls, in fact makes call recipients more likely to vote as compared to individuals who were not contacted (see also Nickerson et al. 2007). In a second study, a professional call center contacted potential voters using a script that was explicitly designed to ensure that interactions with call recipients were personalized and engaging, once again leading to increases in voter turnout (Nickerson 2007). By focusing explicitly on phone banking and varying the degree of personalized calls that voters received, these studies help identify a key mechanism through which this get-out-the-vote strategy can trigger increases in turnout.[11]

Finally, strategy (3) analyzes the effect of a treatment on distinct outcomes in the same context to explore both causal mechanisms and generalizability. These outcomes may consist of the same variable measured at different points in time, or clusters of variables that are expected to be causally related. For example, although a large literature in development economics suggests that investments in the poor can improve living standards and economic outcomes, recent work has also explored the effect of these investments on recipients' political behavior. In Uganda, Blattman et al. (2014) find that providing the poor with microfinance grants led to increases in their future earnings and the likelihood that they were employed. Blattman et al. (2018) go on to probe the political ramifications of the intervention. They find that these grants *did not* lead to greater support for the incumbent political party, suggesting that programmatic government programs targeting the poor may lead to income gains that break clientelistic linkages with poor voters. By exploring the effect of the same treatment on related outcomes, this study sheds light on how government programs can shape individuals' political behavior through increases in economic welfare. In a parallel vein, researchers can also examine outcomes over the medium and long term to explore the persistence of causal effects.[12]

In sum, researcher control over the implementation of a treatment in RCTs can facilitate cumulative learning using strategies (1), (2), and (3). It is important to underscore, however, that unplanned proliferation of experimental studies may not in fact lead to cumulative learning.[13] Substantial recent attention has therefore focused on how experimental research can be better planned and coordinated across research teams to produce cumulative learning—and, ideally, more valid and useful aggregate knowledge. This has been the goal, for example, of the Metaketa Initiative mentioned above. Despite many challenges, coordination in this vein may bring researchers closer to the broader goal of making confident empirical generalizations beyond a specific context or study population and also increasing our knowledge of mechanisms in relations of cause and effect.

[11] An additional example is the alternative treatment arm in the Metaketa Initiative, which allows researchers to embed related, but distinct treatments within otherwise standardized field experiments. See, e.g., Adida et al. (2019); Arias et al. (2019); Platas and Raffler (2019); Buntaine et al. (2019); Lierl and Holmlund (2019); Boas et al. (2019); and Sircar and Chauchard (2019).
[12] For a review of some noteworthy examples of these efforts, see Bouguen et al. (2019).
[13] For an example from the literature on community monitoring, see e.g. Dunning (2019a, Chapter 2).

KNOWLEDGE ACCUMULATION THROUGH NATURAL EXPERIMENTS

Can replication or extension of natural experiments contribute to the dimensions of knowledge accumulation defined in the previous section? This section highlights several paths through which natural experiments can contribute—and have contributed—to cumulative learning. Paralleling the discussion of RCTs in the previous section, Table 1 highlights sets of natural experiments that allow us effectively to leverage empirical strategies (1), (2) or (3) to build knowledge of generalizability and/or mechanisms. For each dimension of knowledge accumulation (first column) and the corresponding empirical strategy (second column), we identify areas of substantive focus (third column) and examples of studies (fourth column) that collectively provide examples of the empirical strategy.

To study similar treatments and outcomes in different empirical settings using strategy (1), for example—and thus make progress on assessing one facet of generalizability—scholars can leverage the diffusion of similar political institutions and social policies across contexts. Though some natural experiments rely on unique features of specific cases to make possible causal identification, others are less idiosyncratic. Indeed, studies often take advantage of institutions that are widely in use across the world. Studies of incumbency advantage that employ close race regression discontinuity (RD) designs provide one prominent example. The spread of democratic elections for both national and local office in recent decades has expanded the potential applicability of close race RDs, permitting studies of incumbency on future vote share and other outcomes in a variety of cases. One notable cluster of recent studies, for example, probes the effect of partisan incumbency in legislative bodies on future electoral outcomes. While research from the Americas[14] and Western Europe[15] finds that incumbents enjoy an advantage in future elections, evidence from other regions[16] suggests that incumbency may at times prove *disadvantageous* to political parties and their candidates. In Table 1, we note a set of studies that collectively allow assessment of the generalizability (and direction) of incumbency effects across the world.

Other political institutions offer similar opportunities for replication across contexts. Table 1 gives examples of studies that allow comprehensive assessment of the impact of similar treatments on similar outcomes in varied contexts, across disparate substantive areas: the effect of gender quotas on women's political participation; the impact of party incumbency on re-election in the legislature; personal (as distinct from partisan) incumbency advantage in cities; and the impact of military conscription on voting behavior and on crime.[17] The use of similar electoral rules or

[14] See, e.g., Lee (2008) on the United States, Kendall and Rekkas (2012) on Canada, and Avelino et al. (2022) and Meireles (2019) on Brazil.

[15] Studies include those on Germany (Hainmueller and Lutz Kern 2008); Ireland (Redmond and Regan 2015); and the United Kingdom (Eggers and Spirling 2017).

[16] Studies that find an incumbency disadvantage include Lee (2020) on India and Ariga et al (2016) on Japan.

[17] Here we refer to "similar" treatments and outcomes for the same reason that work in the Metaketa Initiative refers to "harmonization" of treatment and outcomes (Dunning et al. 2019a,b): exact equivalence across contexts may not be possible, either theoretically or practically. Yet, climbing Sartori's (1970) ladder of abstraction, we can see different operationalizations across contexts as instantiations of the same concept.

other institutions across different contexts allows researchers to build similar natural experimental designs. This then provides opportunities for knowledge accumulation by allowing researchers to generalize about the effects of causes in different settings, despite lacking control over treatment implementation.

Scholars can also leverage the exogenous variation that arises from natural experiments in pursuit of strategy (2)—the study of effects of different classes of a treatment on the same outcome in a single context. As discussed above, this strategy can shed light on both generalizability *and* mechanism. For example, Dunning and Nilekani (2013) examine the effect of caste- and tribe-based electoral quotas through a design similar to a standard discontinuity design, in which they leverage a rotating system of reservations for village council presidents based on population thresholds.[18] They find that caste-based quotas have little effect on the provision of benefits to these minority groups. In related research, Gulzar et al. (2020), using the same outcome measures as Dunning and Nilekani (2013), employ a geographic discontinuity design to analyze the presence of permanent electoral quotas for Scheduled Tribes at the local level.[19] In contrast to Dunning and Nilekani, they report positive effects on the targeting of benefits to marginalized groups when reservations are in effect. These findings suggest that ethnic quotas may have different effects depending on whether they are permanent or temporary (Gulzar et al. 2020). Thus, variations in related treatments across studies with similar outcome measures may shed explanatory light on the mechanisms that lie behind the variation in effects. In addition to this research, Table 1 highlights a second example of this strategy based on work that allows comparison of the effect of partisan and personal incumbency in the US Congress.

Finally, natural experiments can also allow assessment of generalizability and mechanism using empirical strategy (3), i.e. the effects of a similar treatment implemented in a particular context on different outcomes, or on the same outcome at different points of time. Consider, for example, studies that leverage the military draft lottery in the United States during the Vietnam War to explore the causal impact of military service. Studying the effects of the United States' military draft on political attitudes, Erikson and Stoker (2011) find that men with low lottery numbers—who were thus more likely to be drafted into military service—were, on average, more likely to support the Democratic Party, be antiwar, and espouse liberal policy positions. While any number of factors could be driving these results, evidence from Angrist (1990) that draft eligibility caused lower later earnings suggest that Erikson and Stoker's findings may be, at least partially, due to the economic disruptions associated with the risk of conscription (Angrist 1990).[20] Table 1 documents studies in other substantive areas in which natural experiments with similar treatments and varied outcomes can be leveraged to assess generalizability and

[18] Many other scholars explore the effect of caste-based electoral quotas in India, though not all using natural experiments (e.g. Besley et al. 2004, Besley, et al. 2008, Bardhan et al. 2010).
[19] Specifically, both studies use data on the distribution of benefits from the MGNREGA program, along with other outcome measures.
[20] Erikson and Stoker emphasize the risk of conscription associated with a low draft lottery number as their key treatment, whereas Angrist (1990) seeks to pinpoint the impact of military service itself, using draft lottery number as an instrumental variable. Notably, both the economic effects and political effects of the Vietnam draft lottery fade over time (Erikson and Stoker 2001 and Angrist et al. 2011).

mechanism, including research about the effects of gender quotas in India on re-election of female candidates, distribution, and attitudes; the effects of caste quotas on distribution and caste-based violence; and the effects of incumbency on re-election as compared to other (perhaps intermediate) outcomes, such as campaign contributions.

From all appearances, knowledge accumulation through natural experiments is not only possible; it is occurring already. Yet this route to cumulative learning is not without challenges. In the following section, we discuss important limitations on knowledge accumulation through natural experiments and propose several approaches that may further enhance cumulative learning.

Table 1: Select Examples of Knowledge Accumulation Through Natural Experiments

Dimension of Knowledge Accumulation	Empirical Strategy	Substantive Focus	Example studies
Generalizability (effects in varied populations and contexts)	(1) Similar treatment, similar outcome, variation in context	Gender quotas for local executive office and women's political participation	Clayton 2015; Bagues and Campa 2021; Beath et al. 2013
		Partisan incumbency advantage in the legislature	Lee 2008; Eggers and Spirling 2017; Kendall and Rekkas 2012; Hainmueller and Lutz Kern 2008; Redmond and Regan 2015; Ariga et al. 2016; Avelino et al. 2022; Meireles 2019; Golden and Picci 2015; Schiumerini 2015
		Personal incumbency advantage in cities	Trounstine 2011; de Benedictis-Kessner 2018; de Magalhaes 2015; Hyytinen et al. 2018; Weaver 2021.
		Military Drafts and Voting Behavior	Erikson and Stoker 2011; Fize Louis-Sidois 2020; Cáceres-Delpiano et al. 2021
		Military drafts and Crime	Galiani et al. 2011; Lindo and Stoecker 2013; Siminski et al. 2016; Albæk et al. 2017; Lyk-Jensen 2018; Wang and Flores-Lagunes 2020
Generalizability (effects of related treatments); Mechanisms (disaggregating components of treatment)	(2) Different treatment, similar outcome, same context	Targeted benefits in India: Rotating vs. permanent ethnic quotas in villages	Dunning and Nilekani 2013; Gulzar, et al. 2020
		Election in US legislature: Effect of partisan vs. personal incumbency	Lee 2008; Eggers et al. 2015; Fowler and Hall 2014
Generalizability (effects on related outcomes); Mechanisms (assessing active	(3) Similar treatment, different outcome (or outcomes measured at different periods of time), same context	Local gender quotas in India: reelection vs. distribution vs. attitudes	Bhavnani 2009; Beaman et al. 2009; Beaman et al. 2012; Deininger et al. 2015; Chattopadhyay and Duflo 2004; Brulé 2020; Deininger et al. 2020; Iyer et al. 2012;

element of treatment by comparing effects on different outcomes)			Parthasarathy et al. 2019; Turnbull 2019
		Caste quotas in India – distribution vs. attitudes vs. limiting violence	Dunning and Nilekani 2013; Chauchard 2014; Soni 2018;
		Military service; earnings vs. political participation of Vietnam Era Veterans in the US	Angrist 1990; Angrist et al. 2011; Erikson and Stoker 2011
		Incumbency advantage in Brazilian cities: vote share in different levels of government and party building	Novaes 2018; Avelino et al. 2012; Klašnja and Titiunik 2017; Sells 2020; Feierherd 2020
		Incumbency advantage in the US: reelection vs. campaign contributions	Lee 2008; Eggers et al. 2015; Fouirnaies and Hall 2014;

LIMITATIONS AND POSSIBILITIES FOR KNOWLEDGE ACCUMULATION IN NATURAL EXPERIMENTS

Table 1 documents the possibility of successful knowledge accumulation using natural experiments. Yet many areas of research have not seen this type of learning. Why not? There exist a number of limitations to knowledge accumulation using natural experiments, some of which are familiar from both RCTs and observational research.

One set of limitations does indeed stem, to some extent, from the serendipitous nature of natural experiments, as well as the sometimes limited contexts in which researchers can draw on them. Despite our survey in the previous section, which suggests that similar natural experiments may be leveraged across a variety of countries and contexts, it remains the case that studies employing these designs are constrained to those cases where naturally occuring random—or as-if random—variation in treatment assignment is present. This shapes the settings from which researchers can draw inferences about causal effects. For example, though countries across Latin America adopted gender quotas as early as 1991, natural experimental evidence about the effects of these electoral institutions is concentrated in India, where gender quotas have been randomly assigned to single-member electoral districts in some states.[21] Even in settings where a natural experiment is present, additional obstacles—such as a small number of observations—may further circumscribe the feasibility of analysis. It is therefore perhaps no coincidence that many of the close race RD designs referenced in Table 1 were carried out in countries with a large number of districts, such as India, Brazil, and the United States. These challenges limit assessments of generalizability and potentially preclude entirely the natural experimental study of social and political phenomena that do not naturally vary exogenously or do so in quantitatively important ways. We note that these limitations are not exclusive to natural experiments. Field experimental research also focuses disproportionately on contexts where RCTs are more feasible.[22] The availability of data or of variation in treatment status can also sharply limit the "effective sample" in conventional quantitative observational studies, such as those using cross-national or panel fixed-effect regressions (Aronow and Samii 2016). Qualitative research based on techniques such as interviews or shadowing politicians can be constrained by subject willingness (Bussell 2020), and some political contexts (like violent war) may make fieldwork difficult or impossible. However, it may be that natural experiments are especially prone to these limitations.

A further challenge to generalizability, specific to strategy (1), pertains to the comparison of findings from natural experiments in different contexts. Consider two, hypothetical studies of partisan incumbency, one in a context with extensive political patronage and another in a country where parties rely on merit to select political officeholders. What does partisan incumbency mean in these two cases? Do voters associate incumbency with corruption and cronyism, or efficacy and experience? Absent understanding of these aspects of incumbency in each setting, comparisons of observed effects may not be meaningful. Moreover, differences in the design and/or institutional setting in which natural experimental treatments are embedded may lead scholars to use different measurement and coding schema, as well as distinct statistical analyses to draw causal inferences. This can make it difficult to compare findings across different studies, despite causes and effects that are conceptually broadly similar.

There also exist limitations to the accumulation of knowledge through natural experiments using strategy (2), which assesses the effect of different classes of a treatment on the same outcome within a single context. Consider studies that examine the economic effects of military conscription using the draft lottery in the United States during the Vietnam War. We might wonder whether the effects are driven by conscription generally, or depend specifically on conscription during military conflict. However, because the draft lottery was only in place during the Vietnam War (from 1969 to 1973), and thus does not provide exogenous variation in peacetime conscription, we cannot evaluate this hypothesis with the same natural experimental design. In this respect, strategy (2) may be constrained by the absence of different subpopulations from which to test components of a treatment (in the example here, if

[21] The first gender quota in Latin America was adopted in 1991, two years before similar quotas were mandated in India (see Htun 2016 and Chattopadhyay and Duflo 2004).
[22] On this point, see e.g., Humphreys and Weinstein (2009); Blair et al. (2013); Allcott (2015); and Corduneanu-Huci et al. (2021).

conscription had continued after the war, one could have analyzed those conscripted during the war versus those conscripted during peacetime).

Finally, the accumulation of knowledge through strategy (3) must also contend with challenges to inference that arise from analyzing the effects of a treatment on more than one outcome of interest, particularly when the same natural experimental design is used across multiple studies. First, there is no guarantee in a given substantive area that as-if random variation will occur in the presence of clusters of outcomes of scholarly interest. Second, even if it does, estimating the effects of a single treatment on a battery of outcome variables risks observing statistically significant effects that stem from chance variation, rather than representing valid estimates of true causal effects.[23] This risk is of particular concern across studies that leverage the same natural experiment to examine different outcome variables. As the number of outcomes explored across these studies increases, so too does the likelihood of observing statistically significant results due to chance. Moreover, because the submission of pre-analysis plans is not standard among studies using natural experiments and null findings often go unreported, the universe of tests performed in other research is generally unknown. This complicates efforts to employ formal, statistical tools to adjust results to account for multiple statistical comparisons.[24]

Encompassing all three empirical strategies, and building on the final point in the previous paragraph, are challenges that stem from the nature of the academic production of knowledge. As we noted in our definition of knowledge accumulation, the nature of cumulative learning may reflect career incentives or other motivations that researchers have for undertaking particular strands of research. Natural experiments are no exception. One prominent obstacle is publication bias. As is well documented elsewhere, research journals and other academic outlets are more likely to publish apparent evidence of causal effects.[25] This may prevent the dissemination of studies—natural experimental, observational, and experimental—with null findings.[26] Similarly, even in the presence of significant results, it may be difficult to publish studies that assess generalizability and/or causal mechanisms if the findings are consistent with those of existing research. The prioritization of originality and new findings—while understandable and laudable in some ways—can therefore also undercut cumulative learning. These forms of publication bias risk limiting cumulative learning across all three strategies described in Table 1.

How can these limitations best be mitigated? Some are inherent to natural experimental designs, which are constrained by the presence and scope of naturally occurring, exogenous variation. We cannot, after all, hope to leverage a natural experiment where none exists—though enterprising researchers have greatly expanded the empirical scope of these designs in recent years through their discovery of new possibilities for natural experiments.

[23] For a review of the issue of multiple statistical comparisons, as well as its risks to knowledge accumulation generally, see Sterling (1959), Humphreys et al. (2013), and Dunning (2016).
[24] Researchers can adjust p-values to account for multiple comparisons using Bonferroni corrections, the false-discovery-rate correction of Benjamini & Hochberg (1995), among other alternatives.
[25] See, e.g., Gerber et al. (2008b) and Gerber and Malhotra (2008).
[26] For a discussion of this concern in experimental settings, see, e.g. Dunning (2016); Malhotra (2021).

Other limitations, stemming from a lack of consideration of and coordination with existing natural experimental research, may be more tractable, however. While natural experiments are unplanned, there is no need for studies using these designs to follow a similar approach to learning. Indeed, some of the examples of the studies described in this chapter directly seek to engage with and build on the designs, evidence, analysis, and findings of previous work (one clear case is Gulzar et al. 2020's discussion). Yet many engage less explicitly with previous research, leaving it up to other scholars to draw ex-post conclusions about how their findings relate to those of existing work. Promoting greater coordination and transparency across natural experimental research can foster further knowledge accumulation by proactively encouraging scholars to build on one another's work.

Practices that promote transparency can also play an important role in encouraging knowledge accumulation through natural experiments. One such practice is the registration of studies in advance of data collection and analysis. Another is the submission of pre-analysis plans that outline the measures and indices studies will employ and that specify the hypotheses and statistical tests to be conducted. Preregistration and preanalysis plans can encourage researchers to consider directly the ways in which their coding, measurement, and analysis choices relate to those carried out in similar studies. They also make it more straightforward for scholars to replicate the data collection and empirical analyses of prior research, as well as address concerns about multiple statistical comparisons. They can thus facilitate knowledge accumulation by furthering the comparability and replication of studies. While they remain most common in field, survey, and lab experiments, scholars have increasingly advocated their benefits in non-experimental studies.[27]

Proactive coordination of natural experimental research across studies can also foster greater knowledge accumulation. As discussed above, professional incentives that reward originality and new findings can undercut cumulative learning. One potential solution to this problem is to alter these incentives, for example by providing financial support to encourage coordination among clusters of natural experimental studies. There are, of course, unique challenges to this approach, since it requires first the presence of a natural experiment. However, one could imagine offering support to studies that have identified similar natural experiments across empirical cases—for example close race RDs—or within the same case but focused on different outcomes. This financial support could help structure clusters of natural experimental studies in which scholars coordinate their data measurement and empirical analysis, as well as provide spaces for dialogue across studies. It could also facilitate the acquisition of complimentary data, such as from surveys and other qualitative and quantitative sources, to help elucidate causal relationships within and across the coordinated studies. Donors may also view this support as attractive; fostering coordinated natural experimental research furthers knowledge accumulation about causes that are difficult to manipulate experimentally, but that may be especially important for understanding social and political effects and the causes of human welfare. Moreover, supporting these studies may be more cost effective, since researchers are not responsible for

[27] See, e.g., Burlig (2018); Jacobs (2020); and Ofosu and Posner (2021). For a discussion of the benefits of preregistration and preanalysis plans in experimental settings, see Dunning (2016).

treatment implementation in natural experiments. Supporting knowledge accumulation on this domain may thus appeal to both scholars and donors alike.

Overall, we find that the case against cumulative learning through natural experiments should not be overstated. In this chapter, we have documented how natural experimental designs can foster and already have fostered knowledge accumulation, even as we have underscored limitations and challenges in the approach. The three empirical strategies we described help illuminate how natural experimental research may be used to contribute to two key dimensions of cumulative learning—the assessment of generalizability and mechanisms. Natural experiments can boost cumulative learning with respect to social and political phenomena that are not easily manipulated in experimental research, while avoiding some of the challenges for causal inference that beset many conventional observational studies. They can therefore play a critical role in advancing our understanding of social and political phenomena, in tandem with other research designs and methods. Knowledge accumulation through natural experiments should therefore be further facilitated, including through the possibilities and new practices we have outlined here.

REFERENCES

Acemoglu, Daron. 2010. “Theory, General Equilibrium, and Political Economy in Development Economics.” *Journal of Economic Perspectives* 24(3):17–32.

Adida, Claire, Jessica Gotlieb, Eric Kramon, and Gwyneth McClendon. 2019. “Under What Conditions Does Performance Information Influence Voting Behavior? Lessons from Benin.” Pp. 81–117 in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, *Cambridge Studies in Comparative Politics*, edited by C. McIntosh, G. Nellis, G. Grossman, M. Humphreys, S. D. Hyde, and T. Dunning. Cambridge: Cambridge University Press.

Albæk, Karsten, Søren Leth-Petersen, Daniel le Maire, and Torben Tranæs. 2017. “Does Peacetime Military Service Affect Crime?” *The Scandinavian Journal of Economics* 119(3):512–40.

Allcott, Hunt. 2015. “Site Selection Bias in Program Evaluation.” *The Quarterly Journal of Economics* 130(3):1117–65.

Angrist, Joshua D. 1990. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.” *The American Economic Review* 80(3):313–36.

Angrist, Joshua D., Stacey H. Chen, and Jae Song. 2011. “Long-Term Consequences of Vietnam-Era Conscription: New Estimates Using Social Security Data.” *American Economic Review* 101(3):334–38.

Arias, Eric, Horacio Larraguy, John Marshall, and Pablo Querubín. 2019. “When Does Information Increase Electoral Accountability? Lessons from a Field Experiment in Mexico.” Pp. 118–55 in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, *Cambridge Studies in Comparative Politics*, edited by C. McIntosh, G. Nellis, G. Grossman, M. Humphreys, S. D. Hyde, and T. Dunning. Cambridge: Cambridge University Press.

Ariga, Kenichi, Yusaka Horiuchi, Roland Mansilla, and Michio Umeda. 2016. “No Sorting, No Advantage: Regression Discontinuity Estimates of Incumbency Advantage in Japan.” *Electoral Studies* 43.

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

Avelino, George, Ciro Biderman, and Leonardo S. Barone. 2012. “Articulações intrapartidárias e desempenho eleitoral no Brasil.” *Dados* 55:987–1013.

Avelino, George, Ciro Biderman, and Scott Desposato. 2022. “Sources of the Incumbency (Dis)Advantage.” *Brazilian Political Science Review* 16(1).

Bagues, Manuel, and Pamela Campa. 2021. “Can Gender Quotas in Candidate Lists Empower Women? Evidence from a Regression Discontinuity Design.” *Journal of Public Economics* 194:104315.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. “From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application.” *Journal of Economic Perspectives* 31(4):73–102.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. “A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries.” *Science*.

Bardhan, Pranab K., Mookherjee Dilip, and Parra Torrado Monica. 2010. “Impact of Political Reservations in West Bengal Local Governments on Anti-Poverty Targeting.” *Journal of Globalization and Development* 1(1):1–38.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. “Powerful Women: Does Exposure Reduce Bias?” *The Quarterly Journal of Economics* 124(4):1497–1540.

Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova. 2012. “Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India.” *Science*.

Beath, Andrew, Fotini Christia, and Ruben Enikolopov. 2013. “Empowering Women through Development Aid: Evidence from a Field Experiment in Afghanistan.” *American Political Science Review* 107(3):540–57.

de Benedictis-Kessner, Justin. 2018. “Off-Cycle and Out of Office: Election Timing and the Incumbency Advantage.” *The Journal of Politics* 80(1):119–32.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Besley, Timothy, Rohini Pande, Lupin Rahman, and Vijayendra Rao. 2004. “The Politics of Public Good Provision: Evidence from Indian Local Governments.” *Journal of the European Economic Association* 2(2/3):416–26.

Besley, Timothy, Rohini Pande, and Vijayendra Rao. 2008. “The Political Economy of Gram Panchayats in South India.” Pp. 243–64 in *Development in Karnataka: Challenges of Governance, Equity, and Empowerment*, edited by G. K. Kadekodi, S. M. R. Kanbur, and V. Rao. New Delhi: Academic Foundation.

Bhavnani, Rirkhil R. 2009. “Do Electoral Quotas Work after They Are Withdrawn? Evidence from a Natural Experiment in India.” *The American Political Science Review* 103(1):23–35.

Blair, Graeme, Radha K. Iyengar, and Jacob N. Shapiro. 2013. “Where Policy Experiments Are Conducted in Economics and Political Science: The Missing Autocracies.”

Blair, Graeme, Jeremy M. Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A. Blair, Ali Cheema, Ahsan Farooqui, Thiemo Fetzer, Guy Grossman, Dotan Haim, Zulfiqar Hameed, Rebecca Hanson, Ali Hasanain, Dorothy Kronick, Benjamin S. Morse, Robert Muggah, Fatiq Nadeem, Lily L. Tsai, Matthew Nanes, Tara Slough, Nico Ravanilla, Jacob N. Shapiro, Barbara Silva, Pedro C. L. Souza, and Anna M. Wilke.

2021. “Community Policing Does Not Build Citizen Trust in Police or Reduce Crime in the Global South.” *Science* 374(6571):eabd3446.

Blattman, Christopher, Mathilde Emeriau, and Nathan Fiala. 2018. “Do Anti-Poverty Programs Sway Voters? Experimental Evidence from Uganda.” *The Review of Economics and Statistics* 100(5):891–905.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez. 2014. “Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda.” *The Quarterly Journal of Economics* 129(2):697–752.

Boas, Taylor C., F. Daniel Hidalgo, and Marcus André Melo. 2019. “Horizontal But Not Vertical: Accountability Institutions and Electoral Sanctioning in Northeast Brazil.” Pp. 257–86 in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, *Cambridge Studies in Comparative Politics*, edited by C. McIntosh, G. Nellis, G. Grossman, M. Humphreys, S. D. Hyde, and T. Dunning. Cambridge: Cambridge University Press.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng’ang’a, and Justin Sandefur. 2018. “Experimental Evidence on Scaling up Education Reforms in Kenya.” *Journal of Public Economics* 168:1–20.

Bouguen, Adrien, Yue Huang, Michael Kremer, and Edward Miguel. 2019. “Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics.” *Annual Review of Economics* 11(1):523–61.

Brady, Henry E., and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield Publishers.

Brulé, Rachel E. 2020. “Reform, Representation, and Resistance: The Politics of Property Rights’ Enforcement.” *The Journal of Politics* 82(4):1390–1405.

Buntaine, Mark T., Sarah S. Bush, Ryan Jablonski, Daniel L. Nielson, and Paula M. Pickering. 2019. “Budgets, SMS Texts, and Votes in Uganda.” Pp. 188–220 in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, *Cambridge Studies in Comparative Politics*, edited by C. McIntosh, G. Nellis, G. Grossman, M. Humphreys, S. D. Hyde, and T. Dunning. Cambridge: Cambridge University Press.

Burlig, Fiona. 2018. “Improving Transparency in Observational Social Science Research: A Pre-Analysis Plan Approach.” *Economics Letters* 168:56–60.

Bussell, Jennifer. 2020. “Shadowing as a Tool for Studying Political Elites.” *Political Analysis* 28(4):469–86.

Cáceres-Delpiano, Julio, Antoni-Italo De Moragas, Gabriel Facchini, and Ignacio González. 2021. “Intergroup Contact and Nation Building: Evidence from Military Service in Spain.” *Journal of Public Economics* 201:104477.

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

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

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India.” *Econometrica* 72(5):1409–43.

Chauchard, Simon. 2014. “Can Descriptive Representation Change Beliefs about a Stigmatized Group? Evidence from Rural India.” *The American Political Science Review* 108(2):403–22.

Clayton, Amanda. 2015. “Women’s Political Engagement Under Quota-Mandated Female Representation: Evidence From a Randomized Policy Experiment.” *Comparative Political Studies* 48(3):333–69.

Corduneanu-Huci, Cristina, Michael T. Dorsch, and Paul Maarek. 2021. “The Politics of Experimentation: Political Competition and Randomized Controlled Trials.” *Journal of Comparative Economics* 49(1):1–21.

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

Deininger, Klaus, Songqing Jin, Hari K. Nagarajan, and Fang Xia. 2015. “Does Female Reservation Affect Long-Term Political Outcomes? Evidence from Rural India.” *Journal of Development Studies* 51(1):32–49.

Deininger, Klaus, Hari K. Nagarajan, and Sudhir K. Singh. 2020. “Women’s Political Leadership and Economic Empowerment: Evidence from Public Works in India.” *Journal of Comparative Economics* 48(2):277–91.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. “School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools.” *Journal of Public Economics* 123:92–110.

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

Dunning, Thad. 2016. “Transparency, Replication, and Cumulative Learning: What Experiments Alone Cannot Achieve.” *Annual Review of Political Science* 19(1):541–63.

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

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, Gareth Nellis, Claire L. Adida, Eric Arias, Clara Bicalho, Taylor C. Boas, Mark T. Buntaine, Simon Chauchard, Anirvan Chowdhury, Jessica Gottlieb, F. Daniel Hidalgo, Marcus Holmlund, Ryan Jablonski, Eric Kramon, Horacio Larreguy, Malte Lierl, John Marshall, Gwyneth McClendon, Marcus A. Melo, Daniel L. Nielson, Paula M. Pickering, Melina R. Platas, Pablo Querubín, Pia Raffler, and Neelanjan Sircar. 2019b. “Voter Information Campaigns and Political Accountability: Cumulative Findings from a Preregistered Meta-Analysis of Coordinated Trials.” *Science Advances* 5(7).

Dunning, Thad, and Janhavi Nilekani. 2013. “Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils.” *American Political Science Review* 107(1):35–56.

Egami, Naoki, and Erin Hartman. 2022. “Elements of External Validity: Framework, Design, and Analysis.”

Eggers, Andrew C., Anthony Fowler, Jens Hainmueller, Andrew B. Hall, and James M. Snyder Jr. 2015. “On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races.” *American Journal of Political Science* 59(1):259–74.

Eggers, Andrew C., and Arthur Spirling. 2017. “Incumbency Effects and the Strength of Party Preferences: Evidence from Multiparty Elections in the United Kingdom.” *The Journal of Politics* 79(3):903–20.

Erikson, Robert S., and Laura Stoker. 2011. “Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes.” *American Political Science Review* 105(2):221–37.

Feierherd, Germán. 2020. “How Mayors Hurt Their Presidential Ticket: Party Brands and Incumbency Spillovers in Brazil.” *The Journal of Politics* 82(1):195–210.

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

Fize, Etienne, and Charles Louis-Sidois. 2020. “Military Service and Political Behavior: Evidence from France.” *European Economic Review* 122:103364.

Fouirnaies, Alexander, and Andrew B. Hall. 2014. “The Financial Incumbency Advantage: Causes and Consequences.” *The Journal of Politics* 76(3):711–24.

Fowler, Anthony, and Andrew B. Hall. 2014. “Disentangling the Personal and Partisan Incumbency Advantages: Evidence from Close Elections and Term Limits.” *Quarterly Journal of Political Science* 9(4):501–31.

Freedman, David. 2005. *Statistical Models: Theory and Practice*. Cambridge University Press.

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.

Galiani, Sebastian, Martín A. Rossi, and Ernesto Schargrodsky. 2011. “Conscription and Crime: Evidence from the Argentine Draft Lottery.” *American Economic Journal: Applied Economics* 3(2):119–36.

Gerber, Alan S., and Donald P. Green. 2000. “The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment.” *American Political Science Review* 94(3):653–63.

Gerber, Alan S., and Donald P. Green. 2001. “Do Phone Calls Increase Voter Turnout?: A Field Experiment.” *The Public Opinion Quarterly* 65(1):75–85.

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. 2008a. “Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment.” *American Political Science Review* 102(1):33–48.

Gerber, Alan S., Donald Green, and David Nickerson. 2008b. “Testing for Publication Bias in Political Science.” *Political Analysis* 9(4).

Gerber, Alan S., and Neil Malhotra. 2008. “Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals.” *Quarterly Journal of Political Science* 3(3).

Gerring, John. n.d. “Evaluating the Cumulation of Knowledge in the Social Sciences.” in *The Oxford Handbook of Methodological Pluralism*, edited by J. Box-Steffensmeier, D. Christenson, and V. Sinclair-Chapm.

Golden, Miriam A., and Lucio Picci. 2015. “Incumbency Effects under Proportional Representation: Leaders and Backbenchers in the Postwar Italian Chamber of Deputies.” *Legislative Studies Quarterly* 40(4):509–38.

Green, Donald P., and Alan S. Gerber. 2004. *Get Out the Vote!: How to Increase Voter Turnout*. Brookings Institution Press.

Gulzar, Saad, Nicholas Haas, and Benjamin Pasquale. 2020. “Does Political Affirmative Action Work, and for Whom? Theory and Evidence on India’s Scheduled Areas.” *American Political Science Review* 114(4):1230–46.

Hainmueller, Jens, and Holger Lutz Kern. 2008. “Incumbency as a Source of Spillover Effects in Mixed Electoral Systems: Evidence from a Regression-Discontinuity Design.” *Electoral Studies* 27(2):213–27.

Htun, Mala. 2016. *Inclusion without Representation in Latin America: Gender Quotas and Ethnic Reservations*. Cambridge University Press.

Humphreys, Macartan, Samii Cyrus, Alexandra Scacco, Julio S. Solís Arce, and Anna M. Wilke. n.d. “How to Facilitate Learning Across Studies? The ‘Rolling Metaketa’ Approach to Knowledge Accumulation.” in *The Oxford Handbook of Methodological Pluralism*, edited by J. Box-Steffensmeier, D. Christenson, and V. Sinclair-Chapman.

Humphreys, Macartan, Raul Sánchez de la Sierra, and Peter Van der Windt. 2013. “Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration.” *Political Analysis* 21(1):1–20.

Humphreys, Macartan, and Jeremy M. Weinstein. 2009. “Field Experiments and the Political Economy of Development.” *Annual Review of Political Science* 12(1):367–78.

Hyytinen, Ari, Jaakko Meriläinen, Tuukka Saarimaa, Otto Toivanen, and Janne Tukiainen. 2018. “When Does Regression Discontinuity Design Work? Evidence from Random Election Outcomes.” *Quantitative Economics* 9(2):1019–51.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects.” *Statistical Science* 25(1):51–71.

Iyer, Lakshmi, Anandi Mani, Prachi Mishra, and Petia Topalova. 2012. “The Power of Political Voice: Women’s Political Representation and Crime in India.” *American Economic Journal: Applied Economics* 4(4):165–93.

Jacobs, Alan M. 2020. “Pre-Registration and Results-Free Review in Observational and Qualitative Research.” Pp. 221–64 in *The Production of Knowledge: Enhancing Progress in Social Science*, *Strategies for Social Inquiry*, edited by C. Elman, J. Mahoney, and J. Gerring. Cambridge: Cambridge University Press.

Kendall, Chad, and Marie Rekkas. 2012. “Incumbency Advantages in the Canadian Parliament.” *Canadian Journal of Economics/Revue Canadienne d’économique* 45(4):1560–85.

Klašnja, Marko, and Rocío Titiunik. 2017. “The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability.” *American Political Science Review* 111(1):129–48.

Kocher, Matthew A., and Nuno P. Monteiro. 2016. “Lines of Demarcation: Causation, Design-Based Inference, and Historical Research.” *Perspectives on Politics* 14(4):952–75.

Lee, Alexander. 2020. “Incumbency, Parties, and Legislatures: Theory and Evidence from India.” *Comparative Politics*52(2):311–31.

Lee, David S. 2008. “Randomized Experiments from Non-Random Selection in U.S. House Elections.” *Journal of Econometrics* 142(2):675–97.

Lierl, Malte, and Marcus Holmlund. 2019. “Performance Information and Voting Behavior in Burkina Faso’s Municipal Elections: Separating the Effects of Information Content and Information Delivery.” Pp. 221–56 in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, *Cambridge Studies in Comparative Politics*, edited by C. McIntosh, G. Nellis, G. Grossman, M. Humphreys, S. D. Hyde, and T. Dunning. Cambridge: Cambridge University Press.

Lindo, Jason M., and Charles Stoecker. 2014. “Drawn into Violence: Evidence on ‘What Makes a Criminal’ from the Vietnam Draft Lotteries.” *Economic Inquiry* 52(1):239–58.

Lyk-Jensen, Stéphanie Vincent. 2018. “Does Peacetime Military Service Affect Crime? New Evidence from Denmark’s Conscription Lotteries.” *Labour Economics* 52:245–62.

Magalhaes, Leandro De. 2015. “Incumbency Effects in a Comparative Perspective: Evidence from Brazilian Mayoral Elections.” *Political Analysis* 23(1):113–26.

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.

Mahoney, James. 2010. “After KKV: The New Methodology of Qualitative Research.” *World Politics* 62(1):120–47.

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

Malhotra, Neil. 2021. “Threats to the Scientific Credibility of Experiments: Publication Bias and P-Hacking.” Pp. 354–68 in *Advances in Experimental Political Science*, edited by D. P. Green and J. N. Druckman. Cambridge: Cambridge University Press.

Meireles, Fernando. 2019. “Carreiras Políticas na Câmara dos Deputados: Uma Análise Quase-Experimental.” *Dados* 62.

Nickerson, David W. 2006. “Volunteer Phone Calls Can Increase Turnout: Evidence From Eight Field Experiments.” *American Politics Research* 34(3):271–92.

Nickerson, David W. 2007. “Quality Is Job One: Professional and Volunteer Voter Mobilization Calls.” *American Journal of Political Science* 51(2):269–82.

Nickerson, David W., Ryan D. Friedrichs, and David C. King. 2006. “Partisan Mobilization Campaigns in the Field: Results from a Statewide Turnout Experiment in Michigan.” *Political Research Quarterly* 59(1):85–97.

Novaes, Lucas M. 2018. “Disloyal Brokers and Weak Parties.” *American Journal of Political Science* 62(1):84–98.

Ofosu, George K., and Daniel N. Posner. 2021. “Pre-Analysis Plans: An Early Stocktaking.” *Perspectives on Politics*19(4):1–17.

Parthasarathy, Ramya, Vijayendra Rao, and Nethra Palaniswamy. 2019. “Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India’s Village Assemblies.” *American Political Science Review* 113(3):623–40.

Platas, Melina R., and Pia Raffler. 2019. “Candidate Videos and Vote Choice in Ugandan Parliamentary Elections.” Pp. 156–87 in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, *Cambridge Studies in Comparative Politics*, edited by C. McIntosh, G. Nellis, G. Grossman, M. Humphreys, S. D. Hyde, and T. Dunning. Cambridge: Cambridge University Press.

Redmond, Paul, and John Regan. 2015. “Incumbency Advantage in a Proportional Electoral System: A Regression Discontinuity Analysis of Irish Elections.” *European Journal of Political Economy* 38:244–56.

Sartori, Giovanni. 1970. “Concept Misformation in Comparative Politics.” *The American Political Science Review*64(4):1033–53.

Schiumerini, Luis Enrique. 2015. “Incumbency and Democracy in South America.” Dissertation, Yale University, New Haven, CT.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

Sells, Cameron J. 2020. “Building Parties from City Hall: Party Membership and Municipal Government in Brazil.” *The Journal of Politics* 82(4):1576–89.

Siminski, Peter, Simon Ville, and Alexander Paull. 2016. “Does the Military Turn Men into Criminals? New Evidence from Australia’s Conscription Lotteries.” *Journal of Population Economics* 29(1):197–218.

Sircar, Neelanjan, and Simon Chauchard. 2019. “Dilemmas and Challenges of Citizen Information Campaigns: Lessons from a Failed Experiment in India.” Pp. 287–312 in *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, *Cambridge Studies in Comparative Politics*, edited by C. McIntosh, G. Nellis, G. Grossman, M. Humphreys, S. D. Hyde, and T. Dunning. Cambridge: Cambridge University Press.

Soni, Suparna. 2018. “Political Quotas, NGO Initiatives and Dalits’ Human Rights in Rural India.” *Journal of Human Rights Practice* 10(3):388–405.

Sterling, Theodore D. 1959. “Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--Or Vice Versa.” *Journal of the American Statistical Association* 54(285):30–34.

Trounstine, Jessica. 2011. “Evidence of a Local Incumbency Advantage.” *Legislative Studies Quarterly* 36(2):255–80.

Turnbull, Brian. 2021. “Quotas as Opportunities and Obstacles: Revisiting Gender Quotas in India.” *Politics & Gender*17(2):324–48.

Wang, Xintong, and Alfonso Flores-Lagunes. 2020. “Conscription and Military Service: Do They Result in Future Violent and Non-Violent Incarcerations and Recidivism?” *Journal of Human Resources* 0418.

Weaver, Julie Anne. 2021. “Electoral Dis-Connection: The Limits of Reelection in Contexts of Weak Accountability.” *The Journal of Politics* 83(4):1462–77.