# Good Description*

Daniel de Kadt† Anna Grzymala-Busse‡

May 14, 2025

**Abstract**

What distinguishes 'good' description from 'mere' description? We propose a framework for evaluating descriptive empirical social science, premised on the idea that descriptive research – like all empirical social science – should be grounded in theory. First, we articulate the social scientific purpose of description, which provides criteria for assessing descriptive *questions*. Good descriptive questions seek to uncover facts in need of explanation, build theory, or help to evaluate and revise theory. Second, we articulate three ideal characteristics of descriptive *analyses*. Good descriptive analyses are clear, comparable, and complete. Such analyses should be tightly and transparently linked to specific research questions (and thus theory), well contextualized, and as comprehensive – in their measurement and specification – as possible.

*For their very thoughtful comments, we are grateful to Lisa Blaydes, Alex Coppock, Vicky Fouka, Sean Gailmard, Marcus Kreuzer, David Laitin, Andrew Little, Tom Pepinsky, Solé Prillaman, Carlisle Rainey, Melissa Sands, Arthur Spirling, and the participants at the 2024 APSA Foundational Concepts panel and New Directions in Qualitative Research Panel and the Berkeley Methods Workshop.

†Department of Methodology, London School of Economics and Political Science.

‡Department of Political Science, Stanford University.

## 1. Introduction

Description is often seen as a lesser type of social scientific inquiry, both in political science and beyond. In an important intervention on the topic, Gerring (2012) notes that description is often "derided in favor of causal analysis," implying not only that the discipline has a preference for causal research, but an active distaste for "mere" description. This concern has become widely shared in recent years (see, for example Kreuzer (2019) and Holmes et al. (2024)), and new journals have even been established in response (Munger et al., 2021).

Yet at very the same time, important and highly cited descriptive papers are published in leading political science outlets every year. These include studies of democratic transitions, state censorship, representation, international conditionality, political ideology, regime types, and polarization, among many others (e.g. Przeworski and Limongi (1997); Coppedge et al. (2011); King et al. (2013); Bonica (2014); Tausanovitch and Warshaw (2014); Hughes et al. (2019); Davenport (2016); Holland (2016); Benoit et al. (2019); Schimmelfennig and Sedelmeier (2020); Gerring et al. (2021); Jefferson (2023)). Likewise, prominent recent books in political science have made extensive use of description, and have won disciplinary prizes (Beissinger, 2022; Blaydes, 2018; Daly, 2022; Stasavage, 2020; Wang, 2022). Fundamentally, the paradox of descriptive work is that despite its seemingly low status, "much of the empirical work of political scientists and theories that they construct are a direct product of description" (Grimmer, 2015, 80).

We argue that description is not a lesser mode of inquiry, but a fundamental component of the social scientific enterprise. Its status is so contested because social scientists lack a coherent framework for separating the descriptive work that is generally valued from that which is not. This has two implications. Absent a clear articulation of the social scientific purpose of description there remains uncertainty about what makes for good descriptive *questions*. Likewise, absent an articulation of the ideal properties of empirical efforts to answer descriptive questions there is uncertainty about what makes for good descriptive *analyses*. By contrast, there are far more developed perspectives on the

scientific purpose of causal inference, as well as a reasonably clear set of standards against which to measure applied work (see for instance the many textbooks and 'checklist' or 'how to' articles about specific methods).

We develop a framework for the evaluation of both descriptive questions and descriptive analyses. We seek to distinguish 'good' from 'mere' description, in a way that is useful whether we are engaged in description as scholars or assessing it as readers, advisors, reviewers, or editors. We begin from the premise that while descriptive and causal inference do not represent different modes of social scientific inquiry. While different types of social scientific questions may require different research designs, different assumptions, different data, and different statistical or qualitative methods, all of these questions – and subsequent analyses – contribute to the same fundamental scientific goal, the incremental development and evaluation of theories and explanations of the social world.

We argue that description is a research task that contributes to modern social scientific inquiry when it is informed by, motivates, or revises relevant and important theory. Good description articulates and asks relevant and important *questions* about 'how the world is or was,' and provides answers that are useful through well-calibrated and meaningful *analyses*. In this way, good description is no different from good causal inference, yet it stands on its own merits as an important empirical social scientific task, designed to tackle specific types of questions, and with clear criteria for evaluation. How much we are able to learn from empirical social science hinges on whether the empirical exercises we perform are informative about important concepts and theories. While social scientific theories are usually causal (we expand on this in Section 2), many of these derived research questions need not be. Many specific research questions are instead descriptive – questions about 'how the world is or was.' Descriptive questions merit descriptive analyses – empirical analytical exercises designed to provide meaningful, credible, and interpretable insights about 'how the world is or was.'

We offer two specific contributions. First, we articulate the social scientific goals of description, and in doing so explain what makes for a good descriptive question. We argue that the goal of good description should be to engage with phenomena that are meaningful and important. In this way, good

description is invariably theory-led, even if it may serve to highlight the absence of well-developed theory. Good description may help build theory by informing researchers of phenomena in need of explanation, uncovering 'patterns' or 'facts' or 'anomalies' or 'puzzles' that require consideration. Knowing (something about) the state of the world is a necessary, if not always sufficient, empirical task in the process of theory generation. Good description may also help to evaluate theory – its plausibility or its completeness – by assessing whether the implications of theory hold, and assessing the plausibility of different approaches to studying that theory. We detail these features in Section 2.

Second, we articulate the ideal characteristics of good descriptive analyses. We propose three ideal characteristics of empirical exercises designed to answer descriptive research questions. Good description requires analyses that are clear, comparable, and complete. Clear analyses are those are tightly and transparently linked to theory and well-calibrated to the specific question of interest. Comparable analyses are those that are well contextualized, such that any insights learned can be understood in the context of pre-existing (and future) evidence. Complete analyses are those that are as comprehensive – in terms of measurement and specification – as possible. We argue that one should be able to assess any given study in terms of these three criteria, and we articulate them in detail in Section 3. We conclude by providing examples of how good description can improve other analyses, by specifying relevant analytical contexts and examining causal mechanisms in Section 4.

## 2. Theories, Questions, and Description

Empirical social scientific research connects *theories*, research *questions* derived from those theories, and empirical *analyses* that seek to answer those questions. By social scientific *theories* we specifically mean causal theories – theories that are about the causal connections between variables or concepts. Humphreys and Jacobs emphasize that social scientific theory tends to be explicitly causal, a type of explanation that "provides an account of how or under what conditions a set of causal relationships

operate" (Humphreys and Jacobs, 2023, 163).[1]

Throughout this paper we will work with a toy example, a simple causal *theory*: phenomenon A affects phenomenon B in some unspecified way and with some unspecified magnitude. Formally, in the language of directed acyclic graphs (DAGs), we might represent this theory as A $\longrightarrow$ B.

Given this toy theory, researchers could ask a broad range of plausible research *questions*. As we discuss below, they could ask questions about the conceptual nature of A and B – for example, what conceptual properties constitute A and B? They could ask questions about empirical properties or characteristics of A and B – for example, what are they empirically, when do they occur, and how often? They could ask questions about the association between A and B – for example, do A and B typically co-occur or not, or is their co-occurrence conditional on some background variable? They could ask causal questions – for example, does A indeed affect B, is A necessary and/or sufficient for B, under what conditions does this causal relationship hold, or what is the magnitude of any causal relationship? They may also ask mechanistic questions – in what manner does A affect B?

There are two things worth noting about all these questions. First, they are all derived from theory, and that is what gives them scientific meaning and interest. If there is no social scientific theory that features phenomenon A, the value of describing the properties of phenomenon A is generally (though perhaps not always) low. Second, the answers to these questions may provide important insights into the plausibility of the theory as stated, even without an explicit and dispositive test. At a very simple level, if we do not know whether or when A and B occur, then we know very little about the core theoretical claim that A affects B.

Theories, then, can generate both causal and descriptive propositions, which in turn may inspire different research questions. How can we tell the difference between causal and descriptive questions? There is a tradition in the social sciences of trying to define description using natural language. For example, Gerring (2012)[72-73] argues that descriptive questions focus on "*what* questions (e.g., when,

[1]Our conceptualization of theory is not exhaustive: some scholars argue that there are 'ontological' theories, which conceptualize and categorize phenomena (Goertz and Mahoney, 2012).

whom, out of what, in what manner) about a phenomenon or a set of phenomena," distinct from causal questions that "attempt to answer *why* questions" (emphasis in original). Similarly, Holmes et al. (2024) define descriptive questions as those asking "who, what, when, where, and how".

While such language-based definitions are useful, they will always remain fuzzy.[2] We propose a different definition, drawing on Holland (1986). The distinguishing feature of causal and descriptive questions is manipulability: if at least one concept or variable in a given research question is required to be manipulable for the question to be answerable, then that question is causal. If manipulability is not necessary to answer the question, then that question is descriptive. For example, if we ask whether A and B correlate we do not require manipulability of either A or B – we are not engaged in counterfactual reasoning and so the question is descriptive.

Such situations are not purely theoretical – for example, there is a broad consensus that an individual's race cannot affect their wages because an individual's race cannot be manipulated. Racial discrimination in the labor market, which is manipulable, can and does affect an individual's wages (Sen and Wasow, 2016). We might ask whether people of one race group earn less than people of another, understanding that this is a descriptive question precisely because race is not manipulable. The question asks about the state of the world, not the counterfactual state of the world under manipulation. By contrast, we might ask whether certain labor market interventions or arrangements reduce (or exacerbate) any racial earnings gap, thus asking a causal question. From a language-based perspective, we characterize description as addressing questions of the form 'what is the world like?' Causal inference, by contrast, addresses questions of the form 'what would some counterfactual state of the world have been like?'

Returning to our toy theory, many questions we may pose about the concepts or phenomena A and B are what one might consider canonically *descriptive* questions – they seek only to describe the state

[2]For example, one could pose the following 'what' question: what happened when intervention A occurred? Under the definition offered by both Gerring (2012) and Holmes et al. (2024), this would appear to be a descriptive question, yet it is really a canonically causal one.

of the world. Yet the underlying theory of interest, from which these questions derive, is causal.[3] There is a sharp distinction here between the *theory* that a researcher posits or hopes to evaluate, and the specific *research question* that a researcher may hope to assess, either conceptually or empirically. In many research settings this distinction is not properly specified, and so there emerges a tension between the language researchers use to describe the goal of their research and the empirical claims they actually make.

Researchers should focus on this distinction early in the research process, articulating both the theory of interest and the specific research question(s) derived from that theory. While these two may be the same in some cases, they need not be. Empirical analyses should be devised to evaluate a well specified research question, but they need not always directly evaluate the theory in the first instance. Indeed, sometimes this is simply not possible – perhaps there is no satisfying and compelling research design that isolates variation in A. Good description requires the definition of (explicitly descriptive) research questions in clear relation and correspondence to social scientific theory, and executing empirical analyses that are correctly specified so as to target these (important and useful) intermediate research questions.

### 2.1. Theoretically Informed Description

For description to be valuable, then, it must be grounded in social scientific theory. Description demands a clear articulation of precisely what researchers are attempting to describe, and the status of their questions and claims of interest. That is, researchers engaged in description must have a well-formed understanding of the phenomenon of interest – the thing – and a well-formed understanding of the relational context of that phenomenon – the question. This criterion applies whether we are engaged in qualitative or quantitative description (or put differently, 'historical' or 'statistical' description, see (Kreuzer, 2019)). All of this presupposes a well-formed understanding of the broader theory of interest.

[3] To be clear, normative political theory motivates both descriptive and causal analyses and informs what we consider relevant and important phenomena.

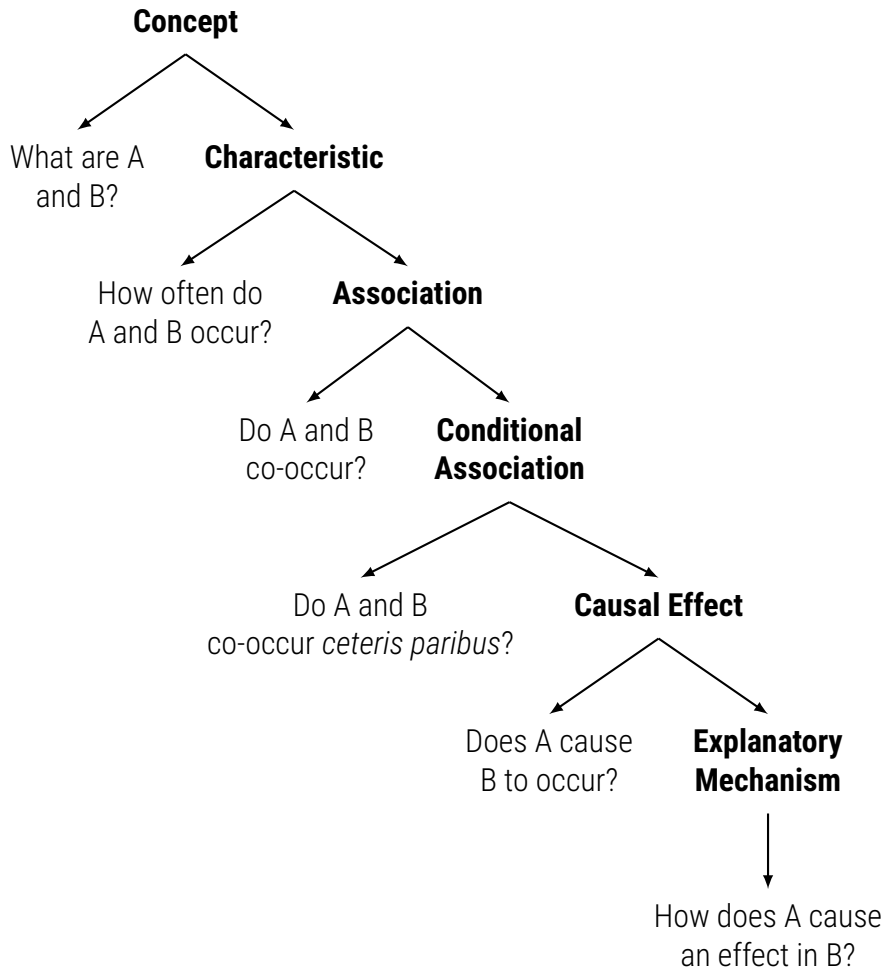


Figure 1: An illustrative ladder of possible research questions, derived from a simple theory: A⟶B.

Let us dive more deeply into the toy example outlined earlier: A ⟶ B. Here we have two concepts of interest that are theoretically derived, thing A and thing B. We offer in Figure 1 an illustrative ladder of some social scientific questions one might ask about those two things. Note two points about this exercise. First, while the rungs (written in bold) are intended to be close to comprehensive, the figure is not exhaustive in that there are many more specific questions that one could formulate for each specific rung. Second, the use of the word ladder here does not presuppose that any question is *more important* than any other, but does presuppose that we cannot answer the 'deeper' questions (those lower on the ladder) without a means to answer the 'shallow' questions (those higher on the ladder).

At the top of the ladder are questions about the conceptual status of the phenomena A and B. For example, researchers might wish to investigate what precisely the phenomena of interest are, be that at a conceptual level (ontological questions about what conceptually constitutes the thing) or at an observational level (an empirical question about what regularly constitutes the thing).[4]

One rung down are questions related to characteristics of the phenomena of interest. For example, we might ask how often each of the things of interest, A and B, occur. A further rung down are associational questions, for example whether A and B can be said to co-occur. Below those are questions of conditional association, for example, conditional on some other factors, does the co-occurrence of A and B still hold. On rung four we find causal relationships, questions such as whether A increases the likelihood of B, or perhaps whether A is necessary and sufficient for B to occur. Finally, at the bottom of the ladder are explanatory mechanisms, questions about how any effect of A on B happens.

As noted above, there is a hierarchy here in the sense that we cannot answer a lower question unless we are in principle able to answer all the higher questions (whether we actually do answer the higher questions is a separate issue, but we should theoretically be able to do so). We can say little empirically about the mechanism by which A affects B unless we are able to empirically show that A causes B. We cannot say whether A and B co-occur unless we are able to say how often A and B occur. Critically, we are not able to answer even this seemingly simple question about frequencies or other characteristics of A and B, unless we have a clear understanding of what A and B are meant to represent in the conceptual space.

Which of these types of questions are typically conceived of as 'descriptive?' Rungs 1, 2, and 3 are perhaps the most clearly descriptive. Rung 4 is an ostensibly ambiguous case where researchers should exercise a great deal of caution. This case reflects the common disposition of social scientists in the period prior to the "credibility revolution," where researchers established *ceteris paribus*comparisons

[4]For example, Chandra (2006) conceptualizes ethnic identity as identities determined by descent-based attributes. Description as an empirical exercise could help to identify the constituent parts of 'ethnic identity' and test whether these constituent parts exist, without playing any role in creating the concept itself.

and used these conditional associations as evidence of causal relationships.

Many questions from rung 4 have been re-posed so that they fall under rung 5 (causal effects). Yet in the absence of such re-wordings, conditional co-occurrences are themselves descriptive questions because they do not require manipulability. For example, in discussions of the gender pay gap, scholars do not often argue that gender *causes* wages. Instead the claim is that there is a systematic co-occurrence of particular genders and (lower/higher) wages even when conditioned on a range of explanatory factors (job type, sector, educational attainment, etc.). That is, there remains a residual difference in wages between men and women that is not accounted for by other factors included in the analysis. This does not mean, and is not intended to imply, that 'gender affects wages,' but instead describes a systematic conditional difference in wages based on gender. These types of questions often seem to fit in a liminal space between descriptive inference and and causal inference. Yet it is the necessity of manipulability that sets the two apart, and sharply so.

This discussion emphasizes how important it is that researchers are explicit and clear about the types of questions they are trying to answer with any given set of analyses. Hernán (2018) notes that there is a tendency for researchers to speak euphemistically about the questions they are interested in answering. Ideally researchers might hope to answer a causal research question (rung 5), but they use the language of description because they do not believe they have an appropriate research design or identification strategy. But the quality of the evidence available cannot change the type of question researchers are interested in asking – that question itself is stated in abstraction from the evidence. If the available evidence can only facilitate answers to descriptive questions, that is an argument for specifying good descriptive questions (e.g. from rung 1, 2, 3, or potentially 4). Good description, then, is not simply second-rate causal inference. Indeed, this euphemistic or "ambiguous" language appears to make interpretation more challenging for readers (Haber et al., 2022).

## 2.2. The Goals of Good Description

Given this analytical ladder, and the central role of theory, our view is that good description is an empirical exercise designed to describe phenomena of interest with one or more of the following goals:

1. Reflect or invoke theory in a conscious and clear way: The things we are measuring or characterizing are theoretically relevant objects.

2. Build theory by describing phenomena we want to explain: Patterns of occurrence such as characteristics, associations, or conditional associations can help to link potential causes to outcomes, and mechanisms to causes, without ever explicitly testing a theory.

3. Evaluate theory by assessing whether the implications of theory hold: Some causal theories have non-causal observable implications, and description can update our beliefs by studying these hypothesized relationships.

The first of these represents measurement, the only role that social scientists very clearly agree upon for description. There is a rich tradition of measuring – and thus describing – the state of the world. Examples include the measurement of democracy, representation, and political cleavages (Przeworski and Limongi, 1997; Tausanovitch and Warshaw, 2014; Huijsmans and Rodden, 2024), but this barely scratches the surface. Importantly, such measurement efforts have clearly always been theoretically guided; graduate students are taught early in their careers about the importance of 'construct validity,' and the ways in which observations are not independent of the theories they are used to test (Kuhn, 1970; Lakatos, 1970; Hall, 2003). Good measures should accurately and validly capture theoretical concepts of interest, whether these arise from positive or normative theory. Many of these concepts are what might be termed 'primitives,' or foundational building blocks of the social world, such as regime type, a nation's wealth, or the level of inequality. Yet even these 'primitives' are important precisely because they are (perhaps trivially) theoretically relevant objects – for example, we care about measuring the prevalence of democracy because at a fundamental level we believe democracy itself is

theoretically important.

Reflecting theory is a necessary (if not sufficient) condition for valuable measurement. Measures that have poor construct validity can offer little to researchers engaged in either descriptive or causal arguments. Indeed, for measurement to be valuable scholars typically expect measures to capture concepts of interest that underpin some theoretical argument. For example, Huijsmans and Rodden (2024) measure urban-rural partisan divides across Western democracies. Doing so requires a theoretical understanding of the concept of partisan voting and partisan loyalties in each context, and why we would expect them to differ between cities and rural areas, a point invoked by our discussion of comparability in the previous section.

Yet description is much more than measurement. It can help researchers build theoretical arguments by describing the social world more accurately or comprehensively, thus creating opportunities to perceive patterns or puzzles they would not otherwise have known to explore or explain. Anomalies and outliers only exist in light of existing theoretical expectations, and accounting for them can help to strengthen theory (or undermine the original theoretical framework). Here, some scholars have called for a recursive (and transparent) moving back and forth between empirics and theoretical or causal models, refining models in light of empirical observations and in effect intertwining observation and theoretical generalization (Geddes, 2003; Tavory and Timmermans, 2014; Yom, 2015; Humphreys and Jacobs, 2023).

Likewise, description can help us evaluate well-developed theories and explanations. Causal theories can have non-causal implications, and description is especially useful in *disconfirming* such implications. One very simple way to evaluate causal theories through good description is to use observational correlations to establish whether other relationships consistent with the causal ones exist. If a theory predicts a certain state of the world (for example, a correlation between two variables), a descriptive analysis can assess whether that state of the world exists. If it does not, or if that state occurs less often than expected, that is an indication the theory is either incomplete or in error. It may be that the prediction does not follow from the theory, or the relationship is conditional on some other fac-

tors that theory did not specify. Here, the disconfirmatory value of not finding a correlation outweighs the confirmatory value of finding the expected correlation: description can help to undermine potential explanations and thus both to evaluate theories and help to build new ones.[5]

While clearly separate from causal inference, there remains a strong connection between the two. A typical admonition in the practice of causal inference (and empirical research more broadly) is for researchers to evaluate and challenge the assumptions underpinning their research design. These assumptions usually relate to the researcher's beliefs about the assignment mechanism – how it is that levels of the causal variable are assigned to units in the given design (Imbens and Rubin, 2015). Yet these assumptions are themselves part of a broader theory, often formally specified through the use of directed acyclic graphs (DAGs), which accounts the researcher's beliefs about the variables that determine, jointly or independently, both A and B (in our toy example). We propose that descriptive evidence can challenge or question that theory, be it by questioning specifically the assignment mechanism, or questioning some other part of the hypothesized DAG.

Descriptive inference can also be useful in assessing the plausibility of assumptions underlying empirical strategies. For example, in spatial regression discontinuity design (RDD), the researcher studies the effect of a policy which only applies within a specific spatial boundary. RDD relies on continuity in potential outcomes around the boundary, which is generally implausible if other things (e.g. other policies) also change at the boundary, as is often the case, or if the boundary is itself a function of the treatment. Yet as Kocher and Monteiro (2016) and Verghese (2024) demonstrate, such assumptions may be rendered untenable through careful historical description of the assignments to treatment. State borders in Africa, for example, are not straight lines drawn randomly by colonizing powers, but reflect a far more complicated process involving negotiations with indigenous rulers and and a consideration of strategic

[5]Related to this point, Blackwell et al. (2024) demonstrate that even well-intentioned analyses in this spirit can mislead when the underlying theory is incorrectly specified. This emphasizes, again, the importance of explicitly stating the theory from which research questions are derived – it is hard to evaluate whether a theory is incorrectly specified unless researchers are clear about what that theory is.

geographic focal points (Paine et al., 2025).

## 3. The Characteristics of Good Description

So far we have argued that the social scientific purpose of description is much the same as the purpose of causal inference: to improve our understanding of the social world. To that end, description should be theoretically informed, and focus on describing theoretically relevant objects. The way researchers approach these objects should be via precisely articulated research questions, and analyses that are well-calibrated to these precise questions. If these are the goals or purpose of description in social science, how can we evaluate whether any given instance of description achieves these goals? We propose that good description is clear, comparable, and complete.

Clear description is tightly linked to theories and explanations of interest, and where the analytical or empirical approach is well suited to the research questions derived from these theories. There should be an apparent correspondence between the theory of interest, the descriptive research question stated, and the descriptive analyses conducted. It should be apparent why the descriptive analysis is valuable in terms of social scientific and/ or normative political theory (why do we want to describe this phenomenon?), precisely what the analysis is intended to evaluate (what is the precise descriptive question?), and how the analysis was conducted (what is the method of description?). In short, descriptive research questions should be posed with clear reference to conceptual objects or theories that connect those objects together, and appropriate methods should be chosen given that research question. Note that many descriptive exercises will serve to build theory, and so theory may be invoked in an inductive or abductive, rather than deductive, way.

Comparable description pays attention to how context affects measurement and meaning, so that comparisons are made on a consistent basis. Researchers should consider explicitly the underlying conditions that are baked into any specific descriptive exercise. It is not obvious that a descriptive exercise conducted in one context (population, time, or place) is comparable with a similar descriptive exercise conducted in a different context (population, time, or place). This is especially important since

many descriptive exercises will be in aid of comparative analyses, whether over different populations, different times, or different places.

For researchers, this implies at least three considerations regarding measurement, meaning, and analysis. First, measurement needs to take into account local data generating conditions. For example, the reliability of data from autocratic and democratic governments may vary. Similarly, measures of wealth from the 12th and 21st century will likely reflect underlying differences in state capacity to measure wealth. Second, measures may have different and incompatible meanings. Take for example, measures of violence against journalists (Coppedge et al., 2019; Little and Meng, 2023). In some contexts a high score on such a measure may suggest democratic erosion and a low score democratic strength. In other contexts, a high score may indicate the weakness of an authoritarian regime and its inability to successfully intimidate journalists. Low scores on violence against journalists can also mean control by other means, such as a successful pattern of state media ownership and manipulation (Carey and Gohdes, 2021). For example, no journalists have been killed in Hungary in recent years (CPJ, 2025), but that may be because over 80% of the media is owned, directly or indirectly, by the government. In short, comparable description is that in which the researcher is careful to engage with whether data from different contexts map onto similar meanings. Third, the same analytical or empirical strategy may not mean the same thing in different contexts. A regression that controls for Z and Q may not mean the same thing in different places if the underlying data generating process that connects (or does not connect) Z and Q to dependent and independent variables of interest is different. Finally, researchers should also be sensitive to different underlying distributions in the data. The central tendency, such as a mean or median, may be an adequate representation of a wealth distribution in relatively equal places such as Norway or Denmark, but in highly unequal places like Brazil or South Africa such quantities could be highly misleading.

Comparability thus means that measurements are consistent enough from place to place and over time, *and* that those measures actually reflect the context that produced the data. That in turn implies that there may be a trade-off between generalizability and context-sensitive, fine grained measurement.

Measurement is not useful if it is idiosyncratic and unique to one place or time, because it does not allow comparison. But measurement also needs to take stock of what is being measured, and the context that produced these estimates. Otherwise, while we may have consistent measurements, they may refer to very different phenomena. This is a hard problem for description, and descriptive work that is able to balance this tension fairly and transparently is highly valuable.

Complete description either answers as many descriptive questions as possible given a theory of interest, or has assessed an appropriate sample given a (super-)population of interest. Good description should endeavor to provide as comprehensive a picture of the phenomena of interest as possible. This could be understood in many ways: we focus on three here. First, completeness may be with reference to the set of descriptive research questions derived from the theory of interest. That is, given a theory of interest, has the descriptive exercise comprehensively studied as many (useful and important) descriptive questions as possible, or do (useful and important) questions remain unanswered? Second, completeness may be with reference to the inferential population of interest. That is, has the researcher specified a theoretically appropriate population, and has the descriptive exercise assessed an appropriate sample from which meaningful population-level inferences can be drawn? Third, completeness may implicitly trade off depth and breadth. A researcher may study deeply a single case at few points in time. Or they may study broadly many cases at many points in time. Either of these approaches tends toward completeness, but a failure to do either would suggest incompleteness.

### 3.1. From Question to Analysis

For descriptive analyses to be clear, comparable, and complete, the analytical approach adopted by a researcher must be carefully mapped to the specific research question. This can be challenging, particularly so for questions on rungs 3, 4, and 5 of Figure 1. Consider for example some of the most academically successful social scientific descriptive work in recent years, that by Chetty et al. (2014) on economic mobility in the United States. Here, one of the key findings is that "a 10 percentile increase in parent income is associated with a 3.4 percentile increase in a child's income." While Chetty et al.

(2014) are clearly characterizing an association (rung 3 or 4 on our ladder, depending on whether the association is conditional), the scientific value of the claim hinges on the specific question and underlying theory being tested. For the authors' descriptive inference to be scientifically valuable given their theory of interest, they attempt to rule out competing explanations by making a *ceteris paribus* type claim (Spirling and Stewart, 2022).

This tension is routinely ingrained in the analytical choices that scholars make, though it is often not an explicit consideration. One common descriptive methodology in quantitative social science, for example, is the use of multivariate regression to establish conditional associations between variables. In this exercise, the researcher attempts to characterize 'what goes with what' (an associational claim, on rung 3 of our ladder), yet invariably the exercise is more akin to 'what goes with what, conditional on other variables' (rung 4). Whether that exercise makes sense will depend – as in the case of Chetty et al. (2014) – on what theory or theoretical concepts the authors are attempting to measure, build, or evaluate.

How should we interpret this exercise? If the researcher firmly believes they are not engaged in counterfactual reasoning – they purely want to 'describe the data' – then the use of multivariate regression is itself a peculiar choice. Regression by design involves projections into counterfactual parts of the data, that is, extrapolation beyond the support of the data. If one hopes to 'control' for variables in a descriptive analysis, it is because one has an underlying theoretical model about how those variables are causally connected. Indeed, Stanley Lieberson famously cautioned against the misuse of controls as anything other than descriptive devices: "The control variable is a perfectly appropriate descriptive device; the problem occurs when control variables are viewed as an analytical procedure. It is one thing to ask, what is the occupational attainment of blacks and whites of a given educational level? It is another to ask, taking into account or controlling for the influence of education on attainment, what is the influence of race on attainment?" (Lieberson, 1987, 213-4).

Such conditioning may be justified on the basis of a causal inference research design such as selection on observables, where an underlying theoretical model asserts that a treatment is 'as-if randomly'

assigned, conditional on pre-treatment covariates. Alternatively, it may be justified without an explicit causal model at all, as in the case of Simpson's paradox, where the sign on a global relationship may in fact reverse when studying the relationship for sub-sets of the data (Simpson, 1951). This type of setting is often used to justify conditioning on variables: Without adjusting for the correct covariate(s), we are misled and draw the 'wrong' conclusion from the data. It is only once we adjust that the 'right' relationship emerges.

Yet this only makes sense as an exercise when it reflects theory. If theory predicts an unadjusted correlation between two variables, conditioning is not necessary. Likewise, if the specified research question of interest is simply 'does X correlate with Y,' conditioning is not necessary. However, if theory suggests a conditional relationship, or a conditional relationship is stated in the research question, then conditioning is necessary. 'Rightness' and 'wrongness' can only be understood in terms of an explicit research question connected to theory – the global sign in a case of Simpson's paradox is not 'wrong' unless one is targeting the conditional sign. The goal of such an exercise is rarely explicitly explained by researchers. It may be an attempt to study conditional associations (rung 4). Yet it may be, and more often is, an exercise in counterfactual reasoning (rung 5) where we control so as to "rule out alternative explanations" (Spirling and Stewart, 2022, 17).

If one wants to purely 'describe the data' then one could do this without regression. Yet invariably the instinct to control dominates. One could go so far as to argue that causal inference studies focused on a single case can be thought of as 'description' of that specific case. That is, a dominant trend in social science has been the use of causal inference tools to analyze specific case studies, not to engage in comparison across multiple cases, whether in any traditional 'comparative method' sense or through meta-analyses. One could think of these case studies, which use the technology of modern causal inference, as descriptive studies themselves: They provide precise estimates or assessments of one data point for future researchers consider when conducting comparative or meta-scientific analyses.

By thinking carefully about where on the 'ladder' in Figure 1 their questions sit, researchers should be able to establish the appropriate methods for studying their questions.

## 4. Good Description in Practice

So far we have articulated three scientific goals that good descriptive questions can help researchers to pursue, and three characteristics of good descriptive analyses. We now explore two ways good description can be applied. First, good description can specify the relevant theoretical and empirical contexts for valuable comparisons and credible identification. By providing baselines and denominators for comparison, description can help to evaluate theory and provide plausible scope conditions. Second, description helps us to understand the unfolding of phenomena over time and place. Description can thus illuminate the pathways and mechanisms underlying causal relationships.

### 4.1. Diversity of Contexts

Good description provides theoretically-relevant context and situates measurements and findings. This sensitivity to context includes a deeper knowledge of culture, norms, reference points, and details that improve the quality of data collection and interpretation (Cirone et al., 2021), and attention to the ways in which phenomena of interest may vary across time and space (Hall, 2003, 383), (Tavory and Timmermans, 2014, 72). This is an important way in which good description both relies on, and makes for, good measurement. Good description also specifies the theoretically-relevant comparative context in which we are analyzing a given outcome: other situations in which an outcome occurs, the distribution to which the outcomes belong, and the chronological timelines or empirical baselines against which we evaluate the outcome. Here, description helps to build theory by specifying both the scope conditions in which the theory may apply, and the variation in the phenomena that the theory seeks to explain.

Different outcomes are often observed in similar contexts that may call into question causal accounts. Consider some examples outside of political science. There is increasing concern about teenagers and their use of cell phones as a source of depression and anxiety, based on data from the United States (Haidt, 2024). One important piece of information for evaluating the plausibility and completeness of this explanation would be whether teenagers with similar rates of phone use in other

wealthy democracies show similar levels of anxiety and depression. If they do not, any narrative that attributes mental anguish to adolescent phone use alone becomes more suspect, or at least more complicated. Indeed, despite similarly high rates of cell phone use, rates of anxiety and depression among European adolescents appear to be lower than in the United States (Sacco et al., 2024). This finding suggests that a monocausal explanation is incomplete. Similarly, studies of automobile deaths show that while purchases of ever-larger suburban utility vehicles (SUVs) skyrocketed after 2000 in the United States, Canada, Australia, and New Zealand, automotive deaths decreased everywhere except the US (Burn-Murdoch, 2024). Increasing size or prevalence of SUVs, then, is not the (single, independent, unconditional) cause of automobile deaths. If we observe similar outcomes across different contexts, that, too, can give us valuable clues about potential relationships across the data. For example, were we to observe that parents everywhere favor their children, then it is unlikely that a higher or lower levels of 'individualism' alone can account for their behavior (Henrich, 2020). Description can thus not only provide a context for comparison but weaken our confidence in some potential causes or demand more nuanced theory and subsequent testing. A careful examination of context can suggest an incompleteness, either in terms of unspecified confounders that mask (amplify) the true relationship outside (inside) of the USA, or in terms of treatment effect heterogeneity where the effect only emerges when other background factors also exist, or where other countervailing factors are absent. [6]

Similarly, most political phenomena are the result of multiple and interacting causes, beyond single and independent causes. An explanatory variable A may be one of many causes of changes in Y, but it may also be a *conditional* cause, operating in the presence of or at some level of other causal factors. Good description can help us to specify these potential conditional relationships, for example through the use of moderators or interaction terms (Clark et al., 2006; Hainmueller et al., 2019). If Z is the conditioning or modifying variable, the strength of any effect of A on B may vary depending on the

[6] To be clear, we are *not* suggesting that a lack of observable correlation necessarily implies a lack of causation. Instead, it may suggest that the researcher's theory of how either the causal variable is assigned or how the dependent variable is generated is under-specified or incomplete in important ways.

value of Z. Questions about heterogeneous treatment effects are themselves descriptive – Z is not required to be manipulable to answer the question. Good description both measures A, B, and Z, and helps to specify the ways in which any marginal effect of A is conditional on Z, only some of which may be consistent with the original theory (Berry et al., 2012, 654).

Another way of describing context is to assess the plausible distributions from which the outcome may be drawn. We want to know how specific or general a given concept is, and how stable or variable are its attributes across contexts (Kreuzer, 2023, chapter 5). Empirically, relevant questions include whether the observed cases or relationships are drawn from around the mean of a normal distribution of outcomes, or whether it is an outlier, several standard deviations away from the mean. In other words, how unusual is the relationship or outcomes that we are observing? Such description can be useful theoretically and empirically. For example, a bimodal bunching of outcomes not only calls for different tests, but it can also indicate substantive outcomes of interest such as fragmentation, polarization, social distancing, and so on. More prosaically, if an outcome is itself exceedingly rare (as determined by an appropriate descriptive analysis), this may raise questions both about the scientific importance or completeness of the theory, and about the plausibility of any treatment effect that may simply reflect noise in the measurement of the outcome.

Distributions themselves are highly informative. For example, outcomes can follow a 1/x power-law distribution, such as Zipf's law or Pareto's law. These distributions have fatter 'tails' than normal distributions, the workhorse of social science, and that has considerable implications for both theory building and testing. First, we are more likely to find far more rare events (such as very large cities, common words, or very rich people) than we would expect with a normal distribution. Second, it suggests that such outcomes may be self-similar: small conflicts are simply a fractal version of large wars, and vice versa. Third, while normal distributions are the result of the addition of independent variables, power-law distributions result from the variables or events being multiplied or correlated (West, 2017; Miller and Page, 2009). Thus, positive feedback may result in a power-law distribution, not a normal one. In building theory, this implies we need to specify the kind of interdependence that leads to these

patterns. In such cases assumptions of independence may be violated, and empirical models that rely on normal distributions are misleading.

Finally, denominators, historical baselines, and base rates can provide critical context for empirical researchers. Trivially, if we ask whether a given phenomenon is on the rise or the decline, then we need to know the historical baseline. If we measure a phenomenon in absolute terms we also need to know the number of units and any changes in that number over time. For example, an increasing number of democracies in the world may indicate the triumph of liberal democracy – or simply reflect the change in the number of countries and new states (the denominator). Recent debates about democratic backsliding revolve partly around the indicators of democracy – but also question whether the historical trends indicate a democratic crisis or continued resilience (Little and Meng, 2023; Treisman, 2023). Unless we have a clear understanding of the values of 'democracy' relative to other regimes, and to democracies in the past, it is difficult to conclude whether democracies are resilient or fragile. Similarly, (Chang and Wang, 2024) examine the broadcasting of state power through the presence of specific state institutions such as police units, public prosecutor offices, courts, and tax agencies. They find that there are fewer agencies in densely populated regions, the result of decreasing marginal costs of governance. This is an important reminder to specify the population as a denominator before making claims about the relative efficiency of state authority.

### 4.2. Time, Place, and Unfolding Phenomena

The analysis of how phenomena unfold over time and place is another critical contribution of good description. There is already a wealth of important and related work on process tracing (Bennett and Checkel, 2015; Checkel, 2008; Kittel and Kuehn, 2013; Collier, 2011) and event history modeling (Box-Steffensmeier and Jones, 1997; Boehmke, 2005; Blossfeld et al., 2014). Prominent contributions in this vein are explicitly concerned with establishing causation (Collier et al., 2010; Collier, 2011; Mahoney, 2012). Yet more modest empirical efforts may also be valuable: specifically, describing how events occur and historical legacies unfold without necessarily establishing causal links.

It is often useful to describe how a given event or phenomenon spreads (or retreats) across time, territory, social networks, or institutions. Other scholars have called for greater attention to geographic and historical trends, transformations over time, and continuities and discontinuities as part of causal claims (Kreuzer, 2023), but such analysis fits just as comfortably in a descriptive framework.

First, good descriptive measurement is needed to identify the theoretically relevant units of time and territory, institutions, or networks, subdividing broader phenomena into their constituent parts (Cirone and Pepinsky, 2022, 255). This may mean days, years, or decades, units of man-made time (such as religiously mandated periods of mourning or holidays, electoral cycles, or census periods), or specific periods when given events unfolded (armed conflict and wars, famines, or regime collapses). This measurement needs to be relative to the relevant population. By tracing the presence of a given variable or event across both time and space, we may learn how sustained it was, when and how it peters out and attenuates, and we may be able to eliminate some potential underlying causes or mechanisms. We can then trace whether and how a given variable or mechanism is transmitted over time and place. Some relevant question here include: who adopted the practice and when? When do contiguous units start to show the same characteristics? At what point a given environment is saturated? To answer these questions we need to know where and when we see a given phenomenon, and be able to evaluate evidence of the mechanism of 'contagion' or spread.

Second, good description can measure maintenance, growth, or contraction of a phenomena over time and place. In historical legacy arguments, including long-run persistence studies, researchers establish a causal effect of a historical phenomenon on a contemporary phenomenon of interest, whether using the tools of causal inference or mechanistic approaches that trace the interaction of agents and the reproduction of a given legacy (Gailmard, 2024; Simpser et al., 2018). Examples of these historical arguments include studies linking patterns of colonial settlement to contemporary economic and political outcomes (Acemoglu et al., 2001), autocratic rule to democratic competitive patterns (Grzymala-Busse, 2002; Riedl, 2014; Wittenberg, 2006), the Nazi regime to subsequent economic development (Charnysh, 2019; Charnysh and Finkel, 2017; Homola et al., 2020), or historical slavery to modern social

trust and attitudes (Nunn and Wantchekon, 2011; Acharya et al., 2016). Accounts invoking historical legacies often argue that phenomena can persist long after the conditions that gave rise to them have themselves disappeared (De Kadt, 2017; Simpser et al., 2018).

Critics point out that such arguments often do do not specify *how* the long-term outcomes arise, leaving the mechanism of transmission underspecified. Good description can provide the "abundance of intermediate outcome data and multiple qualitative studies that meet stringent case selection criteria" (Cirone and Pepinsky, 2022, 254) needed for mechanistic accounts. We can show that a historical cause was sustained over time, with similar levels of the causal factor over time and place, or that it gave rise to another causal factor, which we can then measure over time. We come closer to demonstrating how a given cause was sustained and reproduced over time.

Historical legacy arguments are also less than specific about the half-lives of the invoked legacies: when such legacies attenuate, where do they do so, and whether some are more durable than others. They are also circumspect about the *distribution* of the impact of the legacies: where and when these legacies played the greatest role, or show the greatest association. Similarly, we would want to know not only how the territorial diffusion of state authority or imperial rule diffuses from the center – but also whether and how it *contracts*. Here, good description can measure whether the legacy is sustained over time and place, through multiple and repeated measurements of the observable levels of the antecedents of the historical legacy.

Good description can help to discover (or challenge) certain historical explanations. For example, Voigtländer and Voth (2012) tie medieval pogroms to 20th century antisemitism. They suggest that this may be the result of socialization within families and the imitation of parental norms. Similarly, Homola et al. (2020) argue that proximity to Nazi concentration camps has led to an increase in voting for far-right parties, the result of exposure to Nazi institutions and the cognitive dissonance this experience produced. Much as in Voigtländer and Voth (2012), family socialization and parental norms are the mechanism that transmit beliefs across time. Yet this account has also been questioned on the basis of descriptive knowledge. Pepinsky et al. (2024b) argue that Germany's Länder vary greatly

in terms of school curricula and civic education, which likely affect modern day social beliefs and behaviors, and may act as a confounder. They thus propose that Länder fixed effects should be included in the conditioning set (see Homola et al. (2024) and Pepinsky et al. (2024a) for a continuation of this debate). More broadly, as Voigtländer and Voth (2012) point out, we need to specify the conditions under which historical factors are preserved and transmitted, and when they attenuate in response to shocks, political evolution, or cultural exposure via trade or education.

Third, carefully attending to the territorial and temporal unfolding of a phenomenon can help in isolating the causal mechanisms that might be at play, and more specifically, which elements are unlikely given the observed patterns. Elements of temporality, such as duration, tempo, timing, and sequencing help to differentiate causal mechanisms: "rapid institution building is unlikely to involve deliberation, consensus building among wide constituencies, or multiple veto players" (Grzymala-Busse, 2011, 1271). Instead, it is more likely to involve existing skills, networks, and templates. Such rapid change may also preclude other phenomena from occurring: for example, quick repression may prevent protest from evolving into a revolution.

Measuring growth can also identify exponential patterns, of the kind observed in large-scale behavioral tipping seen in the shift of social and political norms (Mackie, 1996; Nyborg et al., 2016). For example, attitudes towards same-sex marriage suddenly shifted radically over the course of the early 2000s, tipping into majority approval by 2011 (Gallup, 2025). One critical property of such geometric growth is that the initial increase may be very hard to observe at the early stages, consisting as it does of the multiplication of very small quantities, so that by the time we observe the huge and sudden expansion, the mechanism had been unfolding for a considerable period of time. Such mechanisms are especially critical in epidemiology (and why containing the initial spread of highly infectious diseases is both critical and difficult), finance (bank runs tend to take this form), revolutions (and why we can only observe that they "are not made: they come", in the words of Wendell Philips) and the spread of ideas.

Tracing the sequence of events is particularly important in path-dependent processes, where the outcomes depend on the ordering of the events leading up to either an outcome in one period, or a

long-run equilibrium (Page et al., 2006). Distinguishing between path-dependent processes and simple persistence necessitates documenting the sequence of events and the accumulations of externalities over time. The description-based counterfactual reasoning we outlined earlier in Section 2, and exercises such as placebo tests, may also show how these sequences and accumulations precluded other paths.

## 5. Conclusion

Social science has seen several broad methodological developments in the last few decades, ranging from statistical, computational, and qualitative advancements that have deepened and broadened the analytical toolkit, to formal modeling that has taught us to specify the micro-foundations and strategic interactions in our theories, to the identification revolution which emphasizes the importance of underlying research designs and assumptions for causal inference. One of the features of all these advances is the development of clear evaluative criteria. We try to build that same framework for description, which connects and underpins all these enterprises.

Description is the scientific art of connecting well specified theory, which is invariably causal, to appropriate computational, statistical, and qualitative analyses to offer insights on important questions. Description is therefore a theoretically-informed mode of social scientific inquiry. When researchers attempt to answer the questions they have derived from theory, they do so by conducting empirical analyses. Description is both fundamental first step for various forms of inference, and a valuable social scientific exercise in its own right. As researchers pursue descriptive work, we encourage them to argue clearly and explicitly about where the social scientific value of their work lies.

'Good description' engages theory, by invoking existing theories, motivating new theoretical models and their refinement, and in some cases even evaluating theory. It does so by specifying phenomena of interest, pointing to departures from expected distributions of data, or demonstrating patterns that ought (not) to exist given our theory. It serves to specify the scope conditions and missing aspects of theory, and even assess some observable implications of causal theories. Good description can help

us to specify the distribution, context, and unfolding of a variety of important phenomena. In applied settings, such description may include specifying relevant contexts, and describing how phenomena unfold across time and space. Here, mapping and situating the phenomenon of interest through good description may be as important as identifying any causal effect.

We propose that both the producers and readers of such work can evaluate the quality of descriptive research by examining how clear, comparable, and complete it is. Good description clearly articulates how it engages theory, it specifies the context that is invoked (through comparisons, distributions, or through tracing change over time and place), and encompasses plausible measurements and specifications. Such 'good description' goes hand in hand with other models of inference – and when done well, represents a valuable and fundamental form of modern social scientific inquiry.

## References


Acemoglu, D., S. Johnson, and J. A. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *American economic review 91*(5), 1369–1401.

Acharya, A., M. Blackwell, and M. Sen (2016). The political legacy of american slavery. *The Journal of Politics 78*(3), 621–641.

Beissinger, M. R. (2022). *The revolutionary city: Urbanization and the global transformation of rebellion*. Princeton University Press.

Bennett, A. and J. T. Checkel (2015). *Process tracing*. Cambridge University Press.

Benoit, K., K. Munger, and A. Spirling (2019). Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science 63*(2), 491–508.

Berry, W. D., M. Golder, and D. Milton (2012). Improving tests of theories positing interaction. *The Journal of Politics 74*(3), 653–671.

Blackwell, M., R. Ma, and A. Opacic (2024). Assumption smuggling in intermediate outcome tests of causal mechanisms. *arXiv preprint arXiv:2407.07072*.

Blaydes, L. (2018). *State of Repression: Iraq under Saddam Hussein*. Princeton University Press.

Blossfeld, H.-P., A. Hamerle, and K. U. Mayer (2014). *Event history analysis: Statistical theory and application in the social sciences*. Psychology Press.

Boehmke, F. J. (2005). Event history modeling: A guide for social scientists. *Perspectives on Politics 3*(2), 366–368.

Bonica, A. (2014). Mapping the ideological marketplace. *American Journal of Political Science 58*(2), 367–386.

Box-Steffensmeier, J. M. and B. S. Jones (1997). Time is of the essence: Event history models in political science. *American Journal of Political Science*, 1414–1461.

Burn-Murdoch, J. (2024). Why are american roads so dangerous?

Carey, S. C. and A. R. Gohdes (2021). Understanding journalist killings. *The Journal of Politics 83*(4), 1216–1228.

Chandra, K. (2006). What is ethnic identity and does it matter? *Annu. Rev. Polit. Sci. 9*(1), 397–424.

Chang, C. and Y. Wang (2024). The reach of the state. *Comparative Political Studies 57*(8), 1243–1275.

Charnysh, V. (2019). Diversity, institutions, and economic outcomes: Post-wwii displacement in poland. *American Political Science Review 113*(2), 423–441.

Charnysh, V. and E. Finkel (2017). The death camp eldorado: political and economic effects of mass violence. *American political science review 111*(4), 801–818.

Checkel, J. T. (2008). Process tracing. In *Qualitative methods in international relations: A pluralist guide*, pp. 114–127. Springer.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics 129*(4), 1553–1623.

Cirone, A. and T. B. Pepinsky (2022). Historical persistence. *Annual Review of Political Science 25*(1), 241–259.

Cirone, A., A. Spirling, et al. (2021). Turning history into data: data collection, measurement, and inference in hpe. *Journal of Historical Political Economy 1*(1), 127–154.

Clark, W. R., M. J. Gilligan, and M. Golder (2006). A simple multivariate test for asymmetric hypotheses. *Political Analysis 14*(3), 311–331.

Collier, D. (2011). Understanding process tracing. *PS: political science & politics 44*(4), 823–830.

Collier, D., H. E. Brady, and J. Seawright (2010). Outdated views of qualitative methods: time to move on. *Political Analysis 18*(4), 506–513.

Coppedge, M., J. Gerring, D. Altman, M. Bernhard, S. Fish, A. Hicken, M. Kroenig, S. I. Lindberg, K. McMann, P. Paxton, et al. (2011). Conceptualizing and measuring democracy: A new approach. *Perspectives on politics 9*(2), 247–267.

Coppedge, M., J. Gerring, C. H. Knutsen, J. Krusell, J. Medzihorsky, J. Pernes, S.-E. Skaaning, N. Stepanova, J. Teorell, E. Tzelgov, et al. (2019). The methodology of "varieties of democracy"(v-dem). *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 143*(1), 107–133.

CPJ, C. t. P. J. (2025). Hungary. Technical report, https://cpj.org/europe/hungary/.

Daly, S. Z. (2022). *Violent victors: Why bloodstained parties win postwar elections*. Princeton University Press.

Davenport, L. D. (2016). Beyond black and white: Biracial attitudes in contemporary us politics. *American Political Science Review 110*(1), 52–67.

De Kadt, D. (2017). Voting then, voting now: The long-term consequences of participation in south africa's first democratic election. *The Journal of Politics 79*(2), 670–687.

Gailmard, S. (2024). *Agents of empire: English Imperial governance and the making of American political institutions*. Cambridge University Press.

Gallup, T. G. O. (2025). Lgbtq+ rights. Technical report, https://news.gallup.com/poll/1651/gay-lesbian-rights.aspx.

Geddes, B. (2003). *Paradigms and sand castles: Theory building and research design in comparative politics*. University of Michigan Press.

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

Gerring, J., T. Wig, W. Veenendaal, D. Weitzel, J. Teorell, and K. Kikuta (2021). Why monarchy? the rise and demise of a regime type. *Comparative Political Studies 54*(3-4), 585–622.

Goertz, G. and J. Mahoney (2012). Concepts and measurement: Ontology and epistemology. *Social Science Information 51*(2), 205–216.

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics 48*(1), 80–83.

Grzymala-Busse, A. (2011). Time will tell? temporality and the analysis of causal mechanisms and processes. *Comparative Political Studies 44*(9), 1267–1297.

Grzymala-Busse, A. M. (2002). *Redeeming the communist past: The regeneration of communist parties in East Central Europe*. Cambridge University Press.

Haber, N. A., S. E. Wieten, J. M. Rohrer, O. A. Arah, P. W. Tennant, E. A. Stuart, E. J. Murray, S. Pilleron, S. T. Lam, E. Riederer, et al. (2022). Causal and associational language in observational health research: a systematic evaluation. *American journal of epidemiology 191*(12), 2084–2097.

Haidt, J. (2024). *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness*. Penguin Press.

Hainmueller, J., J. Mummolo, and Y. Xu (2019). How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis 27*(2), 163–192.

Hall, P. A. (2003). Aligning ontology and methodology in comparative research. *Comparative historical analysis in the social sciences 374*.

Henrich, J. (2020). The weirdest people in the world: How the west became psychologically peculiar and particularly propserous. *New York: Farrar, Strauss, and Giroux*, 2–3.

Hernán, M. A. (2018). The c-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health 108*(5), 616–619.

Holland, A. C. (2016). Forbearance. *American political science review 110*(2), 232–246.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association 81*(396), 945–960.

Holmes, C. E., M. K. Guliford, M. A. S. Mendoza-Davé, and M. Jurkovich (2024). A case for description. *PS: Political Science & Politics 57*(1), 51–56.

Homola, J., M. M. Pereira, and M. Tavits (2020). Legacies of the third reich: Concentration camps and out-group intolerance. *American Political Science Review 114*(2), 573–590.

Homola, J., M. M. Pereira, and M. Tavits (2024). Fixed effects and post-treatment bias in legacy studies. *American Political Science Review 118*(1), 537–544.

Hughes, M. M., P. Paxton, A. B. Clayton, and P. Zetterberg (2019). Global gender quota adoption, implementation, and reform. *Comparative Politics 51*(2), 219–238.

Huijsmans, T. and J. Rodden (2024). The great global divider? a comparison of urban-rural partisan polarization in western democracies. *Comparative Political Studies*, 00104140241237458.

Humphreys, M. and A. M. Jacobs (2023). *Integrating Inferences: Causal Models for Qualitative and Mixed-Method Research*. Cambridge University Press.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Jefferson, H. (2023). The politics of respectability and black americans' punitive attitudes. *American Political Science Review 117*(4), 1448–1464.

King, G., J. Pan, and M. E. Roberts (2013). How censorship in china allows government criticism but silences collective expression. *American political science Review 107*(2), 326–343.

Kittel, B. and D. Kuehn (2013). Introduction: Reassessing the methodology of process tracing. *European political science 12*, 1–9.

Kocher, M. A. and N. P. Monteiro (2016). Lines of demarcation: Causation, design-based inference, and historical research. *Perspectives on Politics 14*(4), 952–975.

Kreuzer, M. (2019). The structure of description: Evaluating descriptive inferences and conceptualizations. *Perspectives on Politics 17*(1), 122–139.

Kreuzer, M. (2023). *The grammar of time: A toolbox for comparative historical analysis*. Cambridge University Press.

Kuhn, T. (1970). The nature of scientific revolutions. *Chicago: University of Chicago 197*(0).

Lakatos, I. (1970). History of science and its rational reconstructions. In *PSA: Proceedings of the biennial meeting of the philosophy of science association*, Volume 1970, pp. 91–136. Cambridge University Press.

Lieberson, S. (1987). *Making it count: The improvement of social research and theory*. Univ of California Press.

Little, A. T. and A. Meng (2023). Measuring democratic backsliding. *PS: Political Science & Politics*, 1–13.

Mackie, G. (1996). ending footbinding and infibulation: A convention account. *American Sociological Review December*, 999–1017.

Mahoney, J. (2012). The logic of process tracing tests in the social sciences. *Sociological Methods & Research 41*(4), 570–597.

Miller, J. H. and S. E. Page (2009). *Complex adaptive systems: an introduction to computational models of social life: an introduction to computational models of social life*. Princeton university press.

Munger, K., A. M. Guess, and E. Hargittai (2021). Quantitative description of digital media: A modest proposal to disrupt academic publishing. *Journal of Quantitative Description* (1), 1–13.

Nunn, N. and L. Wantchekon (2011). The slave trade and the origins of mistrust in africa. *American economic review 101*(7), 3221–3252.

Nyborg, K., J. M. Anderies, A. Dannenberg, T. Lindahl, C. Schill, M. Schlüter, W. N. Adger, K. J. Arrow, S. Barrett, S. Carpenter, et al. (2016). Social norms as solutions. *Science 354*(6308), 42–43.

Page, S. E. et al. (2006). Path dependence. *Quarterly Journal of Political Science 1*(1), 87–115.

Paine, J., X. Qiu, and J. Ricart-Huguet (2025). Endogenous colonial borders: Precolonial states and geography in the partition of africa. *American Political Science Review 119*(1), 1–20.

Pepinsky, T. B., S. W. Goodman, and C. Ziller (2024a). Causation and history in legacy studies: A reply to homola, pereira, and tavits. *Social Science Research Network*.

Pepinsky, T. B., S. W. Goodman, and C. Ziller (2024b). Modeling spatial heterogeneity and historical persistence: Nazi concentration camps and contemporary intolerance. *American Political Science Review 118*(1), 519–528.

Przeworski, A. and F. Limongi (1997). Modernization: Theories and facts. *World politics 49*(2), 155–183.

Riedl, R. B. (2014). *Authoritarian origins of democratic party systems in Africa*. Cambridge University Press.

Sacco, R., N. Camilleri, J. Eberhardt, K. Umla-Runge, and D. Newbury-Birch (2024). A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in europe. *European Child & Adolescent Psychiatry 33*(9), 2877–2894.

Schimmelfennig, F. and U. Sedelmeier (2020). The europeanization of eastern europe: the external incentives model revisited. *Journal of European public policy 27*(6), 814–833.

Sen, M. and O. Wasow (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science 19*(1), 499–522.

Simpser, A., D. Slater, and J. Wittenberg (2018). Dead but not gone: Contemporary legacies of communism, imperialism, and authoritarianism. *Annual Review of Political Science 21*(1), 419–439.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological) 13*(2), 238–241.

Spirling, A. and B. M. Stewart (2022). What good is a regression. Technical report, Technical report.

Stasavage, D. (2020). *The decline and rise of democracy: A global history from antiquity to today*. Princeton University Press.

Tausanovitch, C. and C. Warshaw (2014). Representation in municipal government. *American Political Science Review 108*(3), 605–641.

Tavory, I. and S. Timmermans (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Treisman, D. (2023). How great is the current danger to democracy? assessing the risk with historical data. *Comparative Political Studies 56*(12), 1924–1952.

Verghese, A. (2024). Randomized controlled history? *Available at SSRN 4696790*.

Voigtländer, N. and H.-J. Voth (2012). Persecution perpetuated: the medieval origins of anti-semitic violence in nazi germany. *The Quarterly Journal of Economics 127*(3), 1339–1392.

Wang, Y. (2022). *The rise and fall of imperial China: The social origins of state development*. Princeton University Press.

West, G. (2017). *Scale: The universal laws of life, growth, and death in organisms, cities, and companies*. Penguin.

Wittenberg, J. (2006). *Crucibles of political loyalty: Church institutions and electoral continuity in Hungary*. Cambridge University Press.

Yom, S. (2015). From methodology to practice: Inductive iteration in comparative research. *Comparative Political Studies 48*(5), 616–644.