Good Description *

Daniel de Kadt[†] Anna Grzymala-Busse[‡]

November 20, 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 uncover facts in need of explanation, build and refine theory, or help to assess and revise theory. Second, we articulate three ideal characteristics of descriptive *analyses*: clarity, comparability, and completeness. Good description is thus tightly and transparently linked to specific research questions (and thus to theory), well contextualized, and as comprehensive in measurement and specification as possible.

*For their very thoughtful comments, we are grateful to Lisa Blaydes, Alex Coppock, Jill Frank, Vicky Fouka, Sean Gailmard, Miriam Golden, Sebastian Karcher, Marcus Kreuzer, David Laitin, Andrew Little, Tom Pepinsky, Solé Prillaman, Carlisle Rainey, Melissa Sands, Arthur Spirling, Klaudia Wegschaider 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, particularly in political science. Gerring (2012) observes 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 disdain for "mere" description. This concern has become more widespread (see for example Kreuzer (2019) and Holmes et al. (2024)), even prompting the creation of new journals in response (Munger et al., 2021).

At the same time, leading journals continue to publish important and highly cited descriptive papers. 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, recent prominent and even prize-winning books in political science have relied heavily on description (Beissinger, 2022; Blaydes, 2018; Daly, 2022; Stasavage, 2020; Wang, 2022). Entire disciplinary debates are sometimes descriptive in nature – for example, recent disputes about democratic resilience and backsliding (Little and Meng, 2024; Knutsen et al., 2024; Gorokhovskaia, 2024; Bergeron-Boutin et al., 2024; Treisman, 2023, 2024; Levitsky and Way, 2023; Miller, 2024; Baron et al., 2024; Weitzel et al., 2024). 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).

The contested status of description reflects the absence of a coherent framework for distinguishing descriptive work that is valuable from that which is not. This has two consequences. First, absent a clear account of the social scientific purpose of description, we are uncertain about what makes for good descriptive *questions*. Second, without shared standards about empirical efforts to answer even good descriptive questions, it is unclear what makes for good descriptive *analyses*. By contrast, causal inference benefits from both a well-developed understanding of its purpose, and a reasonably clear set

of standards against which to measure applied work, supported by many textbooks and 'checklist' or 'how to' articles.

We begin from the premise that descriptive and causal inference are not different modes of social scientific inquiry. Different questions may require different research designs, assumptions, data, and statistical or qualitative methods, but all of these questions and subsequent analyses serve the same fundamental goal: the incremental and cumulative development and evaluation of theories about the social world. The value of empirical research depends on whether and how it generates insights into theory and concepts. In this way, good description is no different from good causal inference. It is an essential empirical social scientific task, aimed at answering distinctive questions, and with clear criteria for evaluation. Description is not a lesser mode of inquiry, but foundational to social science.

We propose a framework that distinguishes 'good' from 'mere' description, useful both for conducting descriptive research and for assessing it as readers, advisors, reviewers, or editors. Our contribution is twofold: first, we articulate the social scientific goals of description and explain what makes for a good descriptive question. Second, we specify the characteristics of good descriptive analysis.

Good description, we argue, is a theory-driven exercise: it is informed by, motivates, or revises relevant and important theory, imposing structure on the social world. Good description highlights facts, patterns, anomalies, or puzzles in need of explanation, and signals gaps or flaws in existing theory. Knowing (something about) the state of the world is necessary, if not always sufficient, for constructing theory. Good description can highlight the absence of well-developed theory, and it can be generative: scholars can create meaningful analytical entities when they choose to describe a phenomenon.

Good description can also be used to refine and 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. Good description provides answers to relevant and important *questions* through well-calibrated and meaningful *analyses*, which produce credible and insights about 'how the world is or was.' Critically, while social scientific theories are usually causal, many of the derived research questions and analyses need not be. We develop these arguments in Section 2.

Second, good description has three ideal features: clarity, comparability, and completeness. Clear analyses are tightly and transparently linked to theory and well-calibrated to the question at hand. Comparable analyses are well contextualized, so that findings can be understood in the context of existing and future evidence. Complete analyses are as comprehensive as possible, in both measurement and specification. Descriptive studies can be evaluated in terms of these three criteria, detailed in Section 3. We conclude by providing examples of how good description can improve other forms of analysis, by clarifying contexts and informing the study of causal analysis in Section 4.

2. Theories, Questions, and Description

Most 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 about the causal connections between variables or concepts. Humphreys and Jacobs emphasize that social scientific theory tends to be causal, offering "an account of how or under what conditions a set of causal relationships operate" (Humphreys and Jacobs, 2023, 163).[1]

To illustrate, we adopt a toy example of a simple causal *theory*: phenomenon A affects phenomenon B in some unspecified way and magnitude. In the language of directed acyclic graphs (DAGs), we might represent this theory as $A \longrightarrow B$.

Given this toy theory, researchers could pose a range of research *questions*. They could ask questions about the conceptual nature of A, B, or both A and B, such as what properties constitute A (B)? They could ask questions about empirical properties or characteristics of A, B, or both A and B: what is A (B) in practice, when do they occur, or how often? They could ask questions about association between A and B: do A and B typically co-occur? Is that co-occurence conditional on other variables? They could ask causal questions: does A affect B, is A necessary and/or sufficient for B, under what conditions

[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).

does this relationship hold, or what is its magnitude? Finally, they may ask mechanistic questions: how does A affect B?

Two observations follow. First, all these questions are derived from theory, and that is what gives them scientific meaning. If no theory features phenomenon A (B), the value of describing the properties of phenomenon A (B) is generally limited. Second, answers to these questions can shed light onto the plausibility of the theory, even without an explicit and dispositive test. This is a not uncontroversial claim, but consider a very simple example: if we do not know whether or when A and B occur, then we can learn little about the core theoretical claim that A affects B.

Theories, then, can generate both causal and descriptive propositions, which inspire different research questions. How do we distinguish between causal and descriptive questions? One view is that descriptive questions can be identified using natural language. For example, Gerring (2012)[72-73] argues that descriptive questions focus on "*what* questions (e.g., when, 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".

Such language-based definitions are useful but imprecise.[2] We propose a different definition, following Holland (1986). The distinguishing feature of causal and descriptive questions is manipulability: if answering a research question requires that at least one concept or variable is manipulable, that question is causal. If manipulability is not required, then that question is descriptive. For example, asking whether A and B correlate does not require that we manipulate A or B: the question is descriptive.

This distinction is not purely theoretical. Consider race and wages. There is a broad consensus that because an individual's race cannot be manipulated, their race cannot be said to affect their wages. Asking whether members of one racial group earn less than others is thus a descriptive question: it concerns the state of the world, not the counterfactual state of the world under manipulation. But

[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.

racial discrimination, which is manipulable, can and does affect wages (for a detailed discussion see Sen and Wasow, 2016). Thus, asking whether certain labor market interventions reduce (or exacerbate) the racial earnings gap is asking a causal question. From a language-based perspective, description addresses questions of the form 'what is the world like?' Causal inference addresses questions of the form 'what would a counterfactual state of the world have been like?'

Returning to our toy theory, some questions about A and B are what one might consider canonically *descriptive* questions – they seek only to describe the state 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 underlying *theory* that a researcher posits, and the specific *research question* that a researcher assesses. This distinction is often blurred, creating tension between the language researchers use to describe their research goals and the empirical claims they actually make.

Researchers clarify this distinction early in the research process, articulating both the underlying theory of interest and the specific research question(s) derived from it. While these may coincide, they often do not. Empirical analyses should evaluate a well-specified research question, even if they cannot directly test the underlying theory. Indeed, sometimes this is simply not possible: perhaps no compelling research design exists to isolate variation in A. Good description requires explicitly descriptive research questions clearly linked to theory, and empirical analyses that are correctly specified to target these (important and useful) research questions.

2.1. Theoretically Informed Description

For description to be valuable, it must be grounded in social scientific theory. This requires articulating precisely what researchers are attempting to describe, and the status of their questions and claims. Researchers engaged in description must have a well-formed understanding of both the phenomenon of interest – the thing – and the relational context of that phenomenon – the question. This criterion

[3]Normative political theory and philosophy motivates both descriptive and causal analyses and informs what we consider relevant and important phenomena.

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 at stake.

Consider the toy example outlined earlier: A $\longrightarrow$ B. Here we have two theoretically-derived concepts or phenomena, thing A and thing B. Figure 1 shows an illustrative ladder of some social scientific questions one might ask about those two things. Two points are important. First, while the rungs (labeled in bold) capture broad categories, the they are not exhaustive: many more specific questions could be posed at each rung. Second, the 'ladder' metaphor does not imply that any question is *more important* than any other. Rather, 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 – be that A by itself, B by itself, or both. Researchers might investigate what precisely A or B *are*, whether at a conceptual level (ontological questions about what constitutes the phenomenon) or at an observational level (an empirical question about what constitutes it). Chandra (2006), for example, conceptualizes 'ethnic identity' as determined by descent-based attributes. Description could help to identify the constituent elements of 'ethnic identity' and test whether these exist. At the same time, by choosing to measure ethnic identity at all, researchers may help to create or reinforce it as a meaningful category, both in research and in the real world. Description can thus structure the empirical world.

One rung down are questions about characteristics: for example, how often A and/ or B occur. Here again, description can create new concepts or refine existing ones, imposing new categories and classifications on the social world. Good description must be alert to the possibility that the very act of describing phenomena invariably imposes our own (mis)understandings on the world, importing and reinforcing those misunderstandings. Conversely, good description may also reveal to us where our underlying conceptualizations are misguided.

A further rung down are associational questions, such as whether A and B occur together. Below those are questions of conditional association: conditional on some other factors, does the co-

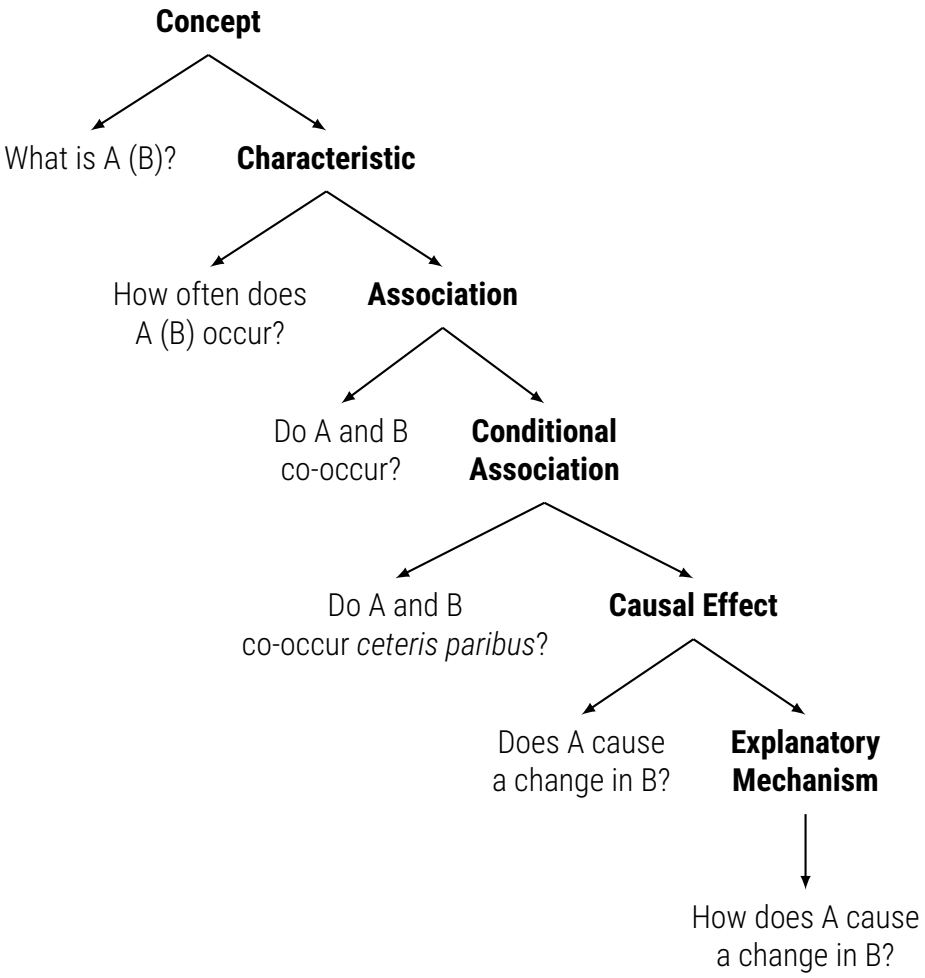

Figure 1: An illustrative ladder of possible research questions, derived from a simple theory: A⟶B.

occurrence of A and B still hold? On rung four we find causal relationships: does A change B, and is it a necessary and/ or sufficient causal factor? Finally, at the bottom of the ladder are questions about mechanisms: how any effect of A on B happens.

There is inter-dependency here: we will struggle to answer a lower question unless we are able to answer all the higher questions (whether we actually answer those questions is a separate issue, but it should be possible to do so). We cannot empirically assess mechanisms unless we first establish causation. We cannot establish co-occurrence without knowing how often A and B occur. And we cannot even describe frequencies unless we first know what A and B conceptually represent.

Despite this inter-dependency, researchers often skip conceptual and characteristic questions, and move directly to analyzing co-occurrence and causation. Yet choices made at the higher rungs – whether explicit or implicit – have consequences at every rung below. Consider the ongoing debate about the status of democracy worldwide. Some scholars advocate for minimalist definitions of democracy (e.g. focused on alternations) and measure it accordingly (e.g. Przeworski and Limongi, 1997; Przeworski, 2000). Others advocate more expansive, multidimensional definitions (e.g. Coppedge et al., 2011, 2019; Herre, 2022). These conceptual choices shape measurement, and cascade down the ladder, producing markedly different conclusions about the status of democracy – exactly as in the case in Little and Meng (2024) and Knutsen et al. (2024). Fundamentally, this is a conceptual dispute, not just an empirical one.

Returning to the ladder, which of these types of questions are typically 'descriptive?' Rungs 1, 2, and 3 are the most clearly descriptive. Rung 4 is more ambiguous, and researchers should exercise a great deal of caution. Before the 'credibility revolution' researchers often treated such conditional association as evidence of causal relationships through *ceteris paribus* comparisons. Many questions from rung 4 have been re-posed so that they fall under rung 5 (causal effects). Yet absent such explicit re-wordings, conditional associations are descriptive: they do not require manipulability. For example, in discussions of the gender pay gap, scholars rarely claim that gender *causes* wages. Rather, they argue that wages vary systematically by gender, even after conditioning on job type, sector, educational attainment, etc.

This result demonstrates a residual wage gender gap without implying causation. Such questions often occupy a liminal space between descriptive and causal inference. It is the necessity of manipulability that distinguishes the two.

The broader lesson is that researchers must be explicit about the types of questions they are trying to answer. As Hernán (2018) notes, researchers often speak euphemistically about the questions they are answering. They may aspire to answer a causal research question (rung 5), but use descriptive language because they do not believe they have the appropriate research design or identification strategy. But the quality of the evidence available does change the nature of the question itself. If the available evidence can only support descriptive claims, researchers should ask good descriptive questions (e.g. from rung 1, 2, 3, or potentially 4). Good description, then, is not simply second-rate causal inference. Indeed, ambiguous language can make interpretation more challenging (Haber et al., 2022).

2.2. The Goals of Good Description

Given this analytical ladder and the central role of theory, good description is an empirical exercise designed to describe phenomena of interest with three main goals:

1. Reflect or invoke theory in a conscious and clear way: The things we measure or characterize should be theoretically relevant. Our questions connect clearly to theory, and demand a descriptive answer.

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

3. Evaluate theory by assessing its implications: Some causal theories generate non-causal observable implications. Good 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. Across the social sciences, and certainly in political science, there is a rich tradition of measuring the state of the world. These measurement efforts have clearly always been theoretically guided. Graduate training emphasizes '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 accurately capture theoretical concepts of interest, whether derived from positive or normative theory. Many of these concepts are 'primitives,' or foundational building blocks, such as regime type, a nation's wealth, or the level of inequality. They are important precisely because they are theoretically relevant. For example, we measure democracy because we believe that democracy itself is a fundamental theoretical (and / or ethical) concern.

Reflecting theory is a necessary (if not sufficient) condition for valuable measurement. Measures with poor construct validity contribute little to either descriptive or causal arguments. Indeed, for measurement to be valuable, measures need to capture concepts that underpin theoretical arguments. For example, Huijsmans and Rodden (2024) measure urban-rural partisan divides across Western democracies. Doing so requires a theoretical understanding of partisan voting and partisan loyalties in each context, and why they might vary between cities and rural areas, a point we return to in Section 3 when discussing comparability.

Yet description is much more than measurement. It can help to build theory by describing the social world more accurately or comprehensively, enabling researchers to perceive patterns or puzzles they would otherwise have not explored or explained. Anomalies and outliers only exist in light of existing theoretical expectations, and accounting for them can help to strengthen or revise theory. Scholars have accordingly called for a recursive (and transparent) moving back and forth between empirics and theoretical models, refining each in light of the other (Geddes, 2003; Tavory and Timmermans, 2014; Yom, 2015; Humphreys and Jacobs, 2023).

Description also helps evaluate theories. Causal theories often have non-causal implications, and description is especially useful in *disconfirming* such implications. For example, one simple evaluation

of causal theories through good description is to see whether the predicted correlations exist. If they do not, or occur less often than expected, that suggests the theory is either incomplete or flawed. The prediction may not follow from the theory, or the relationship may be conditional on some other factors ignored by the theory. The disconfirmatory value of not finding the expected correlation often outweighs the confirmatory value of finding it. Description thus helps to reject weaker explanations and to build and refine new ones.[4]

Such descriptive analysis is distinct from causal inference, but the two are strongly connected. A typical admonition in causal research is to evaluate the assumptions underpinning the research design. These assumptions typically related to beliefs about the assignment mechanism – how levels of the causal variable are assigned to units (Imbens and Rubin, 2015). These assumptions are part of a broader theory (often formally specified via directed acyclic graphs (DAGs) (Pearl, 2009)) that states the researcher's beliefs about the variables that determine both A and B (in our toy example). Descriptive evidence can challenge or question that theory, by questioning the assignment mechanism or some other parts of the hypothesized DAG.

Descriptive inference can, by implication, be used to assess the plausibility of the assumptions underlying empirical strategies. For example, in spatial regression discontinuity designs (RDD), the researcher studies the effect of some treatment, such as a policy, that only applies within a specific geographic boundary. This design relies on continuity in potential outcomes across the boundary, which is implausible if other things (e.g. other policies) also change at the boundary, or if the boundary is itself a function of the treatment. Historical description can interrogate these assumptions, as demonstrated by Kocher and Monteiro (2016) and Verghese (2024). For example, African state borders, often used for spatial RDDs (see McCauley and Posner, 2015), were not randomly drawn by colonizing powers, but reflect strategic geographic focal points and complex negotiations with indigenous rulers (Paine et al., 2025).

[4]Blackwell et al. (2024) also argue that well-intentioned analyses can mislead when the underlying theory is incorrectly specified. This emphasizes the importance of clear and precise theoretical statements – it is hard to evaluate a theory if researchers are not clear about what it is.

3. The Characteristics of Good Description

We have argued that the purpose of description, like that of causal inference, is to improve our understanding of the social world. To that end, description should be theoretically informed and focus on theoretically relevant objects. Researchers should precisely articulate research questions, and calibrate analyses accordingly. How, then, can we evaluate whether a descriptive exercise achieves these goals? We propose three criteria: clarity, comparability, and completeness.

These three characteristics or criteria follow from the goals of description. To measure or generate concepts, we must be clear about their connection to theory and to potential analytical methods. If we choose to describe A, then both the concept A and the methods used to describe it should be clearly linked to our scientific goal. If description is to build theory by pointing to patterns of occurrence, the measurements across units have to be comparable to each other, and consistent across time and space. Finally, measures should be as complete as possible, so they can fairly evaluate theoretical implications.

Clear description is securely linked to theory. It is evident why the descriptive analysis is valuable theoretically (why do we want to describe this phenomenon?), precisely what the analysis is to evaluate (what is the precise descriptive question?), and how the analysis was conducted (what is the method of description?). Descriptive research questions should be posed with clear reference to theory, and methods chosen are appropriate to the research question. Clarity also helps to build better theory via induction: for example, theories of democracy and democratic backsliding gained from discussions of descriptive research questions and debates over measurement (Przeworski, 2000; Boix and Stokes, 2003; Little and Meng, 2024; Knutsen et al., 2024; Gorokhovskaia, 2024; Bergeron-Boutin et al., 2024; Treisman, 2023, 2024; Levitsky and Way, 2023; Miller, 2024; Baron et al., 2024; Weitzel et al., 2024).

Our second criterion is comparability. Comparable description pays attention to how context affects measurement and meaning, ensuring consistent comparisons. Many questions are meaningless without comparability: for example, we cannot answer how many democracies exist in the world to-

day without criteria that are meaningful and applicable across space and time. Therefore, researchers should consider the underlying conditions embedded in their descriptive exercise. Data produced in one context (population, time, or place) may not be comparable to data produced in another. This is especially important since many descriptive exercises feed into analyses that compare distinct populations, times, or places.

Comparability implies at least three considerations regarding measurement, meaning, and analysis. First, measurement needs to consider how data is generated. For example, the reliability of data from autocratic and democratic governments may vary. Measures of wealth from the 12th and 21st century likely reflect differences in state capacity to measure wealth. Governments measure national wealth differently (Herrera, 2011), and count infanticides as stillbirths (Drixler and Matsuzaki, 2025). Official statistics are often poorly collected, politicized, and selectively presented (Huff, 1954). Sensitive numbers, such as victims of wartime conflict or illicit activities, are especially prone to distortion (Greenhill and Andreas, 2011).

Second, measures may have different and incompatible meanings. A high score on measures of violence against journalists (Coppedge et al., 2019) can suggest democratic erosion and autocratic terror. It can also indicate the weakness of an authoritarian regime and its inability to successfully intimidate journalists. Low levels of violence may mean strong media protections–or 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.

Third, the same analyses may not mean the same thing across contexts. A regression that controls for Z and Q may mean different things if the underlying data generating process that connects Z and Q to dependent and independent variables is different. Researchers should be also sensitive to different data distributions. The central tendency, such as a mean or median, may be an adequate representation of a wealth distribution in egalitarian Norway or Denmark, but highly misleading in highly unequal places like Brazil, the United States, or South Africa.

Comparability thus means that measurements are consistent across contexts, and reflect the context that produced the data. This may generate a tension between generalizability and sensitivity to context. Measurement is not useful if it is idiosyncratic and unique, limiting comparison. But measurement also needs to incorporate contexts, lest it refer to disparate phenomena and obscure substantive differences. Descriptive work that manages this trade-off transparently is highly valuable.

Our final criterion is completeness. Complete description answers as many descriptive questions as possible given a theory of interest and/ or assesses an appropriate sample given a relevant and well-defined (super-)population. The more comprehensive our answer to 'what is A?', the more reliably we can answer questions about how often A occurs, whether A and B occur together, and so on, including the causal questions lower on the ladder.

Completeness can be understood in at least three ways. First, given a theory, has the descriptive exercise studied as many (useful and important) descriptive questions as possible? Have we fully measured the aspects of A specified by the theory? Second, has the researcher specified the theoretically appropriate population, and assessed the appropriate sample for meaningful population-level inferences? If one hopes to describe A, but has only a biased (and uncorrected) sample, analyses may be incomplete. Third, completeness may implicitly trade off depth and breadth. Researcher may study deeply a single case at potentially only few points in time. Or they may study broadly many cases at potentially many points in time. Either of these approaches can be complete; failure to do either suggests incompleteness.

The measurement of democracy illustrates these issues. Minimalist (e.g. Przeworski and Limongi, 1997; Przeworski, 2000) and expansive (e.g. Coppedge et al., 2011, 2019; Herre, 2022) definitions imply different standards of completeness. The population of interest – all the countries in the world, or a subset thereof (say, post-industrial societies) – also shapes what counts as complete. Here, note the trade-off: minimalist descriptions may lack conceptual coverage, while expansive ones may sacrifice geographic or temporal coverage.

3.1. From Question to Analysis

For descriptive analyses to be clear, comparable, and complete, the analytical approach must be carefully matched to the specific research question. This can be challenging, particularly for questions on rungs 3 and 4 of Figure 1. Consider some very successful descriptive work, 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 this claim characterizes an association (rung 3 or 4 on our ladder, depending on whether the association is conditional), its scientific value hinges on the specific question and underlying theory. For their descriptive inference to be valuable given their theory of interest, the authors attempt to rule out competing explanations by making a *ceteris paribus* claim (Spirling and Stewart, 2022).

This tension is ingrained in the analytical choices that scholars make, often without acknowledgment. One common descriptive methodology in quantitative social science, for example, is multiple regression to establish conditional associations. Researchers make an associational claim of 'what goes with what' (rung 3 of our ladder), yet they actually analyze 'what goes with what, conditional on other variables' (rung 4). Whether this is appropriate depends on the theory or theoretical concepts the authors are attempting to measure, build, or evaluate.

How should we interpret this exercise? If the researcher claims they are not engaged in counterfactual reasoning, but only want to 'describe the data,' then multiple regression itself is a peculiar choice. Regression by design extrapolates beyond the support of the data, into counterfactuals. If one 'controls' for variables, it is because one has a theoretical model that connects these variables causally. Indeed, Stanley Lieberson cautioned against the misuse of controls as 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 the model asserts that a treatment is 'as-if randomly' assigned, conditional on pre-treatment covariates. Alternatively, without a causal model, as Simpson's paradox, where a global unconditional relationship reverses in subsets of the data (Simpson, 1951). Without adjusting for the 'correct' covariate(s), we are misled and draw the 'wrong' conclusion from the data.

Yet this exercise only makes sense when it reflects theory. If theory predicts an unconditional correlation, conditioning is not necessary. Likewise, if question is simply 'does X correlate with Y,' conditioning is not necessary. However, if theory suggests a conditional relationship, then conditioning is necessary. 'Rightness' and 'wrongness' is defined by the theory and the research question. The global relationship in the case of Simpson's paradox is not 'wrong' unless the target is the conditional sign. Often, however, researchers are not explicit about their goal. It may be a study of conditional associations (rung 4). Yet it is often counterfactual reasoning (rung 5), where controls "rule out alternative explanations" (Spirling and Stewart, 2022, 17).

If one wants to simply 'describe the data' one could do this without regression. Yet the instinct to control dominates. One might even argue that causal inference studies of a single case are themselves descriptive. That is, much recent work in social science uses causal inference tools to analyze specific case studies, not to engage in comparison. These studies provide precise estimates or assessments of one data point, useful for future comparative or meta- analyses. By explicitly situating their questions on the 'ladder' in Figure 1, researchers can pinpoint 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 advance, and three criteria for good descriptive analyses. We now illustrate how good description can work in practice. First, good description can specify the relevant theoretical and empirical contexts for valuable comparisons and credible identification. By providing baselines and denominators, it helps to evaluate theory

and provide plausible scope conditions. Second, description can illuminate how phenomena unfold over time and place, clarifying the pathways and mechanisms that underlie causal relationships.

4.1. Diversity of Contexts

Good description provides theoretically-relevant context and situates measurements and findings. Sensitivity to context requires knowledge of culture, norms, reference points, details that improve the quality of data collection and interpretation (Cirone et al., 2021; Greenhill and Andreas, 2011), and attention to how phenomena vary across time and space (Hall, 2003, 383), (Tavory and Timmermans, 2014, 72). Such contextual knowledge also makes it possible to correctly interpret data-driven claims. For example, we might conclude that media freedom in Russia has improved because fewer journalists are now killed, when in fact the Putin regime increased repression so that it no longer views journalists as a threat (Gorokhovskaia, 2024, 182). Good description thus both relies on, and makes for, good measurement.

Good description also specifies the comparative context: other situations in which an outcome occurs, the distribution to which the outcome may belong, and the chronological timelines or empirical baselines against which we evaluate the outcome. If we ask whether democracy worldwide is robust or declining, for example, we need to know the status quo ante (Little and Meng, 2024; Knutsen et al., 2024). Here, description helps to build theory by specifying scope conditions and the variation in the phenomena that the theory seeks to explain.

Such context-sensitive description can demand more nuanced theory and reveal incompleteness. For example, US data suggest that mobile phones as a source of depression and anxiety in teenagers (Haidt, 2024). Yet in Europe, despite similarly high rates of cell phone use, rates of adolescent anxiety and depression among European adolescents are lower (Sacco et al., 2024). This finding suggests that a monocausal explanation is incomplete.[5] Sensitivity to context can expose incompleteness: unspeci-

[5]To be clear, a lack of observable correlation does not mean a lack of causation. Instead, the theory of how either the causal variable is assigned or how the dependent variable is generated is likely underspecified or incomplete.

fied confounders that mask (amplify) the true relationship outside (inside) of the US, or treatment effect heterogeneity where the effect emerges only under some conditions.

Similarly, most political phenomena arise from multiple and interacting causes. An explanatory variable A may cause change in Y, but it can be a conditional cause, operating only in the presence of other factors. Good description can specify these conditional relationships, for example by incorporating moderators or interaction terms (Clark et al., 2006; Hainmueller et al., 2019). If a theory proposes that Z is a conditioning or moderating variable, then the effect of A on B should vary depending on the value of Z. Researchers can propose descriptive questions about heterogeneous treatment effects – for example, does the effect of A on B vary by levels of Z? These are descriptive questions about Z (which is not required to be manipulable), but causal questions about A and B. Good description both measures A, B, and Z, and clarifies how any marginal effects of A might be conditional on Z (Berry et al., 2012, 654). Yet again, description is valuable only if what we measure, and the descriptive analyses we conduct, are clearly connected to a theory that specifies conditional relationships.

Another way context develops comparable and complete description is by situating outcomes within plausible distributions. We want to know how unique or general a given concept is, and how stable its attributes across contexts (Kreuzer, 2023, chapter 5). Are observed cases or relationships drawn from around the mean of a normal distribution of outcomes? Or are they outliers, several standard deviations from the mean? For example, some scholars argue that China and India are two powerful outliers that drive conclusions about democratic decline (Treisman, 2023). An outcome that is exceedingly rare raises questions about the importance or completeness of the theory, and apparent treatment effects here may simply reflect noise. It is thus important to know just how unusual these cases are, and how much they might contribute to our conclusions.

Distributions themselves provide insights. A bimodal bunching of outcomes not only calls for different statistical tests, but can also indicate fragmentation, polarization, social distancing, and so on. Outcomes that follow a 1/x power-law distribution, such as Zipf's law or Pareto's law, have fatter 'tails' than normal distributions, with implications for both theory building and testing. First, rare but big

events (such as very large cities, common words, or very rich people) are more common than with a normal distribution. Second, outcomes may be self-similar: small conflicts mirror large wars, and vice versa. Third, power-law distributions arise from multiplicative or correlated processes, while normal distributions arise from additive independence (West, 2017; Miller and Page, 2009). For example, positive feedback may result in a power-law distribution, not a normal one. Assumptions of independence may fail, and models relying on normal distributions mislead. Building theory calls for specifying the independence or interdependence that leads to these patterns.

Finally, denominators, historical baselines, and base rates are critical. To know whether a phenomenon is rising or declining, we need a historical reference point. To measure in absolute terms, we need to count the units and changes over time. For example, more democracies globally may indicate the triumph of liberal democracy – or the change in the number of countries and new states (the denominator). Debates about democratic backsliding hinge on indicators and their baselines (Treisman, 2023; Little and Meng, 2024). For example (Miller, 2024) argues that democratic backsliding should be evaluated only within democracies, tracking year to year changes. Descriptions that is clear, comparable, and complete allows context-sensitive and principled comparative research.

4.2. Time, Place, and Unfolding Phenomena

The analysis of how phenomena unfold over time and place is another critical contribution of good description. A wealth of important and related work has examined 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 seek to establish causation (Collier et al., 2010; Collier, 2011; Mahoney, 2012). Scholars have called for more attention to geographic and historical trends, transformations over time, and continuities and discontinuities as part of causal claims (Kreuzer, 2023). Yet more modest empirical efforts – charting how events occur across time and space and how historical legacies unfold–can also yield valuable insights without making causal claims.

First, good descriptive measurement identifies the 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 (religiously mandated periods of mourning, holidays, electoral cycles, or census periods), or episodes (wars, famines, or regime collapses). This measurement must be relative to the relevant population. By tracing variables or events across both time and space, we can learn how sustained they were, when they faded out, and which potential mechanisms are at work. We also trace how whether and how they diffuse. Key questions include: who adopted the practice and when? When do contiguous units follow? When is a given environment saturated? Answering these questions requires specifying where and when phenomena arise, and evalute any evidence of 'contagion' or spread.

Second, good description can track the maintenance, growth, or contraction of phenomena in reference to theory. For example, historical legacy arguments, including long-run persistence studies, examine how historical events shape contemporary outcomes, using process-tracing, causal inference, or mechanistic approaches (Simpser et al., 2018; Gailmard, 2024). Examples include studies linking colonial settlement to contemporary economic and political development (Acemoglu et al., 2001), autocratic rule to democratic competition (Grzymala-Busse, 2002; Riedl, 2014; Wittenberg, 2006), Nazi rule to postwar economic development (Charnysh, 2019; Charnysh and Finkel, 2017; Homola et al., 2020), or slavery to modern social trust (Nunn and Wantchekon, 2011; Acharya et al., 2016). Such studies show that legacies can persist long after the conditions that gave rise to them have disappeared (De Kadt, 2017; Simpser et al., 2018).

Critics charge that these accounts often underspecify mechanisms of how long-term outcomes arise. 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. Legacy arguments are often also circumspect about the durability of impact, and its distribution or evenness: where and when they had their greatest impact and for how long. Good description would include repeated, context-sensitive measurements across time and place that show

whether the causal factor and mechanisms were sustained, attenuated, or gave way to another link in the causal chain.

Good description can also challenge or refine historical explanations, through clear measurement and contextual comparison. For example, Voigtländer and Voth (2012) link medieval pogroms to 20th century antisemitism via intergenerational socialization. Similarly, Homola et al. (2020) argue that proximity to Nazi concentration camps increased voting for far-right parties, the result of exposure to Nazi institutions and subsequent cognitive dissonance. Much as in Voigtländer and Voth (2012), family socialization transmits beliefs across time. Yet this account has also been questioned on the basis of descriptive knowledge. Pepinsky et al. (2024b) note the variation in school curricula and civic education across German Länder which likely shape beliefs and confound Nazi legacies.[6] This debate underscores the need to specify which historical factors are preserved and transmitted, and when they attenuate in response to shocks, political evolution, or cultural exposure via trade or education (Voigtländer and Voth, 2012).

Third, attention to temporality, such as duration, tempo, timing, and sequencing can help to identify or eliminate causal mechanisms. For example, if democratic backsliding is conceptualized as a slow-moving phenomenon, a rapid drop in democracy indicators suggests a coup rather than slow decay (Knutsen et al., 2024; Weitzel et al., 2024). Similarly, rapid-institution building is more likely to involve existing skills, networks and templates, rather than deliberation, consensus building, or slow learning (Grzymala-Busse, 2011). Rapid change can also preclude outcomes: for example, swift repression can prevent protest from evolving into a revolution. This temporal context provides clues about plausible links between causes and effects.

Descriptive attention to growth dynamics can capture non-linear patterns, such as exponential growth or behavioral tipping in social and political norms (Mackie, 1996; Nyborg et al., 2016). For example, support for same-sex marriage suddenly shifted in the early 2000s, reaching majority approval in

[6] As a solution, Pepinsky et al. (2024b) 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).

the US by 2011 (Gallup, 2025). Geometric growth is very hard to observe initially, with the multiplication of very small quantities that eventually culminate in rapid and intense expansion. Such patterns are 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. Linking measurement to theory about nonlinear dynamics helps to detect them.

Tracing sequences is especially important in path-dependent processes, where the outcomes depend on event ordering and the accumulation of externalities (Page et al., 2006). Distinguishing path-dependence from persistence requires documenting the sequence of events and showing how they foreclosed alternatives. The description-based counterfactual reasoning we outlined earlier in Section 2 and exercises such as placebo tests can reveal how sequences and accumulations blocked other trajectories

Analyses that are clear, comparable, and complete identify the relevant units, track growth dynamics over time, and capture non-linear patterns and sequences. Good description can then challenge historical accounts, identify or eliminate plausible causal mechanisms, and distinguish between path dependence and persistence.

5. Conclusion

In recent decades, social science has undergone several methodological developments. Statistical, computational, and qualitative advancements have deepened and broadened our analytical toolkit. Formal modeling clarified the need to specify the micro-foundations and strategic interactions. The identification revolution underscored the importance of research design and assumptions for causal inference. A unifying feature of these innovations is the development of clear evaluative criteria. Our aim has been to provide a comparable framework for description, which connects to and underpins all of these enterprises.

Description is the scientific art of connecting well specified theory, usually causal in nature, to appro-

priate computational, statistical, and qualitative analyses to generate insight into important questions. Description is therefore a theoretically-informed and empirically rich mode of inquiry. It is both a necessary first step for various forms of inference, and a valuable exercise in its own right. We encourage researchers to argue clearly and explicitly where the value of their work lies.

'Good' description engages theory by invoking and refining existing theories, motivating new theoretical models, and even evaluating theory. It does so by specifying relevant phenomena, highlighting departures from expected distributions, or revealing patterns that ought (not) to exist given our theory. It helps to define scope conditions, identify gaps in theory, and test observable implications. Good description can specify the distribution, context, and unfolding of a variety of important phenomena. In applied work, this can mean specifying relevant contexts, and describing how phenomena unfold across time and space. Here, mapping and situating may be as important as identifying causal effects.

We propose that both producers and readers of descriptive work evaluate its quality by examining how clear, comparable, and complete it is. Good description clearly articulates its connection to theory, specifies relevant context (through comparisons, distributions, or tracing change over time and place), and employs valid and comprehensive measures. Such 'good description' goes hand in hand with other models of inference – and when done well, represents a valuable and fundamental mode of modern social science.

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.

Baron, H., R. A. Blair, J. Gottlieb, and L. Paler (2024). An events-based approach to understanding democratic erosion. *PS: Political Science & Politics 57*(2), 208–215.

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.

Bergeron-Boutin, O., J. M. Carey, G. Helmke, and E. Rau (2024). Expert bias and democratic erosion: Assessing expert perceptions of contemporary american democracy. *PS: Political Science & Politics 57*(2), 184–193.

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.

Boix, C. and S. C. Stokes (2003). Endogenous democratization. *World politics 55*(4), 517–549.

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.

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.

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.

Drixler, F. and R. Matsuzaki (2025). Façade fictions: False statistics and spheres of autonomy in meiji japan. *Politics & Society 53*(1), 57–97.

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.

Gorokhovskaia, Y. (2024). Difficult to count, important to measure: assessing democratic backsliding. *PS: Political Science & Politics 57*(2), 178–183.

Greenhill, K. M. and P. Andreas (2011). *Sex, drugs, and body counts: The politics of numbers in global crime and conflict*. Cornell University Press.

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*.

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.

Herre, B. (2022). The "varieties of democracy" data: how do researchers measure democracy? *Our World in Data*.

Herrera, Y. M. (2011). *Mirrors of the economy: National accounts and international norms in Russia and beyond*. Cornell University Press.

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.

Huff, D. (1954). *How to Lie with Statistics*. Norton.

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.

Knutsen, C. H., K. L. Marquardt, B. Seim, M. Coppedge, A. B. Edgell, J. Medzihorsky, D. Pemstein, J. Teorell, J. Gerring, and S. I. Lindberg (2024). Conceptual and measurement issues in assessing democratic backsliding. *PS: Political Science & Politics 57*(2), 162–177.

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.

Levitsky, S. and L. A. Way (2023). Democracy's surprising resilience. *Journal of Democracy 34*(4), 5–20.

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

Little, A. T. and A. Meng (2024). Measuring democratic backsliding. *PS: Political Science & Politics 57*(2), 149–161.

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.

McCauley, J. F. and D. N. Posner (2015). African borders as sources of natural experiments promise and pitfalls. *Political Science Research and Methods 3*(2), 409–418.

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.

Miller, M. K. (2024). How little and meng's objective approach fails in democracies. *PS: Political Science & Politics 57*(2), 202–207.

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.

Pearl, J. (2009). *Causality*. Cambridge university press.

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. (2000). *Democracy and development: Political institutions and well-being in the world, 1950-1990*. Number 3. Cambridge University Press.

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.

Treisman, D. (2024). Psychological biases and democratic anxiety: A comment on little and meng (2023). *PS: Political Science & Politics 57*(2), 194–197.

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.

Weitzel, D., J. Gerring, D. Pemstein, and S.-E. Skaaning (2024). Measuring backsliding with observables: Observable-to-subjective score mapping. *PS: Political Science & Politics 57*(2), 216–223.

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.