QMMR
Qualitative and Multi-Method Research
Fall 2024 | Volume 22.2

Inside:

ISSN: 2153-6767

QMMR Qualitative & Multi-Method Research

Fall 2024 | Volume 22.2

Table of Contents

QMMR Qualitative & Multi-Method Research

Spring 2024 | Volume 22.1

QMMR Qualitative & Multi-Method Research

Letter from the President

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062766

As a teacher of methods, I tend to emphasize what our methodologies *cannot* do in our efforts to learn things in political science. We must take methodological weaknesses as a given—there is no perfect or ideal approach to any question, because all of our methodological tools are sustained by assumptions that tend to be difficult to demonstrate or "prove" in practice. We must also acknowledge that we, as social scientists, are non-neutral observers of the world. We make choices about our data (Should I code this unit as a 1 or a 0? Did her "no" really mean "yes"?) that inevitably shape what we can know and how we know it.

Certainly, these are not novel claims. These are not *my* arguments. Interpretivists have long acknowledged the (inter)subjectivity of knowledge creation. Many of our innovations and advancements in methodology are driven by our desire to shore up different assumptions. Nevertheless, I find that many of my students, and especially my undergraduate students, end up slightly shocked by the tenuousness of what (I think) we can know in the social sciences.

To me, however, an acknowledgement that our approaches to knowledge-building are inherently limited is incredibly freeing. It means that we—or, at least, I—can observe the world and interact with human beings via different mediums (experiments, surveys, focus groups, interviews) without overly worrying about whether the methodological choices I have made are the "right" ones. In saying this, I do not mean to suggest that we should not be rigorous and systematic in our engagement with the world. There are better and worse ways to undertake all methods, and we should always strive to use any method in the best way possible.

Nevertheless, I reject the idea that there is only one correct or "right" way to do political science—that some gold standard exists that can allow us to get one step closer to the "truth" about the social world and human behavior in general. When we pursue a gold standard—the next, "best" way to model the world—we can lose sight of the world's complexity and messiness. Models are one way to interpolate the world, but we should never commit the mistake of thinking that the model is the world.

I say all of this, because I want to dedicate this short opening letter to the importance and indeed the necessity and utility of poetic license for how we study the world around us. There is quite a bit of "art" in our pursuit of (political, social) science. Rather than mitigate this fact, we should embrace it. We should allow ourselves the freedom to deviate from convention and expectation. The "poetry" that accompanies our scientific endeavors is fruitful and even generative. It fosters creativity and, consequently, drives methodological innovation.

To me, this issue of QMMR, like many that have come before it, reflects the creativity and innovation that many scholars—and especially young scholars—are bringing to the study and use of qualitative and mixed methods. We know that novelty is vital for the longevity of any ecosystem (Lugo, et al. 2018), and so the work of these emerging scholars and methodologists is especially valuable.

Jennifer Cyr
Universidad Torcuato Di Tella

References

Lugo, A. E., K. M. Winchell, and T. A. Carlo. "Novelty in ecosystems." *The encyclopedia of the Anthropocene 3* (2018): 259-271.

QMMR Qualitative & Multi-Method Research

Symposium: Draw Your Argument! Causal Figures in Political Science

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062770

"Draw Your Argument!": Introduction to the Symposium

Ezequiel Gonzalez-Ocantos
University of Oxford

Juan Masullo
Leiden University

"Draw your argument!" Many of us have been told to do this by supervisors and reviewers and have even returned the favor, urging our students and colleagues to do the same. But how exactly should we draw arguments to maximize theoretical clarity and the productivity of subsequent empirical work? While causal chains are central to the work of researchers in the qualitative and mixed methods community, there is surprisingly little guidance on the standards one should comply with when presenting theoretical arguments in graphical form.

We invited five leading methodologists and philosophers of social science to start a conversation about standards for causal graphs in Political Science. We asked them to read "Legacies of the Third Reich: Concentration Camps and Out-Group Intolerance," an excellent paper by Jonathan Homola, Miguel Pereira, and Margit Tavits (2020) published in the *American Political Science Review*, and write an essay explaining how they would draw its core argument. The piece is phenomenal: both theoretically rich and methodologically sophisticated, it triggered important debates in the discipline and contributed to a growing literature on the legacies of repressive institutions. Because the article doesn't offer a graphic summary of its complex theory of historical persistence and associated mechanisms of attitudinal reproduction, it was perfect for our exercise.

All our contributors to the symposium share the view that drawing an argument is key for the success of a research project. Gary Goertz goes one step further, declaring that "a scholar does not have a clear mechanism until it is presented as a figure" (in this symposium, page 24). In addition, all five are of the opinion that it is crucial to disaggregate process theories into constituent parts and engage in piecemeal testing; that is, one part at a time. This also requires a causal graph.

Despite areas of agreement, there are some key differences. In fact, the five graphs look pretty different! What our contributors think should be included and the conventions they think we should establish to represent relationships, differ dramatically. For example, both David Waldner and Rosa Runhardt argue in favor of the Directed Acyclic Graph (DAG) approach; Goertz and (especially) Derek Beach & Rasmus Brun Pedersen disagree. Similarly, while Waldner, Runhardt, and Goertz think of mechanisms as constituent parts of figures, Beach and Pedersen see the figure itself as the mechanism. Unlike Waldner and Goertz, who

focus primarily on offering guidelines for creating these figures, the contributions from Runhardt and from Beach and Pedersen engage with the testing implications of their approach to graphs. In Runhardt's case, she illustrates how to apply the counterfactual theory of causation she thinks lies at the heart of process tracing methods. In line with their previous writings, Beach and Pedersen reject the centrality of counterfactual reasoning in process tracing. Instead, they highlight the practical advantages of using graphs to derive observable implications and collect data accordingly.

Drawing on his book manuscript *Qualitative Causal Inference and Explanation*, Waldner adopts a Directed Acyclic Graph (DAG) approach, emphasizing that even the simplest graphs can provide valuable insights into the nature of hypothesized causal relations. He argues that causal graphs can enhance our research in three ways: heuristically, communicatively, and methodologically. He also outlines three criteria for valid unit-level causal inference—front door, back door, and side door—that should guide our graphing strategies.

Waldner's essay stands out for two central claims, both of which are rooted in his view that variables, mediators and mechanisms are ontologically distinct; he conceptualizes mechanisms as invariant properties that are triggered when variables assume specific values. First is the claim that graphs should only include variables deemed causally relevant to the outcome or subject matter at hand. The second is that graphs should clearly differentiate between variables (represented as nodes) and mechanisms, which are encapsulated in the arrows linking the nodes.

Runhardt's point of departure is "interventionism;" a counterfactual theory of causation she has developed elsewhere. Her essay explores the practical implications of this approach for how we conceptualize and visually represent causal arguments. Before diving into graphical representations, she emphasizes the importance of translating analyses into counterfactual terms. This encourages researchers to clarify causal claims by breaking down the different elements within a causal chain and assessing the presence and significance of each counterfactually. This process is essential for establishing "productive continuity," which refers to the unbroken flow of causal influence from a putative cause to its outcome. In this regard, her approach aligns with Waldner's emphasis on researchers adhering to the "completeness standard" when assessing causal claims via process tracing.

Runhardt's recommendations for constructing causal graphs, however, differ from Waldner's. She argues that graphs should explicitly include both variables and mechanisms, rather than subsuming the latter into unspecified arrows linking nodes. Unlike Waldner, she does not view variables and mechanisms as having different ontological status. This divergence highlights how differing perspectives on the components of a causal chain can significantly impact the representation of theoretical arguments and the structure of causal graphs.

Goertz offers a contrasting perspective by explicitly moving away from the DAG approach. Instead, he introduces the concept of "causal model mechanism figures," which, as the name suggests, places a stronger emphasis on mechanisms. Unlike Waldner and Runhardt, Goertz is skeptical that simply using a DAG focused on the proposed argument is sufficient to clarify the mechanisms at play. He sets a higher bar: figures should not only include causally relevant forces but also explicitly reference alternative explanations. As he puts it, "claims that something is NOT a cause should be included" (Goertz, this symposium, page 24) Given these rigorous demands, his figure looks markedly different from those in Waldner and Runhardt. For those interested in adopting this approach, Goertz provides detailed guidance on what researchers should and should not include, ensuring a more comprehensive representation of causal relations.

Despite important differences between the first three approaches, they share more common ground with each other than they do with Beach and Pedersen's perspective. Beach and Pedersen offer a truly distinct approach to constructing causal graphs, emphasizing a more granular breakdown of mechanisms. Drawing on their extensive contributions to the literature on process tracing, they argue that a high level of granularity is essential to fully leverage the benefits of the method. They

focus precisely on unpacking Waldner's arrows, highlighting that "arrows alone do not explain the nature of the process that actually links one node with another" (Beach and Pedersen, this symposium, page 32). They advocate for specifying arrows in terms of actors and activities. These actors and their actions generate linkages between nodes and should, therefore, be prominently represented in causal graphs. Unlike the previous three contributions, where mechanisms appear as elements within a causal graph (either as arrows in Waldner's framework or as diamonds in Goertz's), Beach and Pedersen argue that the mechanism is what should be explicitly unpacked and visualized in the graph.

Considering these differences, Waldner suggests that "as a research community, we would benefit if we settled on a single standard for how to represent causal claims graphically" (this symposium, page 8). This is precisely the thought that prompted us to organize the symposium in the first place. The essays that follow, however, suggest that reaching consensus on common guidelines will be challenging, if not impossible. An alternative approach might be for researchers to be more transparent about the decisions they make when representing their arguments graphically, clarifying how their approach to graphs aligns with their analytical goals, specific approach to process tracing, and core epistemological assumptions. We hope this symposium provides a valuable framework for making, implementing, and justifying these decisions.

References

Homola, Jonathan, Miguel M. Pereira, and Margit Tavits. 2020. "Legacies of the Third Reich: Concentration Camps and Out-Group Intolerance." *American Political Science Review* 114, no. 2 (May): 573–90.

Waldner, David. 2024. *Qualitative Causal Inference and Explanation. Manuscript: University of Virginia, Charlottesville.*

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062779

Inference and Causal Graphs: Potential Lessons from Qualitative Causal Inference

David Waldner
University of Virginia

Introduction

Directed acyclic graphs —causal graphs, for short— are increasingly recognized as a valuable element of causal inference. While they are primarily used for modeling relations of statistical dependence in quantitative approaches to inference, my own project over the past few years has been to adapt causal graphs to improve qualitative causal inference, a project culminating in my book manuscript, *Qualitative Causal Inference & Explanation* (Waldner 2024).[1]

Causal graphs improve our research in three ways: heuristically (thinking clearly about causal relations), communicatively, and methodologically. Consider the primitive causal graph in Figure 1, below, which we can interpret either as the encapsulation of accumulated knowledge about causal relations or as a posited claim about causal relations that guides future empirical work.

Figure 1. A primitive causal graph

$$X \rightarrow Y$$

Figure 1 has two nodes representing random variables and one directed edge or arrow representing a relationship of statistical dependence of Y on X given the claim that X causes Y. Verbal descriptions of causal relations are often connotatively ambiguous, as scholars substitute a variety of transitive verbs for the simple character of an arrow. But the figure does more than state the claim that X causes Y: it also states that Y does not cause X, that there are no pre-treatment causes of both X and Y, where these pre-treatment common causes are conventionally denoted Z, there is no prior cause of X that would act as an instrument assuming the satisfaction of an exclusion restriction, there are no mediators along the path from X to Y, conventionally denoted M, and there are no other causes of Y that are orthogonal to X — in this essay, these other causes are denoted by W.[2]

A simple graph communicates a remarkable amount of information, and it does so unambiguously, given mutual knowledge about the semantic and syntactical rules of causal graphs. Compare this unambiguous and transparent form of communication to the sometimes tortured language or idiosyncratic diagrams we frequently encounter in published works.[3] The heuristic value of the graph is that to draw the graph, we are forced to think hard about a large inventory of potential

1 For some of the earlier versions, see Waldner (2015a,b).

2 Orthogonal means that elements of W are statistically unrelated to X. Any variable related to X would have to be either a pre-treatment variable, which would be included then in Z and any back-door path that stemmed from it, or post-treatment consequences of X, which would be members of M .

3 For a limited sense of the heterogeneous ways that causal diagrams can be used, see two of the other essays in this symposium, with only the essay by Rosa Runhardt adopting a graph-theoretic perspective using DAGs. The diagrams in the essay by Beach and Pedersen are designed to depict different levels of granularity, with higher levels of granularity used to justify a causal interpretation. Gary Goertz uses what he calls a "causal mechanism figure," which he declares is distinct from a DAG, primarily because his figures exclude "those factors which are neither specific nor substantive" (this symposium, page 24). I am not denying here the potential value of either approach; my point, rather, is that we as a research community would benefit if we settled on a single standard for how to represent causal claims graphically.

causal relations—to draw the graph above, we have to be able to convince ourselves and others that, for example, the vectors *Z*, *M*, and *W* are all empty sets. Finally, we can draw clear methodological guidelines from the graph; if we are convinced by the figure, then a simple difference-of-means test will be sufficient to consider our sample-based estimate to be an unbiased estimator of the relevant estimand. If, on the other hand, *Z* were not an empty set, we could draw on the principles of what is called "d-separation" to devise an empirical strategy aimed at satisfying conditional independence by blocking the backdoor path.[4]

We live in a world of too many variables and limited opportunities to disentangle their inter-relationships to infer the causal effect of one variable, the treatment, on another, the response or outcome variable. One valuable communicative and heuristic tool afforded by the graph-theoretic approach is to divide all variables other than the treatment and outcome variables into one of three types: *Z* variables that are *pre-treatment* common causes of both *X* and *Y*, *M* variables — mediators in graph-theoretic terminology (and, as discussed below, conceptually distinct from causal mechanisms) — that are post-treatment consequences of treatment and that propagate causal influence along a front-door path terminating at *Y*, and *W* or other causes of *Y* that are post-treatment and orthogonal to *X*. The existence of these three types of variables implies three potential criteria for valid, unit-level causal inference: a back-door criterion which, when satisfied, eliminates the threat of systematic bias from pre-treatment common causes, a front-door criterion, which, when satisfied, eliminates the threat of systematic bias due to endogeneity or spuriousness, and a side-door criterion, which, when satisfied, allows us to use the pre-treatment value of *Y* as a proxy measure for the post-treatment, counterfactual value of *Y*.[5] Table 1 summarizes these three, path-specific causal criteria.

Table 1. Three criteria of valid, unit-level causal inference

Path	Criterion	Purpose
Front-door	Identify full set of mediators	Check for endogeneity and spuriousness
Back-door	Identify assignment mechanisms and potential back-door paths	Check for confounding bias
Side-door	Identify alternative causes that are orthogonal to *X*	Check for measurement error when using pre-treatment value of *Y* as counterfactual value of *Y* post-treatment

Contributors to this symposium have been asked to illustrate the value of causal graph by using them to interpret and evaluate an article by Jonathan Homola, Miguel M. Pereira, and Margit Tavits (2020) that discusses possible legacies of the Nazi domination of Europe. I consider this to be exemplary research into a substantively and normatively important topic. If causal graphs can make even an incremental improvement to this article, we can easily extrapolate their value to other research projects. In particular, because I have borrowed so heavily from quantitative approaches to causal inference to improve qualitative methods of inference, I hope to show that using graphs to think about making unit-level inferences using qualitative methods may provide some benefit to quantitative scholars.

Homola, Pereira, and Tavits' verbal description of their argument is, in my opinion, clear and (with a few minor exceptions discussed below), unambiguous. It is also, however, less than ideally compact, with different elements of the argument discussed at different stages of the exposition. It should be interesting, then, to see if my reconstruction of the argument as a causal graph corresponds to the authors' understanding; and if that reconstruction provides any communicative,

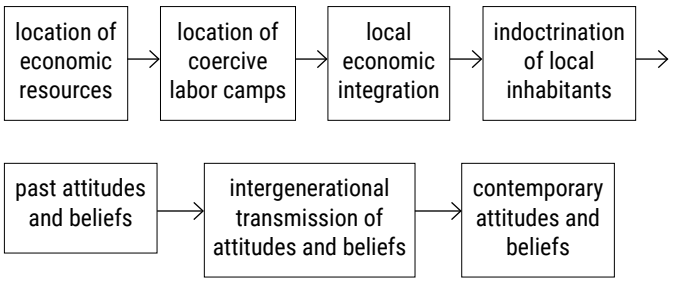
4 For discussion, see Pearl (2000).

5 Unit-level causal inference is generally considered impossible due to the fundamental problem of inference: for a single unit, we cannot simultaneously observe the outcome under treatment and control (Holland 1986).

heuristic, or methodological benefits. Figure 2 represents my interpretation of that argument. As we read Figure 2 sequentially from left to right, keep in mind that causal graphs provide information explicitly, by the inclusion of certain nodes and arrows, and implicitly, by the exclusion of other nodes and arrows.

Let's now interpret Figure 2 by first assigning each node to a type of variable, especially by distinguishing *Z*, *M*, and *W*, and then by considering each of the three criteria briefly introduced above. In "Legacies of the Third Reich," the treatment is the node "location of coercive labor camps." The front-door path thus contains, by my reconstruction, four mediating variables, terminating at the outcome node, "contemporary attitudes and beliefs." Figure 2, however, begins with a prior variable, "location of economic resources"; this is the assignment mechanism that assigns units to treatment (i.e., determines the spatial location).

Figure 2. Implied causal graph of "Legacies of the Third Reich"



As denoted in the graph, the initial node assigns units to values of the treatment, the node "location of coercive labor camps," but the figure does not contain a backdoor path (a path from "location of economic resources" to "contemporary attitudes and beliefs" that does not pass through the treatment); the only path that connects the initial and terminal nodes runs through the treatment. Put differently, units are, according to the graph, assigned exogenously to values of the treatment. Importantly, the graph omits a node labelled "pre-treatment attitudes and beliefs." Were the exclusion of this node not credible, then the entire analysis would be undermined by the high likelihood of endogeneity, with past attitudes and beliefs assigning units to treatment. In short, we can treat the causal graph as beginning with an ideal intervention, *I*, such that *Z* is an empty set—there are no pre-treatment common causes of both treatment and outcome.[6] In that case, I consider Figure 2 to satisfy a backdoor criterion of causal inference.

Figure 2 contains four mediators, the four nodes between "location of economic resources" and "contemporary attitudes and beliefs." I would like to propose that this enumeration of the members of *M* satisfies a front-door criterion; these four mediating nodes are sufficient to establish causal continuity between the treatment and the outcome. A front-door criterion is obviously distinct from a backdoor criterion, and the two criteria are conventionally treated as potential substitutes from one another. In addition to redefining the front-door criterion in terms of causal continuity (in quantitative approaches to causal inference, analysis of the front-door path involves separately estimating the causal effect of *M* on *Y* and the causal effect of *X* on *M*, then weighting these two estimates to arrive at an unbiased estimate of the causal effect of *X* on *Y*), I treat the front-door criterion as a critical complement to the backdoor criterion; given the constraints, using observational data, on ascertaining with confidence that the exclusion restriction is validly satisfied (that is to say, there is no path from the initial to the terminal node that does not pass through treatment), there is no reason to not make every effort to satisfy both criteria and thus guard against multiple sources of systematic bias.

Think of the front-door criterion as executing a series of hypothetical interventions, each of which is roughly analogous to Judea Pearl's "do-operator" (2000). The criterion of causal continuity obligates

6 An ideal intervention sets the value of the treatment variable, leaving all variables not on the front-door path between *X* and *Y* unchanged. Random assignment is one form of such an ideal intervention.

us to provide reasons such that if we were to intervene on each node to set it to its observed value—not, it must be emphasized, a counterfactual value—the consequence of that intervention would be the next node in the causal graph. If a series of such hypothetical interventions transmit causal influence from treatment to outcome, then we have satisfied the front-door criterion.

Homola, Pereira, and Tavits (2020) do not explicitly invoke a front-door criterion and do not explicitly discuss these hypothetical interventions. They expend considerable creativity and energy to make descriptive inferences about those mediators that are present and those that can be credibly omitted from the front-door path. But it is relatively easy to see, from Figure 2, how these hypothetical interventions could be constructed and justified and how they would satisfy a front-door criterion.

That said, there are a couple of controversies surrounding the front-door criterion that need to be addressed. First, I would propose a conceptual distinction between mechanisms and mediators. In graph theory, mechanisms are nodes in a causal graph that form a path between *X* and *Y*, with one arrow entering from the left and one arrow exiting to the right. It has become quasi-conventional to use the term "mechanism" as synonymous with mediator. There is nothing wrong with this usage, as long as the meaning is clear. For reasons I explain at length in *Qualitative Causal Inference & Explanation*, however, I prefer to use the term mediator to describe random variables and the term mechanism or causal mechanism to refer to invariant properties of entities that transmit causal influence between random variables. Conceptually, the mechanisms are the invariant properties represented by the arrows in the causal graph.

To see why the distinction between mediators and mechanisms matters, note that Homola, Pereira, and Tavits explicitly invoke two factors both described as mechanisms. While they consider cognitive dissonance to be the main causal mechanism "triggered by proximity to the camps during the Third Reich," they also consider the inter-generational transmission of beliefs to be "the second part of our theoretical mechanism" (2020, 583). Yet note that Figure 2 contains a node labelled "intergenerational transmission of attitudes and beliefs" but does not contain a node corresponding to cognitive dissonance. This is because these two elements of the causal graph operate differently. Intergenerational transmission of attitudes and beliefs is a variable that, by virtue of its location in the graph, is a mediating variable. We can easily treat it as a random variable taking on different values with a stochastic com- ponent, and we could easily imagine hypothetical (but probably unethical!) interventions to change the value of the variable. Cognitive dissonance operates at a different level; it is a structural feature of human cognition, not a variable on which we might intervene to set it to a new value. Specifically, cognitive dissonance is represented by the arrow between "indoctrination of local inhabitants" and "past attitudes and beliefs." Thus, Figure 2 does not claim that people living near coercive institutions exhibit cognitive dissonance while those living further away do not; cognitive dissonance is not a variable in that sense. Rather, cognitive dissonance is a constitutive element of how humans process information, where information differs based upon spatial proximity to camps. We should, I would argue, reserve the term "mechanism" for this type of property of a causal system.

The second controversy is whether we should invest any intellectual energy into constructing and testing the front-door path. From a Bayesian perspective, evidence gathered along the front-door path is not likely to provide the grounds to adjudicate between rival hypotheses and thus amounts to wasted effort.[7] But from a causal inference perspective, the front-door criterion represents an important check against either endogeneity or spurious association, especially in a world of observational data in which claims about the absence of backdoor paths will always be fallible and thus subject to revision. Furthermore, uncovering the elements of the front-door path

7 For skepticism about the inferential value of a front-door criterion such as the one I have defined here, see Fairfield and Charman (2022) as well as Humphreys and Jacobs (2023). I explicate the front-door criterion and discuss these Bayesian criticisms of it in *Qualitative Causal Inference & Explanation*.

yields higher information content; it helps explain how the world works. It is for this reason that we can observe renewed interest in mediation analysis even among experimentalists.

To summarize, Homola, Pereira, and Tavits (2020) could more explicitly distinguish mechanisms and mediators, and they could make a more explicit claim about how their analysis of the front-door path satisfies the criterion of causal completeness and thus complements their claim that treatment has been exogenously assigned. These are, perhaps, only incremental refinements of already exemplary research; but in a world in which inferring causal relations is fundamentally difficult, even modest refinements should be eagerly adopted.

Finally, let's consider the side-door criterion. Figure 2 deliberately omits a path rep- resenting the vector *W* of post-treatment causes of *Y* that are orthogonal (hence the term "side-door path") to *X*. Existing graph theory does not, to the best of my knowledge, contain any explicit discussion of side-door paths and a side-door criterion; any causes of *Y* that are orthogonal to *X* are, by definition, not potential sources of biased causal inferences; they are no more than sources of noise. Qualitative methods, on the other hand, place great emphasis on so-called "alternative explanations" because the core inferential model of qualitative inference is to use available evidence, or "causal process observations," to determine the best explanation in terms of relative fit between theory and evidence. I have defined a side-door criterion to better synthesize qualitative methods with the tools of causal inference. Specifically, if we can credibly claim that *W* is an empty set, then contingent on the satisfaction of the back-door criterion, we can use the pre-treatment value of *Y* as a reasonable proxy measure of the counterfactual outcome had the unit been assigned to the untreated group. The side-door criterion is thus central to qualitative causal inference that must confront the fundamental problem of causal inference that would otherwise rule out the possibility of unit-level inference.[8]

Homola, Pereira, and Tavits (2020) creatively and rigorously consider alternative explanations in a section they call "alternative mechanisms." I would suggest two clarifying revisions. The first would be to distinguish mechanisms from mediators along the lines I have suggested, and the second would be to assign these alternatives to either *Z*, *M*, or *W* to clarify what is at stake at this stage of the analysis.

Homola Pereira, and Tavits (2020) first consider various ways in which pre-treatment economic conditions could have influenced both the location of coercive institutions and contemporary attitudes. Perhaps camps were located in economically distressed areas that could have fostered political intolerance independent of the treatment effect of location-induced indoctrination; or, alternatively, camps generated an economic boost that persisted into the post- war era, producing patterns of contemporary attitudes and beliefs. From a graph-theoretic perspective, these random variables are potential members of *Z*, or pre-treatment common causes of *X* and *Y*: their importance is that they could potentially introduce systematic bias along a back-door path. They are not alternative explanations per se, but rather potential confounding variables.

A second alternative considered by Homola, Pereira, and Tavits (2020) is the possibility that memorials, documentation centers, and museums erected on the sites of former coercive institutions might generate contemporary cognitive dissonance. As I interpret it, this mechanism is in fact a potential mediating variable, *M*. It is obviously not a pre-treatment variable, but neither is it orthogonal to *X*; instead, it lies along the front-door path between *X* and *Y* and thus relates to claims about the front-door criterion.

Only their third alternative explanation represents a potential member of *W* that could, conceivably, generate a side-door path. Homola, Pereira, and Tavits (2020) note that in the post-war period —hence, post-treatment —the massive relocation of ethnic Germans into West Germany offers an alternative explanation that is entirely unrelated

8 Given space constraints, I can only direct interested readers to *Qualitative Causal Inference & Explanation* for further discussion of the side-door criterion.

to the treatment. If new migrants both settled disproportionately in areas near the former camps and if they carried with them distinctive attitudes, then the contemporary association of location and attitudes may be spurious after all. Note that if this claim were credible, then attitudes would have changed post-treatment, and we could not use the pre-treatment value of the outcome variable as a proxy measure of the unobservable counterfactual outcome. All available evidence, however, suggests that the location of the camps was not associated with post-war migration patterns. Therefore, based on available evidence and theorizing, it is reasonable to conclude that *W* is an empty set, and the side-door criterion has been satisfied.

From this graph-centric interpretation of their analysis, we can I think see why the side-door criterion should be viewed as a valuable addition to our inferential methods. Especially when working with observational data, our inferences are always provisional, tentative, and subject to future revision. It seems methodologically prudent to hedge our epistemic wagers by seeking to satisfy, to the greatest extent possible, both a front-door criterion and a backdoor- criterion. But the example of post-war migration demonstrates forcefully why the side-door criterion is important as well; were it the case that migrants settled disproportionately in areas proximate to former camps, and if they carried specific attitudes with them, then we would have reason to doubt that the causal effect along the front-door path made a substantial difference. Counterfactually, in the absence of treatment, a high level of migrant settlement would have produced, at least approximately, the observed outcome. Conversely, if the side-door criterion has been satisfied alongside of the other two criteria, then we are in a methodologically justified position to claim that there are no post-treatment causes of *Y* other than *X* and therefore the pre-treatment value of *Y* is a viable and defensible proxy measure for the counterfactual value of *Y* for the treated units. The three criteria thus neatly complement one another, which should be methodologically reassuring given our reliance on observational data.

References

Fairfield, Tasha, and Andrew E. Charman. 2022. *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*. Cambridge: Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, no. 396 (December): 945–60.

Homola, Jonathan, Miguel M. Pereira, and Margit Tavits. 2020. "Legacies of the Third Reich: Concentration Camps and Out-Group Intolerance." *American Political Science Review* 114, no. 2 (May): 573–90.

Humphreys, Macartan, and Alan Jacobs. 2023. *Integrated Inferences: Causal Models for Qualitative and Mixed-Method Research*. Cambridge: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Waldner, David. 2015a. "Process Tracing and Qualitative Causal Inference." *Security Studies* 24 (2): 239–50.

Waldner, David. 2015b. "What Makes Process Tracing Good? Causal Mechanisms, Causal Inference, and the Completeness Standard in Comparative Politics." In *Process Tracing: From Metaphor to Analytic Tool*, edited by Andrew Bennett & Jeffrey T. Checkel, 126-52. Cambridge: Cambridge University Press.

Waldner, David. 2024. *Qualitative Causal Inference & Explanation*. Manuscript, University of Virginia, Charlottsville.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062788

An Interventionist Potential Outcomes Framework for *Legacies of the Third Reich*

Rosa W. Runhardt
Radboud University, Nijmegen, The Netherlands

Introduction

This commentary uses interventionism, a counterfactual theory of causation from philosophy of science, to evaluate the causal analysis in the article "Legacies of the Third Reich: Concentration Camps and Out-Group Intolerance" (Homola, Pereira, and Tavits 2020; hereafter "Legacies"). The section 'An Interventionist Potential Outcomes Framework' consists of a condensed presentation of the interventionist theory, while the section 'Translating "Legacies of the Third Reich"' re-expresses the causal claims from "Legacies" in interventionist terms and presents an accompanying causal graph. The section 'Three Stages of Interventionist Causal Analysis' then highlights one of the steps in the causal graph to illustrate the key stages of interventionist causal analysis in qualitative and mixed-method research. In short, researchers are urged to: (1) translate their analysis in counterfactual terms, in order to disambiguate their causal claims; (2) describe a possible intervention which would bring the counterfactual scenario about; and (3) collect evidence for what would happen in this scenario. The conclusion summarizes.

An Interventionist Potential Outcomes Framework

Interventionism is a philosophical theory of causation which evaluates a given causal claim $X \rightarrow Y$ by assessing whether intervening on has an effect on X (Woodward 2003a, 2007). If appropriate changes in X are associated with changes in Y, and some specific technical requirements on the intervention and other variables connected to X and Y are met, the relation is deemed causal. Interventionists represent causal relations, including the processes which connect a putative cause and effect of interest, using causal graphs.[1]

The interventionist literature on the social sciences often focuses on how one may infer causal relations from large-N data sets. For instance, we may ask what causal claims are supported by experimental data from a randomized controlled trial. When interventionism is applied to large-N observational analysis, it takes on a *counterfactual* nature. It asks what *would* have happened to Y, *had* X taken on a different value. In that situation, interventionism becomes a type of potential outcomes approach. As Stephan Morgan and Christopher Winship (2015, 444) put it in their own methodological handbook on potential outcomes modelling, "[f]or observational data analysis (…) we see no way to escape having to assert what-if assumptions about potential outcomes in order to move forward." Yet interventionism also has concrete consequences for how to infer causal relationships in qualitative and mixed-method research, including process-tracing (Runhardt

1 As such, interventionism has much in common with other structural causal modelling theories (cf. Halpern 2016; Morgan and Winship 2015; Pearl 2000; Spirtes, Glymour, and Scheines 1993; Weinberger 2019). To start unpacking the differences between them, see for example Hitchcock (2009).

2015, 2022a). The remainder of this commentary is based on the epistemological assumption that in such research, counterfactual reasoning can be used to infer whether some causal mechanism actually produced an effect of interest.[2]

In the simplest case, an interventionist would formalize a causal mechanism in the following way (cf. Runhardt 2015, 2022a). Imagine mechanism Z is hypothesized to explain the observed link between some putative cause, X, and observed effect, Y. Approaches like process tracing require one to look for the observable implications of Z, which in the simplest case can be formalized as a chain[3] of variables Z_i, $i \in \mathbb{N}$ such that $X \rightarrow Z_1 \rightarrow Z_2 \rightarrow \ldots \rightarrow Y$ and where $Z_i \rightarrow Z_{i+1}$ denotes that Z_i causes Z_{i+1}. Given this simple representation, interventionism requires one to evaluate each step of the chain in turn, providing evidence of what would happen to a Z_{i+1} under an intervention on the preceding Z_i, all other factors in the network being equal. Corroborating an intervention claim takes place in three stages: (1) use a well-crafted counterfactual scenario to help describe the causal "step" under investigation, that is, what is meant by Z_i and Z_{i+1}, as well as the wider network of other potentially related factors; (2) choose an appropriate intervention for constructing this counterfactual scenario; and (3) collect factual evidence of what would happen under this hypothetical intervention.

The next section uses interventionism to analyse the causal analysis in "Legacies" (Homola, Pereira, and Tavits 2020). The section shows that interventionism supports their analysis. In fact, this study proves an instructive example of how *counterfactual* claims can be corroborated with *factual* evidence.[4] It shows that such evidence is multifaceted: It can come from various sources, such as observation, experiments, cross-case analysis, archival data, interviews, and even established theories from other disciplines, all of which jointly contribute to raising our degree of belief in the counterfactual claim of what would happen under intervention on a mechanism's steps.

Translating "Legacies of the Third Reich"

"Legacies" shows that current-day outgroup intolerance, immigrant resentment, and far-right support in Germany is partially explained by how close to Nazi-era labour camps a given German lives. Homola, Pereira, and Tavits (2020) argue that a two-stage causal mechanism is behind this relation: as camps were integrated into the local economy in the Nazi era, communicating a deeply intolerant belief system, local individuals developed (e.g., out-group intolerant beliefs as a cognitive dissonance mechanism, and then transmitted their intolerance to later generations). The result is that "current-day Germans who live closer to Nazi-era concentration camps are *more* xenophobic, *less* tolerant of out-groups— including Jews, Muslims, and immigrants— and *more* likely to support extreme right-wing parties" (Homola, Pereira, and Tavits 2020, 574; emphasis in original).

The argument in "Legacies" itself takes place in two key steps. Firstly, the authors exclude possible confounders for the correlation between distance and out-group intolerance, immigrant resentment, and far-right support. This raises their confidence that the correlations are in fact causal relations. For example, the authors show that the location of the camps was not determined by the degree to which a local community was sympathetic towards

2 This separates my account both from those who are concerned about interventionism as an epistemology for causal modelling (cf. Reutlinger 2012; Russo 2011) and from those that conceptualize mechanisms as systems of entities and activities, in emulation of a mechanistic philosophy popular in the life sciences (see Beach and Pedersen, this symposium, and Beach and Pedersen, 2019; see also Machamer, Darden, and Craver 2000 for the initial philosophical account). My approach is more in line with David Waldner's hope, expressed elsewhere in this symposium, that "using graphs to think about making unit-level inferences using qualitative methods may provide some benefit to quantitative scholars" (Waldner, this symposium, page 10).

3 Many methodologists and philosophers distinguish underlying causal mechanisms from the observable processes these mechanisms produce (cf. Andersen 2014a; 2014b). I will return to this matter.

4 As such, it answers an important criticism levied against counterfactual analysis by amongst others the systems approach to mechanisms (see footnote 2), namely that such an analysis is fundamentally at odds with collecting concrete evidence *because* of the counterfactual nature of the claims.

the Nazi regime ($A \nrightarrow D$; i.e., pre-existing mass political attitudes in the 1930s, *A*, are not a cause of distance to labour camps, *D*). This was a potential confounder because had it been the case that $A \rightarrow D$, and given that those pre-existing attitudes might themselves be a cause of (amongst others) current-day outgroup intolerance, *OI* (i.e., since it may be that $A \rightarrow OI$), this could have explained the correlation between *D* and *OI* (*A* would have been a common cause).

In the second key step, the authors present evidence for the cognitive dissonance (*CD*) and intergenerational transmission (*IT*) mechanisms. They also exclude alternative mechanisms that may explain the relation between distance to the closest camp and out-group intolerance, immigrant resentment, and far-right support, thus raising their confidence that the relevant process connecting e.g., *D* and *OI* is $D \rightarrow CD \rightarrow IT \rightarrow OI$. They show, for example, that current-day economic inequalities do not mediate between *D* and *OI*, because $D \nrightarrow E$.

To illustrate, consider the causal graph in Figure 1.[5] Here, *D* is the distance to the closest labour camp, *CD* is a cognitive dissonance mechanism triggered by proximity to the camps, *IT* is intergenerational transmission of beliefs, and outcome variables *OI*, *IR* and *FR* are out-group intolerance, immigrant resentment, and far-right support, respectively. Arrows (directed edges) represent that the claim that there is a causal relationship between two variables (nodes). This representation makes no distinction between a mechanism and a putative cause.[6]

Figure 1. A simplified causal graph for Homola, Pereira, and Tavits (2020)

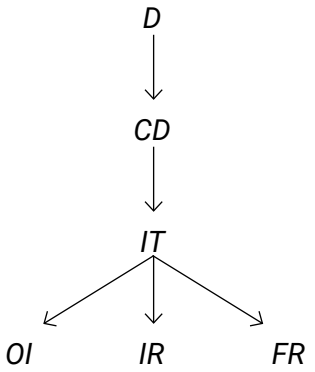


Figure 1 is a simplified graph for the key claims of the article. Not pictured are the confounders and alternative mechanisms which Homola, Pereira and Tavits exclude. This is defensible, since the factors in those confounders and mechanisms are not related to the variables that are pictured in the graph[7]. However, the graph is incomplete in other, indefensible ways: it does not yet include certain variables that *are* causally connected to the various measures of intolerance. The authors discuss two key causes that must be pictured: the current-day local economic situation, here denoted with *E*, and the current-day share of immigrants in the local district, here denoted with *IS*. Both are causally related to out-group intolerance, immigrant resentment, and far-right support. Thus, the full picture "Legacies" defends is that in Figure 2.[8]

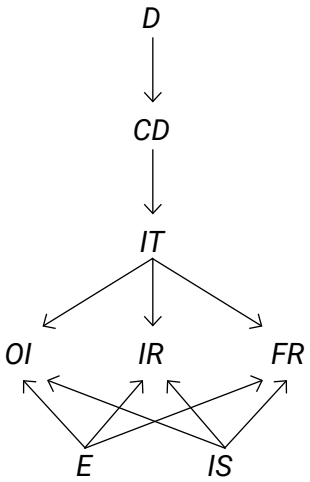
5 Unlike Gary Goertz's contribution elsewhere in this symposium, but like Waldner's contribution, my design of this figure is largely guided by the same principles one uses to construct a directed acyclic graph (DAG). See footnote 15.

6 My philosophical argument for this is that the mechanisms do not have a different ontological status than variables in "Legacies." One could argue that the cognitive dissonance mechanism is a family resemblance term that captures a number of different ways in which psychological tensions are resolved, just like out-group intolerance is a family resemblance term that captures a number of different attitudes and behaviours (see footnote 15). Both factors are considered in their totality despite the nuances one could make at a lower level of analysis. For a similar philosophical view on causal mechanisms see Little (2011).

7 In my approach, the graph is meant to represent the asserted causal relationships for a given field of inquiry only. This fits with Waldner's use of such graphs (this symposium) as comprehensive descriptions of all causal relations. Including unrelated factors as standalone nodes without edges connecting them to the other nodes would be somewhat unorthodox for a DAG, but could serve pedagogical purposes. I am therefore sympathetic to Goertz's decision (this symposium) to include such factors with a slashed arrow indicating the absence of a causal relationship.

8 While Waldner (this symposium) and I agree that one should include only those variables causally relevant to the subject under investigation, I do not include his variable "location of economic resources" here. I have instead decided to start the graph from *D* itself, since I take this to fit with the focus of Homola, Pereira, and Tavits (2020). Moreover, starting from *D* is justified here because, as Waldner himself points out, location of economic resources is not itself connected to other variables in the causal graph (such as *OI, IR* and *FR*).

Figure 2. A complete causal graph for Homola, Pereira, and Tavits (2020)



So far, the graphs pictured are merely helpful representations of the relations found in "Legacies." An interventionist interpretation of them requires inferring from the "Legacies" argumentation and data whether any of the arrows in Figure 2 are genuinely causal (whether they ought to be included in a proper causal graph), in the way described in the previous section. For each arrow $X \rightarrow Y$ in the graph, one ought to design a hypothetical intervention I which changes the value of putative cause X, keeping all other variables in the diagram fixed, and researchers should evaluate whether this (hypothetical) change would affect outcome of interest Y, given certain technical requirements are met. This will be the subject of the next section.

Three Stages of Interventionist Causal Analysis

As we have seen the introduction of the interventionist framework, this foundational theory requires that researchers: (1) translate their analysis in counterfactual terms, in order to disambiguate their causal claims; (2) describe a possible intervention, in order to contrast their claim with others in the literature; and (3) to collect evidence for this intervention, in order to strengthen their conclusions. This section will connect these steps with the causal graph defended in Figure 2. For brevity, we will focus on only one of the relations discussed by the authors, viz., that between distance to the closest labour camp, , and the degree of outgroup intolerance, *OI*.[9]

Step 1. Counterfactual Translation

The counterfactual we are concerned with for the $D \rightarrow OI$ relation is not simply the claim "if the distance to the closest labour camp had been greater, the degree of outgroup intolerance would have been lower." Rather, as argued above, interventionism prescribes that we break the process up into smaller steps and construct associated counterfactuals for each one (keeping all else fixed).[10]

As we have seen in the previous section, one step in the process leading to current-day intolerant beliefs is $D \rightarrow CD$, i.e., the relation that claims that "the likelihood that someone experienced dissonance is higher closer to camps than elsewhere" (Homola, Pereira, and Tavits 2020, 575, f5).[11] As discussed, the associated counterfactual for this step helps us understand what is actually at stake. For one, such a counterfactual scenario needs to be specified at the right level of generality. While the article discusses just one country, the $D \rightarrow CD$ step is nevertheless an average treatment effect claim. It averages out over all individuals, and as such does not tell us (as Homola, Pereira, and Tavits put it) "that everyone close to the camps

9 *OI* is a latent variable the authors constructed using results on the 2008 European Values Survey; for more details on all definitions of the variables, see (Homola, Pereira, and Tavits 2020, 580 and Supplementary Information 3.1).

10 Note that Waldner and I disagree on the use of counterfactual language here. A full technical discussion is beyond the scope of this article; to find out more about the relationship between his (and Judea Pearl's) "do-operator" and my (and Jim Woodward's) use of hypothetical interventions; I can recommend Woodward (2003b), especially his discussion on pages 324-325.

11 Compare this interpretation to Waldner's decision (this symposium) to treat *CD* as an arrow, rather than a node in the graph. I would argue including *CD* as a node in the graph, as I have here, does greater justice to claims by Homola, Pereira, and Tavits like the one cited here. What varies is the *likelihood* of *CD*.

necessarily experienced dissonance" (2020, 575, f5).[12] What is argued, then, is that had the *average distance* been higher, cognitive dissonance would have occurred less frequently as well. It does not make any claims about individual Germans in the Nazi era.[13]

The graph as displayed in Figure 2 is also "higher-level" in a different sense: It specifies the mechanism as just two intermediate steps in the chain between *D* and *OI*, while it could be broken down further (say, by distinguishing intergenerational transmission from parents to children from that between children and grandchildren, or parental transmission from peer transmission). This is in line with the argument in "Legacies" itself.[14]

Step 2. Describing the Intervention

Now consider the next step of the *D* → *OI* relation, the relation between cognitive dissonance and intergenerational transmission (*CD* → *IT*). What intervention might lead to the counterfactual scenario for this step, that is, what intervention could affect the cognitive dissonance of Germans in a way that accords with the demands described by interventionists? We ought to find an intervention variable that meets the following four criteria (similar to Runhardt 2022a, 23; 2015, 1305; cf. Woodward 2003a, 98):

1. *I* should cause the type of cognitive dissonance described by Homola et al., *CD*.
2. *I* should function as a switch for *CD*, that is, make the occurrence of dissonance independent of any other variables in the diagram.
3. *I* should not itself be causally connected to intergenerational transmission of intolerance, *IT*: in particular, *I* should not by itself lead to or prevent *IT*.
4. *I* should be statistically independent of all *other* potential causes of *IT* (besides those related to *CD*); it may be, for example, that the degree to which an individual is conservative itself increases or decreases how often they discuss politics with their children (as well as what they discuss), which would make conservatism a poor intervention variable.

Step 3. Evidence for the Counterfactuals

Despite not concretely specifying a hypothetical intervention that meets the requirements from the preceding subsection, arguably we can provide a rational reconstruction of Homola, Pereira and Tavits' work in those terms. The last step required in the interventionist potential outcomes framework is that Homola, Pereira and Tavits (2020) provide evidence for the existence of an intervention variable that meets the four criteria described in the preceding.

As suggested in Runhardt (2022a), there are few limits on what this evidence should look like. Different sources may corroborate a counterfactual claim of what would happen under intervention, nor is any single piece of evidence sufficient to confirm such a claim. Rather, multiple pieces of evidence will collectively strengthen the causal inference. Despite not specifying a particular *I*, the authors *do* attempt to provide such corroborating evidence for what would happen if an intervention on *CD* were made, that is, what would have happened to intergenerational transmission of intolerance had there been less cognitive dissonance.

12 Contrast this to the decision by Goertz as well as Derek Beach and Rasmus Brun Pedersen in this symposium to focus their analysis (and thus, their causal graphs) at the level of individual Germans.

13 This distinction harkens back to the methodological distinction between case-level and population level causal claims as well as the philosophical distinction between token and type causation. See Runhardt (2022b) for further discussion of levels of generality in mixed-method research.

14 Contrast this to Beach and Pedersen's call, this volume, for unpacking the causal process at a much lower level of abstraction. A further discussion of when it is or is not warranted to break down a process (or mechanism) into even further steps is beyond the scope of this commentary, in part because it requires going into *ontological* details on the distinction between mechanisms and processes (see footnote 3 and 6). For this commentary, the most important question is whether a factor in the graph is "manipulable" (at least in theory). Homola, Pereira, and Tavits (2020) seem to think that the presence or absence of *IT* as a whole (Nazi era to current-day) *can* be considered in its totality.

To find out what evidence is relevant, it is first necessary to make the counterfactual claim precise. For example, the claim here is *not* that parents who are more likely to experience cognitive dissonance are also more likely to discuss politics with their children. Rather, the claim is that had these parents not experienced cognitive dissonance, they would not have instilled as strong an intolerance in their children. This is a more specific claim than showing that closeness to the camps led to more intolerant attitudes at the time; it is about *how* this happened. Arguably, an important part of evidencing this counterfactual is to show that in this counterfactual scenario, it is likely that parents would have more tolerant attitudes. Homola, Pereira, and Tavits (2020) make this likely amongst others by appealing to a wide literature which they believe shows that cognitive dissonance has a real effect on (prejudicial) attitudes. They cite such evidence as the theoretical literature on the historical transmission of intolerance in Germany, theoretical political science on intergenerational legacy effects, and empirical evidence of the effects of political socialization (whether individuals discussed political beliefs with their parents).

It is worth stressing again that the interventionist potential outcomes framework at bottom is not heavily prescriptive about what constitutes good evidence: Rather, it tells us the direction in which we must look (counterfactuals) and gives us the tools to explore this direction (hypothetical interventions). How the interventionist evidence is collected is, in the end, heavily context dependent. And in fact, the plurality of interventionist evidence one can and should use is illustrated well by the article.[15]

Besides appealing to theory, the authors also evidence the counterfactual by varying the cognitive dissonance levels theoretically; after all, they gather statistical evidence on "Germans who resided near concentration camps [and who therefore] had to rationalize a more extreme example of intolerance than other Germans" (Homola, Pereira, and Tavits 2020, 575). However, to use distance from the camps as the intervention variable is only acceptable if it meets the four criteria specified described in Step 2.[16] For instance, the authors evidence this through the statistical analysis, including by showing that distance to the camps was not determined by the degree to which a local community was already sympathetic towards the Nazi regime. Distance is therefore statistically independent of this other potential cause of *IT*, strengthening our faith in criterion 4. One ought to show this for all other potential causes of *IT* in the network.

Conclusion

This commentary gave a condensed presentation of the interventionist theory in order to re-express the causal claims from "Legacies" in interventionist terms. It presented an appropriate causal graph for the two-stage causal mechanism the authors use to explain the link between distance to the nearest Nazi-era concentration camp on the one hand and current-day outgroup intolerance, immigrant resentment, and right-wing support in Germany on the other. One of the steps of this mechanism, between cognitive dissonance and intergenerational transmission, was highlighted to illustrate the key stages of interventionist causal analysis. This illustration showed how a counterfactual causal analysis of mechanisms can be rooted in multifaceted factual evidence despite its counterfactual nature.

15 There is room, within this framework, for the evidence described by Beach and Pedersen in their contribution to this symposium, such as traces of an individual's experience of discomfort in, say, individual diaries. On the other hand, my arguments here require that the search for such evidence is structured rather differently, viz., as evidence of what would happen under an intervention.

16 One may be critical of using *D* as an intervention variable, as it is included in the network. Arguably, one difficulty for Homola et al. from an interventionist point of view, is that there is no real evidence in the paper for a counterfactual scenario in which some Germans (a) reside close to a concentration camp; yet (b) do not experience cognitive dissonance. Specifically, we do not have a view of whether those who were tolerant to begin with will not transmit these attitudes because of another mechanism in the absence of dissonance. There may be a different mechanism that leads them to do so.

References

Andersen, Holly. 2014a. "A Field Guide to Mechanisms: Part I." *Philosophy Compass* 9, no. 4 (April): 274–83.

———. 2014b. "A Field Guide to Mechanisms: Part II." *Philosophy Compass* 9, no. 4 (April): 284–93.

Beach, Derek, and Rasmus Brun Pedersen. 2019. *Process-Tracing Methods: Foundations and Guidelines*. 2nd ed. Ann Arbor: University of Michigan Press.

Halpern, Joseph Y. 2016. *Actual Causality*. Cambridge: MIT Press.

Hitchcock, Christopher. 2009. "Causal Modelling." In *The Oxford Handbook of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Peter Menzies, 299–314. Oxford: Oxford University Press.

Homola, Jonathan, Miguel M. Pereira, and Margit Tavits. 2020. "Legacies of the Third Reich: Concentration Camps and Out-Group Intolerance." *American Political Science Review* 114, no. 2 (May): 573–90.

Little, Daniel. 2011. "Causal Mechanisms in the Social Realm." In *Causality in the Sciences*, edited by Phyllis McKay Illari, Federica Russo, and Jon Williamson, 273-95. Oxford: Oxford University Press.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67, no. 1 (March): 1–25.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. New York: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Reutlinger, Alexander. 2012. "Getting Rid of Interventions." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43, no. 4 (December): 787–95.

Runhardt, Rosa W. 2015. "Evidence for Causal Mechanisms in Social Science: Recommendations from Woodward's Manipulability Theory of Causation." *Philosophy of Science* 82, no. 5 (December): 1296–307.

———. 2022a. "Concrete Counterfactual Tests for Process Tracing: Defending an Interventionist Potential Outcomes Framework." *Sociological Methods and Research* 53, no. 4 (November): 1591-628. https://doi.org/10.1177/00491241221134523.

———. 2022b. "Limits to Evidential Pluralism: Multi-Method Large-N Qualitative Analysis and the Primacy of Mechanistic Studies." *Synthese* 200 (171).

Russo, Federica. 2011. "Correlational Data, Causal Hypotheses, and Validity." *Journal for General Philosophy of Science* 42, no. 1 (May): 85–107.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag.

Weinberger, Naftali. 2019. "Path-Specific Effects." *British Journal for the Philosophy of Science* 70, no. 1 (March): 53–76.

Woodward, James. 2003a. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

———. 2003b. "Critical Notice: Causality by Judea Pearl." *Economics and Philosophy* 19: 321-40.

———. 2007. "Causal Models in the Social Sciences." In *Philosophy of Anthropology and Sociology*, edited by Stephen P. Turner and Mark W. Risjord, 157–210. Amsterdam: Elsevier.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062792

Theorizing and Diagramming Causal Models-Mechanisms: Comment on Homola, Pereira, and Tavits (2020)

Gary Goertz
University of Notre Dame

> Mechanism-process accounts, in contrast, positively welcome history, because their explanatory program couples a search for mechanisms of very general scope with arguments that initial conditions, sequences, and combinations of mechanisms concatenate into processes having explicable but variable overall outcomes.
> **Charles Tilly**

A Personal and Philosophical Prelude[1]

I would like to thank the editors for inviting me to participate in the symposium. In this introductory prelude I provide some philosophical positions on the role of causal mechanisms in causal inference as well as the issue of the role of historical factors, often called legacies, in explaining social phenomena. This prelude pro- vides some background, very brief, before getting to the business at hand which is to draw a causal mechanism or causal model figure of Homola, Pereira, and Tavits (2020), which is the topic of the next section. Because of space constraints I make stark statements with little justification. To provide reasonable justification would push the essay beyond space limits.

This symposium offers me an occasion to go back to some long standing interests. In one of my first books *Contexts of International Politics* (1994) I devoted several chapters to a discussion and analysis of what I called historical factors in statistical models. Coming from the background of quantitative international con- flict studies I was curious about the fact that in all or virtually all of these studies the dependent variable was measured at time t and virtually all if not all of the independent variables were measured at time $t - 1$. So I was quite curious about causal and methodological issues for historical variables measured at time $t - n$. So the Homola, Pereira, and Tavits article (2020) gives me a chance to go back to these con- cerns because they are arguing that something that happened in the 1930-40s Germany influences what is going on in contemporary Germany. So the $t - n$ is here something like $t - 70$.

An interesting set of theories involves the importance of something that hap- pened in the past and its influence on the present. To give this a name one might call it the legacy-persistence literature. A classic example is Stinchcombe's (1968) "historicist" theories. Other examples include the huge literature on path dependence which typically argues that understanding the present requires an analysis of what happened at critical junctures (Pierson 2004). Another example is the importance of colonial legacies on current economic and political factors in the global South (e.g., Acemoglu et al. 2001). Finally, there has been a huge surge in legacy-persistence studies in economics, historical political socially, and political science (for a review see Abad and Maurer 2021). Homola, Pereira, and Tavits (2020) certainly belongs in this literature.

A second long-standing interest is in causal mechanisms. Mechanism talk has become very prevalent in the social sciences. In some ways

1 Thanks to Charlotte Cavaillé, Steph Haggard, Moritz Marbach, Rosa Rundhardt and the editors for comments on earlier drafts.

I have the impression that the term "mechanism" has replaced "theory" in many substantive settings. I was particularly interested in mechanisms in working on my book on multimethod research (Goertz 2017). In my systematic literature review it was clear that people did case studies because they wanted to explore mechanisms in addition to their statistical analysis.

This has become even more important in my new project with Stephan Hag- gard on Large-N Qualitative Analysis (LNQA) (2023) where causal mechanisms and with-in case causal inference play a central role. The causal model-mechanism figure is critical because it provides a template for process tracing and within-case analysis which is where the heavy causal inference lifting occurs.

This is all related to a basic philosophical position on causal inference. Basically it is the position that one cannot do causal inference without serious empirical evidence regarding the mechanism that links cause to affect. In philosophy this is most well known as the Russo-Williamson thesis (Williamson 2019) which requires both correlational as well as mechanism evidence to conclude that there is a causal relationship. To say the least, this is very controversial among experimenters and others who think that randomized controlled trials, natural experiments, instrumental variables, etc., are sufficient to conclude causation.

While working on causal mechanisms for the 2017 book I became convinced that for practical as well as methodological reasons that a scholar does not have a clear mechanism until it is presented as a figure. If one looks at the philosophical literature on causation and mechanisms it draws very heavily on medicine, neu- rology, biochemistry and the like (e.g., Machamer, Darden, and Cramer 2000). All of this very heavily relies on the figures that are used to express these biochemical mechanisms. This means I am happy to engage in a dialogue about diagramming causal mechanisms in the Homola, Pereira, and Tavits (2020) research project. It does mean that I am going to be skeptical about any causal inference that is purely based on statistical analysis without a clear causal mechanism and without evidence for such a mechanism. Of course, I understand that to publish an article one is going to be hard-pressed to deal with just statistical inference issues within space constraints. As a practical matter the authors are going to be pushed to spend more time on statistical inference issues than on causal mechanism ones. Hence the symposium is an opportunity to think more explicitly about the causal mechanisms involved without being forced to spend a lot of space on statistical estimation issues.

A Causal Mechanism Figure for Homola, Pereira, and Tavits

Since the usages and meanings of causal mechanisms vary so widely a few comments about the principles guiding my construction of a figure for Homola, Pereira, and Tavits (2020) are in order.

One might conclude that the demand for a causal mechanism figure would be satisfied by a DAG (see Waldner's contribution in this symposium for a DAG-based approach). As we shall see, this is not the case. This is related to the general principle that only specific, substantive causal mechanism variables will be included in the figure. The point of this essay is to do a causal figure, not to discuss at lengths the pro's and con's of DAGs. Nevertheless, where it seems relevant contrasts with DAGs are included.

It is perhaps useful to start out with what will *not* be included in the figure which falls under the rubric of those factors which are neither specific nor substantive. For example, it is quite common to include instrumental variables in a DAG. But clearly their purpose is to estimate statistically the causal impact of the core causal variable. So they are not relevant to the causal mechanism leading from the core causal variable to the the outcome.

More ambiguous are fixed effects. These are virtually always justified in terms of capturing causal heterogeneity. However, they do not specify specific substan- tive factors in the fixed effects that are causing outcome. This is relevant to the Homola, Pereira, and Tavits (2020) mechanism because in their critique, Pepinsky, Goodman, and Ziller (2023) argue that if one controls for fixed effects by Länder the statistical effects disappear. If one looks at the statistical equation being estimated – $Y_{is} = \beta \text{Distance}_{is} + \gamma_{is} X + \phi_s + \in_{is}$ these fixed effects, ϕ_s, are essentially intercept

terms. These are not of substantive interest to the authors: "Fortunately, because those factors *are not themselves of theoretical interest* but potentially confound the empirical relationship being studied, we can control for them using fixed effects. Länder fixed effects adjust for any factor (observable or not) that varies across German Länder and explains out-group intolerance" (2; emphasis mine).

Even more clearly in the gray zone are other control or causal variables, e.g., $\gamma_{is}X$ in the equation above. Depending on the causal orientation, one woman's control variable is another man's central causal variable. This is going to then de- pend a lot on the theoretical and substantive context of the work in question. So it could be that a particular control variable is really an important alternative ex- planation in which case it should be included in the mechanism figure. However for the purposes of discussion of Homola, Pereira, and Tavits (2020), none of the other causal variables seem to be of particular interest to the parties involved, in fact they are generically referred to as "interwar covariates" and "contemporary mediators," signaling little interest in them as causal factors in their own right. Hence these covariates are not of substantive interest to me in developing a causal mechanism figure, so I exclude them from my causal mechanism figure.

Other causal variables appear in some of the contributions to the symposium. For example, Waldner's "sidedoor" factors are just other causal variables (e.g., "alternative post-treatment causes"; *W* in his notation, this symposium).[2] Some of these might well be included in a DAG of Homola, Pereira, and Tavits. Rundardt (this symposium) includes additional factors *E*, current-day local economic situation, and *IS*, share of immigrants, in her figure. These affect the outcome variables but are separate from the Homoloa, Pereira, and Tavits (2020) mechanism. I do not include these causal factors in my figure because they are not part of Homola, Pereira, and Tavits causal mechanism.

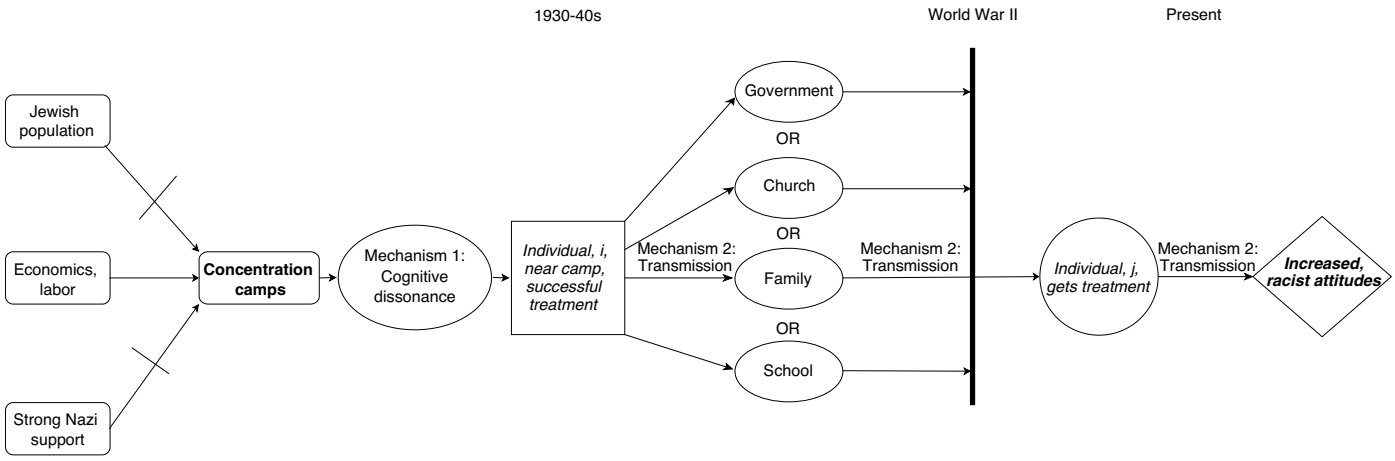
Perhaps the key difference between a causal mechanism figure and a DAG in this setting is that the causal mechanism figure works at the level of the individual. It includes the key factors that influence, and how they influence, the individuals in question and how causation is transmitted over time from individuals i at time $t - 70$ to different individuals j at time t .

It is perhaps best to start the causal mechanism figure from the left-hand side and work our way to the right hand side of the figure. While not always the case I think it is clear to be explicit that this left-right orientation is a temporal one *and* a causal ones. Depending on the DAG this is not always clear. Waldner, for example, says that Ware post-treatment so in my system they would appear to the right of the treatment variable, but before the dependent variable. In Rundardt, her other causal variables appear to the right of the outcome variables, which also does not conform to my system. Homola, Pereira, and Tavits (2020) distinguish between "interwar covariates" (i.e., pre-treatment) and "contemporary mediators," which in my sys- tem would be stacked vertically at the same time as the treatment. In addition, in my framework all variables to the left of the treatment are called "antecedent" variables. This includes confounders and endogeneity concerns.

This allows me to introduce an important principle for causal mechanism figures: important claims that something is not a cause should be explicitly included in the figure. For example, in DAGs if a variable is not considered a cause then it is not included (Waldner notes this in his contribution). So by definition if it is not in the DAG it is not an important causal variable. If one is explicitly arguing against a given causal claim that is theoretically important in the literature then it should be in the figure because that negative causal claim is a critical feature of the causal model. For Homola, Pereira, and Tavits (2020), this is most clear in claims about the endogeneity of the treatment. Since these are critical they should be in the figure. More generally,

2 Waldner requires that these be orthogonal (i.e., independent) of the other causes. This is possible in designed experiments, but in observational data this is almost never the case: there are almost always causal relationships among these other causal factors. For example, the correlations between *X*, the treatment, and other causal factors are usually not zero which they should be if *W* were independent from *X*.

Figure 1. Causal mechanism: concentration camp 1930-40s leads to intolerant attitudes 2015



this would be also meaning including central alternative explanations in the figure. This I do not do: (1) because they do not figure significantly in Homola, Pereira, and Tavits, and (2) it would encumber the figure even more.

Central to the Homola, Pereira, and Tavits setup is the claim that there are certain factors which are exogenous to the treatment. So these arguments must appear in the figure.

There are two causal claims that they stress are not related to treatment: (1) presence of significant Jewish population and (2) degree of Nazi ideology in the area: "The fact that (a) the site selection was mostly driven by economic rather than socio-demographic or attitudinal reasons, and (b) the camp location was exogenous to the Jewish population and Nazi party support, makes Germany an attractive case for identifying the effect of camps on contemporary attitudes." (2020, 577). So in Figure 1 I include those causal claims by having an arrow with a slash through it to indicate that they are not a result of those two variables. On the other hand, they make a strong point that the reasons for for the location of the concentration were economic considerations, notably the need for labor in various economic sectors. So that deserves a positive arrow in the figure. So there are three antecedent (potential) causal factors of the treatment. The treatment itself is in bold to show that it is different from the potential selection factors.

The Homola, Pereira, and Tavits causal mechanism really has two important parts, really two mechanisms, each of which deserves significant discussion.

This is clear in the structure of the article where they have a first section called "Camps and cognitive dissonance" followed by another section called "The persistence of political attitudes." The first causal mechanism relates to what was going on in the 1930s and 40s in regions that had concentration camps. The mechanism question is how does the presence of camps produce intolerant attitudes in individuals in areas near camps?

It is important to keep in mind that the subject or unit of analysis is individuals. So this theory needs to be some kind of socio-psychological theory. The mechanism they focus on, cognitive dissonance, is clearly an individual-level psychological phenomenon. Their discussion in the article is basically one column long, which does not give them much space to develop the mechanism. It appears that they are adopting fairly standard cognitive dissonance theories (this is beyond my area of expertise so I cannot comment how they use cognitive dissonance theories). Beach and Pederson, in their contribution to this symposium, focus their Figure 1 on explicating this mechanism, and each box has "individual" in it, also stressing that is the fundamental unit of analysis. This is certainly moving in the right direction.

The key mechanism point is that:

> The authors must convince us that the cognitive dissonance mechanism generates the attitudes in the individuals, *i*, in the areas near the camps otherwise the mechanism stops in its tracks.

The mechanism which I think could be developed further is that there was extensive interaction between individuals in the area and prisoners in the concentration camp. How this works needs specification. Is it somehow just the presence of the camps? Is it because individuals interact with prisoners who are providing labor for them? Homola, Pereira, and Tavits (2020) suggest that these interactions produced more intolerant opinions. This mechanism needs significantly more development and justification. Ideally there would be enough to do a causal mechanism figure at the individual level since it is a core mechanism in the argument. In any case, to make the causal arrow clear in Figure 1, I have indicated that it is cognitive dissonance that is producing anti-Jewish attitudes in these regions.

For purposes of argument let us assume that this is been done. We now move to the second core mechanism which is how these attitudes are transmitted over time to individuals in the "present" in their section "The Persistence of Political Attitudes." For short, we can call that the transmission mechanism.

This is perhaps the huge challenge of what might be called the legacy literature which is to explain how something which occurred in the perhaps distant past is still a cause of something happening in the present. It is not surprising that in the qualitative literature this is often called process tracing: tracing the process by which those attitudes in the 1930s and 40s are a cause of what is happening decades later.

Because the unit of analysis is the individual it seems like we are in the domain of the political psychology of political attitudes. What are the mechanisms by which political attitudes are formed that we can locate in the (distant) past. There are four generic ones that most of the literature focuses on. Most of these focus on institutions which continue to exist over time and hence provide a conveyor belt of attitudes over time. Generically we can call them family, education, church, and government. These are all potential mechanisms by which the effects in the 1930-40s can be transmitted over time to the present.

Because I think these are potential mechanisms I include them in the figure. Depending on the setting one or more than might be relevant in explaining why individuals *j* in the present have more intolerant attitudes.

This particular case is quite interesting in the sense that I think most of these mechanisms are excluded. The authors cite Acharya, Blackwell, and Sen's work on the US South (2018). Here I think it is quite clear that all the mechanisms were present: families, churches, schools, and government all actively maintained, promoted, and transmitted intolerant attitudes over time. In contrast, World War II presents a dramatic break in German history, and one can include the division of Germany after the war as part of the dramatic break in transmission mechanisms. The governments (East, West, and unified), churches, and schools actively rejected the Nazi heritage. Homola, Pereira, and Tavits seem to take this position: "In our case, the institution (i.e., concentration camp) was removed together with the Nazi regime and was not replaced with any alternative institution reinforcing out-group subjugation. This rules out the institutional channel" (2020, 575).

So the remaining mechanism is really transmission via families: "Rather, we argue that individuals transmitted their attitudes via *family ties and social interactions*, which led to the differences in out-group intolerance measurable even today" (2020, 576). Figure 1 includes family as a transmission mechanism since that relatively clear. "Social interactions" as mechanism could be included but it is not a clearly specified mechanism, and essentially not discussed at all.

One core question involves who is getting the treatment in 2010 and can we track that back to the pre-World War II era? If the treatment is transmitted only via family then that becomes much more problematic. So one needs to flesh out this transmission mechanism.

Critical in Homola, Pereira, and Tavits (2020) and other legacy studies is what one might call "treatment tracking" over time. This works at the level of individuals in Homola, Pereira, and Tavits who get the treatment at time *t* − 70. The question then is how a successful treatment is transmitted over time to those who get the treatment at time *t*.

Assuming for purposes of argument that the treatment is passed down from parents who successfully got the treatment at time *t* − 70. Successfully treated individuals *i* each have children who get the treatment from their parents. For the transmission to work, these parental treatments must be successful across multiple generations to time *t*.

A related issue is one might call demographic sorting. There is movement of successfully treated individuals in the region to non-treated regions. The out-movement of successfully treated are not included in the statistical analysis, so that dilutes the estimated treatment effect.[3]

Treatment tracking involves following those individuals that were successfully treated that then successfully treat their children and so forth over at least a few generations. All children of the successfully treated get the treatment but for some or all of them the treatment is not successful, hence treatment transmission stops. An alternative mechanism that would be important to consider here would be those children who rebel and object to the treatment from their parents. This could be particularly critical in Germany where in the 1960s and 70s in particular there were massive movements in politics and society among young people rejecting these treatments. So treatment tracking asks the question about how many people in the region of the statistical analysis actually receive the treatment via family? Given population movement, failure of treatment transmission, and youth revolt against the treatment, my sense is that they are not many people in the region in the statistical analysis that are actually treated via the mechanism postulated. In short, perhaps only a quite small percentage of people in a region near the camps were successfully treated at time *t*.

In short there are two very large mechanism hurdles which need to be overcome to make the statistical results convincing to someone of the Russo-Williamson school of causal inference. The first mechanism is connecting the camps with at- titude change in the areas near camps and the second mechanism is how those successful treatments are transmitted via families to individuals living in the area decades later and causing increased intolerant attitudes.

There is much more that could and should be said about the actual design of causal model figures. Simple rectangles and arrows are not enough. If possible one should specify the nature of the causal relationships indicated by the arrow, for example, positive or negative, or necessary–sufficient. For example, mediator causal relationships are quite different from treatment or confonders and others. This should be made clear in the figure. There are issues of aggregation when there are multiple causal factors that are involved in the mechanism both over time as well as over the various independent variables. As indicated in Figure 1 there important claims that something is not a cause. Another example in my figure is that I signaled the core treatment and dependent variable using bold fonts. The two mechanisms are a complex set of causal relationships between the two.

The goal is to make the figure clear as possible. In my experience it is often necessary to do a lot of analysis of the text to figure out how to interpret

3 Marbach (2023) focuses on a similar kind of issue of what he calls "post-treatment sorting." The general set up is similar to the Homola, Pereira, and Tavits (2020) study in the sense that between the treatment (*t* − 70) and the outcome (*t*) individuals can move from treated to nontreated geographical regions. While there are a number of significant differences between his framing of the problem and mine, perhaps a key one is that in my framing the transmission mechanism requires that treatment is successful for each generation. So it is not only important that people are moving in and out of treated regions but that the treatment is *successful* across generations. We both come to a similar conclusion is that the treatment effect is going to be diluted. This is because there can be successful transmission across generations in nontreated regions once a successful treated moves to a nontreated region.

the figure. Text, equations and figures ideally work together. The more complex the theory —like here with multiple mechanisms stretching over decades— the more necessary a figure is to help the reader sort out the theory and connect it with empirical evidence about all the causal claims in the figure.

References

Abad Leticia Arroyo, and Noel Maurer. 2021. "History Never Really Says Goodbye: A Critical Review of the Persistence Literature." *Journal of Historical Political Economy* 1 (1): 31–68.

Acemoglu, Darren, Simon Johnson, and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91, no. 5 (December):1369–401.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. *Deep Roots: How Slavery Still Shapes Southern Politics*. Princeton: Princeton University Press.

Brady, Henry E, and David Collier (eds.). 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. New York: Rowman & Littlefield.

Goertz, Gary. 1994. *Contexts of International Politics*. Cambridge: Cambridge University Press.

———. 2017. *Multimethod Research, Causal Mechanisms, and Case Studies: An Integrated Approach*. Princeton: Princeton University Press.

Goertz, Gary, and Stephen Haggard. 2023. "Large-N Qualitative Analysis (LNQA): Causal Generalization in Case Study and Multimethod Research." *Perspective of Politics* 21:1211–39.

Homola, Jonathan, Miguel M. Pereira, and Margit Tavits. 2020. "Legacies of the Third Reich: Concentration Camps and Out-Group Intolerance." *American Political Science Review* 114, no. 2 (May): 573–90.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67, no. 1 (March): 1–25.

Marbach, Moritz. 2023. "Causal Effects, Migration, and Legacy Studies." *American Journal of Political Science* 68, no.4 (October): 1447-59.

Pepinksky, Thomas B., Sara Wallace Goodman, and Conrad Ziller. 2023. "Modeling Spatial Heterogeneity and Historical Persistence: Nazi Concentration Camps and Contemporary Intolerance." *American Political Science Review* 118, no. 1 (February): 519-28.

Pierson, Paul. 2004. *Politics in Time: History, Institutions, and Social Analysis*. Princeton: Princeton University Press.

Stinchcombe, Arthur L. 1968. *Constructing Social Theories*. New York: Harcourt, Brace & World.

Williamson, Jon. 2019. "Establishing Causal Claims in Medicine." *International Studies in the Philosophy of Science* 32 (1): 33–61.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062801

Using Disaggregated Process Theories to Supplement Regression-Based Analysis

Derek Beach
Aarhus University

Rasmus Brun Pedersen
Aarhus University

Introduction

In this contribution, we argue that analysts need to go beyond simple causal graphs in order to reap fully the analytical benefits of using within-case methods such as process tracing as a supplemental tool that can strengthen confidence in the causal nature of findings from quantitative regression-based analyses (e.g., Seawright 2016). Instead, we contend that analysts need to unpac the causal structure of a causal process in more granular detail in order to trace the process empirically, as well as explore the contextual conditions underlying a correlation found using regression-based cross-case analysis. We argue that using simple causal graphs that draw arrows between independent, intervening and dependent variables does not enable the researcher to engage in the type of careful, withn-case detective work that is required to critically assess whether a correlation is actually causal and how the linkages in-between actually work.

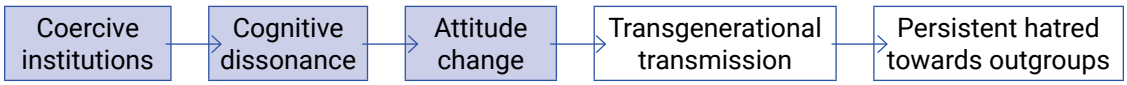
We illustrate the analytical benefits of working with more granular processual theories (i.e., at a lower level of theoretical abstraction) when using process tracing as a supplementary method alongside regression analyses. We use the example of the regression-based article "Legacies of the Third Reich: Concentration Camps and Out-Group Intolerance," focusing on the "cognitive dissonance" process (Homola, Pereira, and Tavits 2020). In the article, the authors demonstrate a strong correlation between higher levels of out-group intolerance in current (2008, 2016) citizen attitudes in areas close to Nazi-era concentration camps within Germany (Homola, Pereira, and Tavits 2020). We show the analytical value-added of engaging in using more granular theories through our identification of several serious issues related to the validity of claiming that the correlation is actually causal that are found through critical assessment of one part of the process (cognitive dissonance).

Using Process Tracing as a Supplemental Method to Regression-Based Analysis

Seawright (2016) points to a number of functions that process tracing can play in strengthening confidence that correlations found using regression analyses are actually causal. In this essay, we explore the analytical value of working with more granular process theories for two of the functions: 1) producing evidence of the causal pathway that is hypothesized to produce the causal effect identified in a regression analysis (Seawright 2016, 64), and 2) assessing the key assumptions behind the regression, focusing in particular on tracing the causes of a treatment to reduce the risks of confounders and missing scope conditions (Seawright 2016, 68-9).

At a very aggregate level, the overall causal pathway hypothesized by Homola, Pereira, and Tavits (2020, 574) can be depicted as a form of simple causal graph (Figure 1). Note that we focus on the *causal* elements of the explanation that the authors put forward (i.e., the "mechanisms," to use Goertz's terminology; this symposium). Not surprisingly, our simple graph is very similar to Goertz's mapping of the relationship. In contrast, other authors in this symposium develop causal graphs that focus on static elements that can vary, such as "the location of coercive labor camps" (see Waldner, this symposium).

Figure 1. Simple causal graph of how coercive institutions can produce persistence hatred, based on Homola, Pereira, and Tavits (2020, 574)



However, where we diverge from the other contributors to this symposium is in our argument for further disaggregating the causal explanation in order to deploy process tracing as a supplemental tool.[1] For space reasons, we focus in the following on the first part of the hypothesized causal process (aka causal mechanism) that links cognitive dissonance, theorized to link exposure to coercive institutions (i.e., an individual is exposed to violence against outgroups during the Nazi era) with attitude change (i.e., stronger outgroup hatred of individual than previously) (shaded grey in Figure 1). However, a fuller analysis would unpack each of the steps in the causal process.

Working with More Granular Process Theories at a Lower Level of Aggregation

Describing a chronological narrative of events in-between the occurrence of a cause and outcome is not a causal theory, and it risks conflating a merely *temporal* sequence with a *causal* process (Sayer 2000, 141). Additionally, drawing arrows between boxes as in a causal graph only denotes that there is some kind of causal relationship, but it does not explain the nature of the process that actually links one node to the other. Waldner (this symposium) states that the arrows are "..invariant properties of entities that transmit causal influence between random variables," but what these invariant properties are, and how causal influence is transmitted is not made clear.

To provide a *causal explanation*, we argue that a causal process theory needs to explain in more detail what is going on *within* the causal arrow(s). One way of doing this is by theoretically unpacking the actors and activities that provide the linkages in the process (Craver and Darden 2013; Beach and Pedersen 2019). In this respect, a causal explanation of a process goes beyond a causal graph because it explains *who* does *what,* and *why* actions are linked together instead of merely drawing an arrow.

Theorizing involves making explicit the actors (*who*) and the activities they perform (*what*), as well as hypothesizing why they are linked together in a causal sense (*why*). Activities are what actors do in a process; they are what binds them to other actors in a causal relationship. To be part of a *causal* process, the activities of one actor must trigger an action in response from other actors, or if at the individual psychological level, a response from the same actor to the preceding activity. As we will discuss below, unpacking the causal dynamics of a process is not done for purely theoretical reasons. Making theorized linkages more explicit in our process theory also forces the analyst to attempt to trace them empirically instead of merely assuming that some unobservable 'mechanism' binds them together.

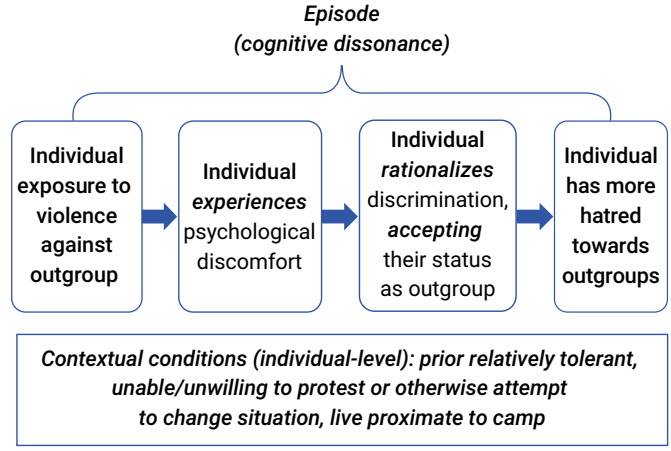
A good process theory should describe in more abstract terms the *causal structure* of the process, but in enough detail that it is possible to empirically assess whether the process actually worked as hypothesized. Causal structure can be thought of in terms of "episodes," understood as a set of interactions between actors—or within actors if processes are at the psychological, individual level) that has to take place for a process to move forwards, and where we have theoretical reasons to expect that the process might have played out differently if they had not occurred (Steel 2008, 88-92). Additionally,

1 Note that Goertz does suggest that the mechanisms could be "developed further" (this symposium, page 24). Goertz also suggests that scholars should "specify the nature of the causal relationships indicated by the arrow, for example, positive or negative, or necessary–sufficient" (PAGE NUMBER). We agree with this, but argue in this contribution that we should go even further by specifying the interactions between actors that provide the linkages.

context matters for how causal processes play out within cases (Falleti and Lynch 2009; Goertz 2017). By making the activities and linkages explicit, it is easier to detect relevant contextual conditions because they are the things that should be present for an activity to produce a particular response instead of other responses.

Figure 2 depicts a more granular version of the first episode in the overall causal process, with the theory at the individual-level because they are the unit of analysis at which the process operates.[2] The "cognitive dissonance" episode takes place at the individual level after initial exposure of a relatively tolerant person to violence against outgroups at camps in Nazi Germany (Homola, Pereira, and Tavits 2020, 575).

Figure 2. The "cognitive dissonance" episode



Note that while the authors do describe these distinct steps in the article, they then lump them together into a "cognitive dissonance" mechanism that they assume provides a causal linkage between exposure and more hatred in an individual. Given that we want to work with the more disaggregated version, we use the term "episode" because it is more compatible with the ambition to trace interactions—although note that here they are mental "conversations" *within* a given individual.

Theorizing activities and linkages in a more granular causal process theory, as undertaken above, have two advantages in relation to performing the supplemental functions of evidencing a causal pathway and exploring key assumptions behind a correlation reflecting a causal relationship.

Function 1: Evidencing a Causal Pathway That is Consistent with the Causal Effect

Assessing whether there is within-case evidence that supports the claim that a regression correlation is causal requires empirically tracing the causal steps in-between X and Y (Seawright 2016, 57). The first advantage of working with a more granular process theory is that they are easier to operationalize empirically because activities and linkages should leave some form of empirically observable traces in cases, at least in theory.

The "cognitive dissonance" process theorized by Homola, Pereira, and Tavits (2020) deals mostly with internal psychological processes, which makes operationalization much more difficult because we are dealing with reasoning within the minds of individuals instead of actions and responses between actors. This does not mean that the process disappears completely into a black box. Indeed, the psychological version of process tracing methods emerged out of an attempt to trace reasoning processes *within* the minds of individuals (see Schulte-Mecklenbeck, Kühnberger, and Ranyard 2011).

Typically, more indirect evidence can be used to evidence rationalizations. In the case in hand, creative historical detective work would enable the operationalization of observables that could make us more or less confident in the causal claim. In particular, given their extreme nature, internal mental activities of "experiencing psychological discomfort" and "rationalizing discrimination" should leave at least some form of empirical traces that could be detected either through testimony of interviews with individuals who experienced exposure in the years immediately after the events, or other traces such as private diaries or other forms of testimony. Ideas about what these types of potential observables might look like is included in Table 1.

2 Goertz's contribution clearly flags that the cognitive dissonance process (aka mechanism) works at the individual level, whereas the original argument by Homola, Pereira, and Tavits (2020) is at the group level.

Table 1. Potential empirical observables of the cognitive dissonance episode

Theoretical level	Exposure of individual to violence of outgroup	Individual experiences 'discomfort'	Individual rationalizes violence	Individual has more hatred of outgroup
Potential empirical observables	• Account of individual mentioning that they witnessed violence (diary or interviews) • Traces of pre-exposure views (voting behavior, diaries)	• Account of individual expressing emotional reaction (anger, shock, etc) in diaries, letters, interviews • Account of close family or peers that individual expressed reaction to them	• Account of individual expressing justifications for treatment (diaries, letters, interviews) • Account of close family or peers that individual rationalizes treatment to them	• Views expressed by individual (post-exposure) in diary, letters, interviews. • Comparison of pre/post-exposure attitudes

If the empirical analysis found confirming evidence in multiple case studies of individuals, we could have greater confidence that the causal effects found in the regression analysis were actually causal instead of being correlational (Seawright 2016, 64). Unfortunately, the article by Homola, Pereira, and Tavits (2020) relies solely on regression analyses, focusing solely on controls for other potential causes. The article could have been strengthened by providing some form of within-case evidence that supports the claim that the processes or mechanisms underlying the posited causal relationship actually were causal. Even relatively weak, indirect evidence of causal linkages is better than no evidence.

Function 2: Assessing the Assumptions Behind the Regression

The second advantage of using more granular process theories is that the focus on activities and context helps assess the key assumptions behind a regression result is produced by confounding factors (Seawright 2016, 69-9). This includes assessing the causes of the "treatment" (i.e., X) to detect whether there might be unknown confounders and whether there are missing contextual or scope conditions. In the example of "cognitive dissonance," for the process to be triggered the following conditions had to be present: The individual being exposed had relatively tolerant prior attitudes; the individual was actually exposed to violence and other forms of inhumane treatment of outgroups by witnessing events in person (i.e., proximity) over some period of time;[3] and arguably a scope condition that the individual was unable to protest or prevent the violence due to the repressive nature of the system, which then led them to have to attempt to rationalize the inhumane treatment instead of trying to do something to stop it.

Empirically assessing the conditions triggering causes and processes requires considerable case knowledge about the context within which theorized cause and process took place. In this instance, historical knowledge about the context and how the camp system evolved from 1934 until 1945 can help shed light on whether the found correlations between exposure and hatred in areas proximate to main camps could have actually been produced by the "cognitive dissonance" process.

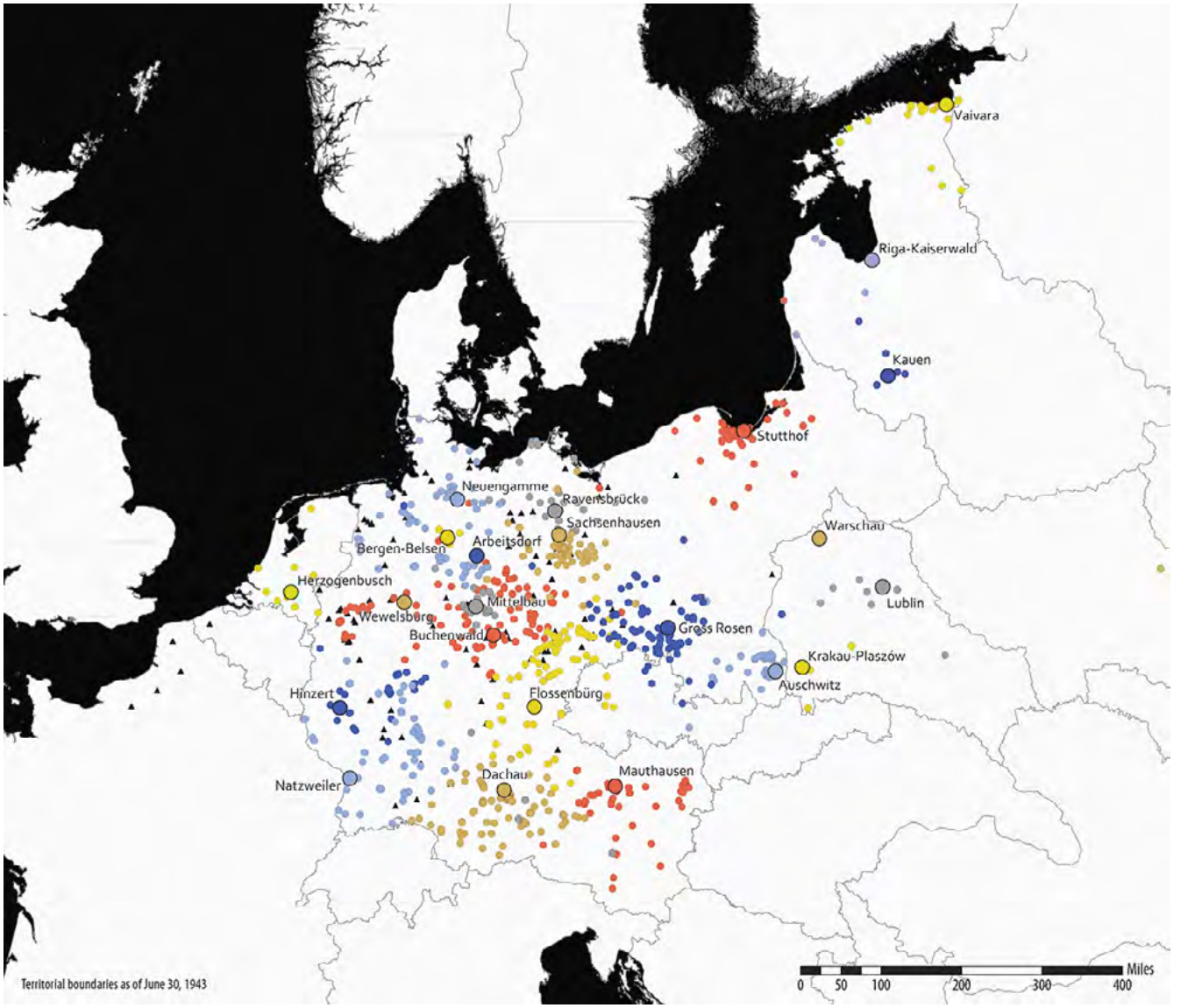
One particularly important historical factor was that the nature of the concentration camp system

3 The amount of time required for more hatred to be produced in any individual is unspecified in the article. This is important because while all of the main camps are treated similarly in the article, a small camp listed as a main camp in the article—Arbeitsdorf (ca 1000 prisoners)—was only operational from April to October 1942. In contrast, large camps like Dachau (200,000 prisoners) were operational from March 1933 to the end of the war. See Megargee (2009) and Knowles, Cole, and Giordano (2014) for more.

in Germany changed character in early 1942 (Knowles, Cole, and Giordano 2014, 41-3). Before 1942, concentration camps were used primarily to house political prisoners, Jews, homosexuals, and Roma, amongst others. It is in this context in which the "cognitive dissonance" process could conceivably functioned.

However, as the war against the Allies expanded from 1942, the camp system was transformed. Jews and other groups were sent eastwards to extermination camps, whereas camps *within* Germany became work camps filled with prisoners of war from the Soviet Union, Poland and other occupied territories. While most prisoners in the period prior to 1942 were housed in a handful of main camps, by late 1942 there was an extensive network of satellite camps extending through every region of Germany —often far from the main camp (see Figure 3, below). The size of satellite camps varied from small groups in tents within factories to large satellite camps with between 5,000 and 10,000 prisoners. Taken as a whole, there were also significantly more prisoners in the last years of the war than in the 1930s in both the main camps and especially their satellites (Knowles, Cole and Giordano, 2014: 41-43).

Figure 3. Map of main camps and satellite camps, 1943

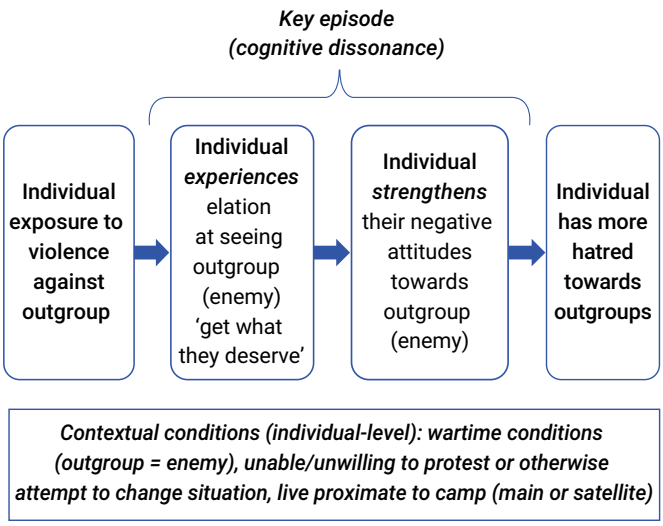


Source: Birkbeck College (http://www.camps.bbk.ac.uk/maps/location-of-main-camps.html), based on Knowles, Cole, Giordano (2014, 36). Large dots are main camps, whereas smaller dots are satellite camps. The triangles are other forms of imprisonments.

The shift in context has two implications in relation to the assumptions behind the correlations found using regression. First, the increase of geographic scope due to satellite camps greatly increased the potential exposure of individuals across all of Germany. However, the regression analysis used in the article only uses *the proximity to a main camp*, thereby missing the much greater geographic scope of potential exposure in the later period.

Second, given the expansion in the size of the prisoner population after 1941, most individuals exposed to violence would have witnessed atrocities against prisoners from countries at war with Germany; especially from the Soviet Union. In a context of massive war propaganda, bombings across Germany, and large numbers of Germans killed in war, it is not difficult to hypothesize that many individuals who were first exposed to violence against outgroups in the later years of the war would not experience the same type of psychological discomfort as their compatriots did in the 1930s. Instead, individuals in the later period might perceive the prisoners as "enemies" upon whom vengeance was being delivered. Therefore, exposure to violence against outgroups might still produce more hatred, but it would not be through "cognitive dissonance," but instead through a "vengeance mechanism." A potential pathway is depicted in Figure 4, below.

Figure 4. A hypothetical "vengeance mechanism," 1941-1945



Taken as a whole, this means that the changing context in the later period includes a range of potential missing confounders and contextual or scope conditions that potentially impinge on the validity of the found correlation between exposure and hatred.

Conclusions

Working with simple causal graphs that depict causal processes as a series of boxes with arrows drawn between them does not unpack theories in enough detail that they can be easily operationalized. We argued in this article that unpacking causal processes at a lower level of abstraction strengthens the ability of the analyst to use process tracing as a supplemental tool to critically assess the causal nature of correlations found using regression analysis. We briefly illustrated how it can help assess whether a hypothesized causal process actually took place, as well as helping shed more light on the conditions that potentially triggered the process. We found that it would have been possible through careful historical detective work to provide at least some indirect contemporaneous evidence to substantiate that the hypothesized cognitive dissonance mechanism was actually causal. Further, our findings related to the changing nature of the causes and conditions of "exposure of individuals" (X) from 1942 onwards raises questions about whether the cause and cognitive dissonance mechanism can actually account for most of the exposure and more hatred produced amongst Germans during the Nazi era.

References

Beach, D. and R.B. Pedersen. 2019. *Process-tracing – foundations and guidelines*. 2nd Edition. Ann Arbor: University of Michigan Press.

Craver, Carl F. and Lindley Darden. 2013. *In Search of Mechanisms*: Chicago: University of Chicago Press.

Falleti, Tulia G., and Julia F. Lynch. 2009. Context and Causal Mechanisms in Political Analysis. *Comparative Political Studies* 42 (9): 1143–66.

Goertz, Gary. 2017. *Multimethod research, causal mechanisms, and case studies: An integrated approach*. Princeton University Press.

Knowles, Anne Kelley, Tim Cole, and Alberto Giordano. 2014. *Geographies of the Holocaust*. Bloomington: Indiana University Press.

Megargee, Geoffrey P, ed. 2009. *The United States Holocaust Memorial Museum Encyclopedia of Camps and Ghettos, 1933-1945. Volume 1: Early Camps, Youth Camps, and Concentration Camps and Subcamps under the SS-Business Administration Main Office (WVHA)* Bloomington: Indiana University Press.

Sayer, A. 2000. *Realism and Social Science*. London: Sage Publications.

Schulte-Mecklenbeck, Michael, Anton Kühberger, and Rob Ranyard, eds. 2011. *A Handbook of Process Tracing Methods for Decision Research : A Critical Review and User's Guide*. New York: Psychology Press.

Seawright, Jason. 2016. *Multi-method Social Science*. Cambridge: Cambridge University Press.

Steel, Daniel. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

QMMR Qualitative & Multi-Method Research

Symposium: Emerging Methodologists Workshop

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062803

Emerging Methodologists Workshop: Introduction to the Symposium

Diana Kapiszewski
Georgetown University

Hillel David Soifer
University of California, Berkeley

Healthy social science disciplines require and encourage the continuous evolution and development of methods for conducting research. As the empirical world evolves, and as scholars study new problems, and bring new perspectives to bear on existing ones, they must innovate methodologically, or depend on the methodological innovations introduced by other scholars. Unfortunately, the groups of scholars who develop, write on, and teach research methods in our discipline continue to lack a key source of this innovation: scholars from under-represented groups (Achen 2014, Shames and Wise 2017, Barnes 2018). In part this lack of diversity results from a "pipeline problem" in which junior scholars from under-represented groups who are interested in and talented with methods are not encouraged and actively mentored to develop existing or new methods, publish methodological scholarship, or teach methods.

The annual "Emerging Methodologists Workshop-Qualitative and Multi-Method Explanatory Research" (hereafter EMW-QMER, http://sigla.georgetown.domains/emworkshop/) has contributed since 2023 to addressing this diversity deficit. In each one-day workshop, six advanced political science graduate students and junior faculty who are based at U.S. institutions, many from under-represented groups, present and receive feedback on a paper focusing on methods for gathering or analyzing qualitative data, and/or strategies for integrating qualitative and quantitative methods, that are aimed at explanation. Each presenter is paired with a "Methods Mentor" who works with and supports them in the months before and after the workshop, with the goal of workshop papers being submitted to and published in peer-reviewed journals.

The EMW-QMER's broader aims are to strengthen existing scholarly networks, promote new networks, and build an inclusive intellectual qualitative and multi-methods research community. Though hardly the first effort of its type (see, e.g., Dion 2014), other initiatives and related scholarship have focused more intently on the gender gap within the groups of scholars who work on research methods (though see Smith, Gillooly and Hardt 2022; and

Perry, Zuhlke, and Tormos-Aponte 2023). The EMW seeks to address the various types of diversity challenges that persist in the discipline, and within the methods community in particular.

This symposium introduces the work presented at the second EMW-QMER, held during the APSA annual meetings in September 2024 in Philadelphia, PA. The workshop featured a wonderful set of scholars and papers examining follow-up question in interview research (Gupta), the representation of interview-based research in top journals (Tuncel Gulek), infusing interviews with an experimental logic (Park), selecting negative cases (Erdoğdu), strategies for abductive analysis (Noor), and a mixed-methods approach to text as data (Morse). Two key themes to emerge from the workshop discussion – reflected in the paper summaries that follow – were the promise and perils of iteration in qualitative research, and the value of systematic attention to less visible methodological practices. Presenters are now revising their papers based on feedback from the workshop and preparing to submit them for peer review in the near future.

The workshop is supported by generous funding from the National Science Foundation's Accountable Institutions and Behavior program. We also thank the American Political Science Association's Qualitative and Multi-Method Research section for its support of this initiative. We owe a debt of gratitude to EMW Steering Committee members Alan Jacobs and Chloe Thurston for their wise counsel and warm encouragement. We also deeply appreciate the intellectual generosity and commitment of the faculty who served as Methods Mentors this year: Erik Bleich, Agustina Giraudy, Michelle Jurkovich, Danielle Lupton, Lauren MacLean, and Juan Masullo.

We encourage advanced political science graduate students and junior faculty based at U.S. institutions who are writing a paper focused specifically on developing, critiquing, challenging, or enhancing a method for gathering or analyzing qualitative data, or a technique for multi-method research, to submit proposals for consideration for presentation in future EMWs. Formerly held on the Wednesday before the annual APSA meeting begins, in the future the EMW will take place each June in association with the Institute for Qualitative and Multi-Method Research at Syracuse University. More information on the EMW can be found here: http://sigla.georgetown.domains/emworkshop/

References

Barnes, Tiffany D. (2018) "Strategies for Improving Gender Diversity in the Methods Community: Insights from Political Methodologists and Social Science Research." *PS: Political Science & Politics* 51(3): 580–87. https://doi.org/10.1017/S1049096518000513

Dion, Michelle L. "An Effort to Increase Women's Participation: The Visions in Methodology Initiative." *The Political Methodologist* 21, no. 2 (2014): 6–8.

Perry, Brittany, Sarah Zuhlke, and Fernando Tormos-Aponte. "Building Infrastructure to Enhance Diversity in Political Methodology." *PS: Political Science & Politics* 56, no. 1 (2023): 178–82. https://doi.org/10.1017/S1049096522000907.

Shames, Shauna L., and Tess Wise. (2017) "Gender, Diversity, and Methods in Political Science: A Theory of Selection and Survival Biases." *PS: Political Science & Politics* 50(3): 811–23. https://doi.org/10.1017/S104909651700066X

Smith, Amy Erica, Shauna N. Gillooly, and Heidi Hardt. "Assessing Racial/Ethnic and Gender Gaps in Political Science PhD Students' Methodological Self-Efficacy." *PS: Political Science & Politics* 55, no. 1 (2022): 165–70. https://doi.org/10.1017/S1049096521001116.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062813

What's in an interview? Follow-ups as moments of clarity and information in qualitative research

Shagun Gupta
American University

Introduction

Qualitative interviews are an important component of multi-method research, offering unparalleled insights into the lived experiences and intricate relationships that shape much of the social world. Interviews are particularly invaluable for political scientists who navigate complex socio-political landscapes where understanding local context and descriptive richness is crucial. Semi-structured interviews, characterized by their flexible yet focused nature, are a popular tool used in multi-method research designs. They allow researchers to go deeper into specific topics while leaving room for the exploration of emergent themes and unexpected findings (Creswell and Poth 2016; Creswell 2018; Cyr and Goodman 2024). This methodological flexibility makes semi-structured interviews particularly useful for graduate students and early-career researchers engaged in multi-method projects, where the depth of qualitative data complements the breadth of quantitative analysis.

Despite their widespread use and acknowledged value, qualitative semi-structured interviews present several methodological challenges. Among these, the formulation and deployment of follow-up questions stands out. While there is extensive literature guiding researchers on how to conduct interviews as part of qualitative research (notably Mosley (2013), Fujii (2017), and Brinkmann (2013; 2022)), there is a notable lack of consensus on best practices for crafting effective follow-up inquiries (Rubin and Rubin 2011). This gap is particularly pronounced when considering whether follow-up questions should be pre-designed or arise spontaneously during the interview. In the semi-structured context, researchers often encounter the challenge of ensuring the responses are clear and comprehensive. It is understood that the semi-structured interview is inherently more flexible, the scope of "semi" varies widely with crucial implications for data quality and research ethics. Pre-designed follow-up questions offer consistency and facilitate ethical design (Tracy 2019). However, these questions may lack the flexibility needed to probe deeper into unexpected responses, particularly in dynamic field settings (Rubin and Rubin 2011).

Drawing on 15 months of immersive fieldwork conducted in the informal settlements of Delhi, India, I propose a novel methodological approach that integrates pre-designed and spontaneous elements through the strategic use of clarity probes and information loops. Clarity probes are targeted follow-up questions designed to elicit further detail or clarification on specific points, ensuring the accuracy and depth of the data collected. They help researchers *know better*, by identifying moments where the lack of a probe might lead to misunderstanding or ambiguous data. Information loops involve summarizing and reflecting back what the interviewee has said to confirm understanding and encourage further elaboration. They help researchers *know more*. This approach aims to systematically identify and address moments in interviews that necessitate elaboration, thereby enhancing both the richness and reliability of the data.

An Integrated Methodological Framework

Clarity probes are specific, targeted questions designed to clarify ambiguous or incomplete responses. These probes are typically reactive,

arising from the need to clarify ambiguous or complex statements in real-time, preventing misunderstandings and enhancing the quality of the interview. Meanwhile, **information loops** are questions that help with summarizing and reflecting back the interviewee's responses to identify patterns, make connections between different parts of the conversation, and enhance overall detail and richness. Clarity probes are designed to elicit detailed and precise responses, while information loops involve summarizing and reflecting participants' statements to confirm comprehension and encourage further elaboration. This dual strategy enhances the interviewer's ability to capture the full spectrum of participant responses, addressing the limitation of pre-designed questions which may not always align with the interviewee's unique context (Kvale and Brinkmann 2009).

The Framework

The framework below outlines the decision-making process for when to use a clarity probe versus an information loop during an interview.

Step 1: Interviewer Asks Question

↓

Step 2: Receive Interviewee's Response

↓

Step 3: Assess Response for Clarity and Completeness

Is the response clear and complete? Does the interviewer feel they can move on to the next question because the response answers the question adequately in verbal or non-verbal ways?
Yes: Proceed with the next question or explore new themes.
No: Proceed to step 3.

↓

Step 4: Determine Nature of the Issue

Is the response ambiguous or incomplete? Does the interviewer feel they need to clarify something or need more information to elicit detail?

Ambiguous/Need to Clarify: Use a Clarity Probe. See Step 5.1.
Incomplete/Need More Information: Use an Information Loop. See Step 5.2.

↓

Step 5: Deploy Follow-up Technique

5.1 Clarity Probe: Ask a specific question to clarify ambiguity (e.g., "Can you explain what you mean by 'unfair treatment'?").
5.2 Information Loop: Summarize and reflect back (e.g., "You mentioned that you felt 'unfairly treated.' Could you tell me more about that experience?").

↓

Step 6: Evaluate the Response

Is the issue resolved?
Yes: Proceed with the interview.
No: Reassess and determine if another follow-up is needed.

Probing Techniques

1. **Paraphrase**: Restate the participant's response in your own words to confirm understanding.
   Example: "So, what I'm hearing is that you felt excluded because of your religious background. Is that correct?"

2. **Direct Ask**: Directly ask participants to explain unclear terms or concepts.
   Example: "You mentioned feeling 'out of place.' Can you describe what that means for you?"

3. **Summarize**: Summarize key points periodically to check for accuracy and comprehension.
   Example: "To make sure I understand, you're saying that the policy change had both positive and negative impacts on your neighborhood. Is that right?"

Looping Techniques

1. **Build:** Continuously ask questions that build on previous responses.
   Example: "You mentioned your supervisor was supportive. How did that support manifest in your daily work?"

2. **Layer:** Start with broad questions and progressively narrow down to specifics.
   Example: "Can you describe your daily routine in the clinic?" followed by "What challenges do you face in addressing the needs of patients here?"

3. **Reflect:** Encourage participants to reflect on their responses and provide further insights.
   Example: "Looking back, how do you feel about the decisions you made during that period?"

4. **Validate:** Repeat key points and ask for specifics to ensure accuracy and detail.
   Example: "You said that the team collaboration improved after the training. Can you give an example that illustrates this improvement?"

Using Probes and Loops in Semi-Structured Interviews

Researchers can anticipate the need for potential probes and loops based on existing knowledge of the field context and add these to their protocols for ethical review as well. This can be done by developing a set of potential questions for probing and looping techniques discussed in sections 2.2 and 2.3 based on your interview questionnaire. While it can be difficult to anticipate when you'll need to paraphrase or summarize for clarity or build for information, language for the same can be drafted. Figure 1 provides an example of using other techniques.

Figure 1

Question: Do households try to collectively solve neighborhood problems or does each household deal with their own specific issues?

Potential Follow-ups for Clarity

Paraphrase: So, you're saying households collective solve problems. Is that correct?
Direct Ask: What is your meaning of "collective" and "solve" for you?
Direct Ask: You said yor personally contact the local authorities if you have a problem. What do you mean by contact? Do you call them or visit their office?

Potential Follow-ups for Information

Layer: You said households come together to solve problems / you deal with problems yourself. Can you give describe how households organized collectively in recent months / how you dealt with these problems in recent months?
Layer: In this example, was the process determined by the nature of the problem?
Reflect: Do you think the process or the steps you took to solve your specific issue were easy for you?
Validate: You said you call the local office if you have a problem. Can you give me an example, say the last phone call you had with the office? What did you say and what was their response?

Concluding Remarks

This article presents a methodological strategy for using follow-up questions in qualitative interviews, integrating clarity probes and information loops to enhance data quality and participant engagement. Based on extensive fieldwork conducted among residents living in informal settlements in Delhi, India, this approach demonstrates the effectiveness of balancing design and spontaneity in semi-structured interviews. By identifying moments that require clarification or elaboration, using techniques to convert those moments

into data, and fostering a collaborative interview process, researchers can uncover hidden nuances and gain deeper insights into complex human experiences. This framework not only addresses a critical gap in existing literature on qualitative research, specifically semi-structured interviews, but also offers practical guidance for researchers working in similar contexts, contributing to a richer and more ethical understanding of socio-political phenomena.

Future research could explore the applicability of these methodological techniques across diverse cultural and geographical settings. Comparative studies could investigate how varying socio-cultural contexts influence the efficacy of clarity probes and information loops in qualitative interviews. Additionally, longitudinal studies could assess the long-term impact of integrating these strategies on data reliability and participant rapport over extended research periods. Furthermore, expanding the application of these methods beyond urban informal settlements to other settings, such as rural communities or different demographic groups, would broaden the scope of their utility and enrich our understanding of their potential limitations and adaptations. Addressing these avenues for further inquiry would advance methodological innovation in qualitative research, improve researcher ability to collect high quality data, and foster more nuanced research in the social sciences.

References

Brinkmann, Svend. 2013. *Qualitative Interviewing*. OUP USA.

———. 2022. *Qualitative Interviewing: Conversational Knowledge Through Research Interviews*. Oxford University Press.

Creswell, John W. 2018. *Designing and Conducting Mixed Methods Research*. Third edition. Thousand Oaks, California: SAGE.

Creswell, John W., and Cheryl N. Poth. 2016. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. SAGE Publications.

Cyr, Jennifer, and Sara Wallace Goodman. 2024. *Doing Good Qualitative Research*. Oxford, United States: Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/aul/detail.action?docID=31177111.

Fujii, Lee Ann. 2017. *Interviewing in Social Science Research: A Relational Approach.* Oxford: Taylor & Francis Group.

Kvale, Steinar, and Svend Brinkmann. 2009. *InterViews: Learning the Craft of Qualitative Research Interviewing*. SAGE.

Mosley, Layna. 2013. *Interview Research in Political Science*. Cornell University Press.

Rubin, Herbert J., and Irene S. Rubin. 2011. *Qualitative Interviewing: The Art of Hearing Data*. SAGE.

Tracy, Sarah J. 2019. *Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact*. John Wiley & Sons.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062819

Elite Interviewing in Political Science: Insights from Leading Journals

Ozlem Tuncel
Georgia State University

Introduction

Interviewing elites provides unique insights and generates reliable data for investigating political complexities (Berry 2002).[1] Scholars often revere this tool and underscore the advantages of its use in developing and testing theories, but there is a limited discussion on designing and reporting elite interviews. This article draws on an original dataset of articles from major political science journals (2000–2023) to analyze the use and evolution of elite interviewing and assess reporting practices. Given the iterative nature of data collection and analysis in interview research, thorough reporting is crucial for enhancing transparency and accountability (Bleich 2013). Therefore, researchers should prioritize robust reporting practices to maximize the potential of elite interviews, focusing on aspects such as transparent sampling and clear documentation.

This article presents two key findings. First, while elite interviews make up a small share of published research in our field, most of these studies appear in Comparative Politics journals. Second, there has been a promising shift towards better reporting practices, particularly in recent years. This includes greater attention to ethical issues and increased transparency through online appendices, which provide details on the interview process, including anonymity and ethics. This article seeks to further the ongoing efforts to improve reporting standards, especially for elite interviewing in qualitative and mixed-methods research.

The dataset of elite interviewing practices

For this meta-analysis, I created an original dataset of articles that utilized elite interviews from major political science journals between 2000 and 2023 – a period marked by renewed interest in qualitative and mixed-method research (Bennett and Elman 2007). Books, edited volumes, and research notes were excluded, focusing solely on articles that employed elite interviews as a research tool.[2] I manually coded each article, with particular attention to details such as sampling and recruitment procedures, descriptions of interview subjects, modes and conduct of interviews, anonymization practices, and ethics committee approvals.

Elite interviewing coverage in political science journals

The dataset has 145 unique articles, 63 percent single-authored and 37 percent co-authored. Around 57 percent of the articles used a qualitative approach, while the rest relied on mixed methods. About one percent of 14,870 articles published between 2000 and 2023 in these journals use elite interviews. The presence of elite interviews varies significantly across journals, with Comparative

1 Elites hold or have held privileged positions in society and likely influence political outcomes more than the general public (Richards 1996).

2 The dataset includes the following 13 journals: *American Political Science Review, American Journal of Political Science, British Journal of Political Science, Comparative Political Studies, Comparative Politics, Democratization, International Organization, International Security, Political Research Quarterly, The Journal of Politics, World Politics, International Studies Quarterly, and Journal of Peace Research.*

Politics journals featuring more articles. Almost 48 percent of these articles were published in just three journals: *Comparative Politics*, *Comparative Political Studies*, and *Democratization*.

Various topics are covered in these articles. Political regimes and conflict research account for the largest share, with political elites, government officials, policymakers, and bureaucrats as primary subjects. Research on institutions, global governance, electoral politics and party politics accounts for 39 percent of the articles. Geographically, these interviews come from various parts of the world, with almost half conducted in African and European countries. Most interviews focus on single-country studies (66 percent), which is expected given the required linguistic skills, context knowledge, and logistical resources.

Assessment of best practices in elite interview research

In addition to journal-level information, I provide an assessment of interview practices focusing on nine main issues: sample size, recruitment strategies, modes of conducting interviews, interview structure, sample description, anonymity, ethical considerations, data sharing, and the adoption of supplementary appendices.

While some influential publications suggest that increasing the sample size can improve research quality (King, Keohane, and Verba 1994), recent critiques argue that the relevance of the sample matters more than its size (Gonzalez-Ocantos and Masullo 2024). In my dataset, sample sizes varied widely, with an average of 68 interviews. Surprisingly, around 28 percent of the articles did not mention the sample size.

Explicit reporting of sampling strategies and sample frames enhances research transparency and reader confidence. However, around 71 percent of the articles did not discuss recruitment procedures in the main text or supplementary materials. The lack of recruitment information is particularly concerning given that elites are considered a hard-to-reach population. In 11 percent of the articles, snowball sampling was used, while other articles employed purposive, non-purposive, or positional sampling. While this information can help the readers (especially the graduate students or novice interviewers) garner know-how, only a few articles provided detailed recruitment information in supplementary appendices.

Traditional fieldwork research (i.e., in-person interviews) was the dominant approach (80 percent) in this sample. Phone interviews (8 percent) and online interviews (5 percent) were less common. Furthermore, around 59 percent of the articles did not specify the interview structure. Among those that did, most research used semi-structured interviews.

Given that the definition of the elites is contested, I also examined how researchers described these elites. Surprisingly, the definition of elites remained undefined in most articles (around 82 percent). However, having an online appendix positively correlated with the reporting of interviewee lists and questions ($p < 0.01$).

Safeguarding respondent confidentiality while providing detailed portrayals of elites' roles is crucial. Authors predominantly anonymized interviewees using pseudonyms or numbers (70 percent). Some revealed the elites' names (23 percent), though they rarely explained whether they obtained permission. Notably, 88 percent of the papers did not discuss the reasons behind their anonymity decisions.

Lastly, protecting human research participants through Institutional Review Boards (IRBs) or equivalent ethics committees is essential. Although IRB information was rarely mentioned in the articles (17 percent), there has been increased attention to ethics and research transparency in recent years. The use of supplementary information has become more popular recently, and appendices have positively correlated with the reporting of IRB information ($p < 0.01$).

Concluding remarks

My analysis shows that elite interviewing is more common in subfield journals than in mainstream ones, reflecting broader trends in the discipline. However, progress in reporting these practices has been uneven. Key details about the interview process are often poorly reported, creating ambiguity and limiting readers' ability to evaluate

the quality of the research. On a positive note, the use of supplementary materials has encouraged researchers to share more insights about their interview methods. Moving forward, I plan to propose practical solutions to help researchers and journals improve reporting on elite interviewing.

References

Bennett, Andrew, and Colin Elman. 2007. "Qualitative Methods: The View from the Subfields." *Comparative Political Studies* 40, no. 2 (February): 111–121.

Berry, Jeffrey M. 2002. "Validity and Reliability Issues in Elite Interviewing." *PS: Political Science & Politics* 35, no.4 (December): 679-682.

Bleich, Erik, and Robert Pekkanen. 2013. "How to Report Interview Data." in *Interview Research in Political Science*, edited by Maria Elayna Mosley. Ithaca: Cornell University Press, 84–105.

Gonzalez-Ocantos, Ezequiel, and Juan Masullo. 2024. "Aligning Interviewing with Process Tracing." *Sociological Methods & Research*. https://doi.org/10.1177/00491241241258822

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

Richards, David. 1996. "Elite Interviewing: Approaches and Pitfalls." *Politics* 16, no. 3 (September): 199–204.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062823

Randomized Priming Interviews (RPI): Integrating Qualitative Interviews with Experimental Logic

Eun-A Park
University of California, Riverside

Introduction

In this paper, I propose a mixed-method design called "randomized priming interviews" (RPI) that combines the strength of priming experiments with the richness of in-depth interview methods. Unlike existing mixed-method designs that tend to sequentially combine qualitative and quantitative methods, the RPI integrates experimental logic with qualitative data collection and analysis ***without*** quantitative components. The common and standard advice for mixed-method designs is to employ either experimentation or quantitative analysis first and *then* qualitative inquiry or vice versa (Bell-Martin 2022; Martin 2013; Seawright 2016). This sequential approach often requires conducting large-N studies to establish causal relations at a macro-level and using qualitative data to assess whether the causal relations hold; therefore, current mixed-method strategies in political science are mostly suitable for quantitatively oriented projects with "many variables" and "many-Ns" that purse law-like regularities in social phenomena.

I argue that ***qualitative projects***, especially those that take interpretive approaches, can implement mixed-method designs without quantitative-statistical components. Using my own research, I demonstrate that methods of qualitative data collection and analysis, especially in-depth interviews, can be connected with and strengthened by elements of experimental designs to compare process of meaning construction and causal patterns given controlled variation in information.

What are Randomized Priming Interviews (RPI)?

The RPI developed through my own research are one way to create mixed-method design for ***qualitatively oriented projects***, whose central goal is to uncover "thicker" narratives and meanings of contextually situated social actions and people's beliefs about the world. A qualitatively oriented researcher may wish to explore whether a cause/factor is associated with the theoretical concept under study and what the causal mechanisms look like, even though their primary goal is not the quantification of variables or estimation of causal effects. Following Dessler's (1991) notion of "an integrative process,"[1] The RPI is an example of an integrative mixed-method design in which two or more methods **are *simultaneously connected*** to advance a unified and concurrent inference bounded to a specific time and context.

The RPI design makes priming experimental logic and in-depth interview method complementary and integrated ***concomitantly*** within a single study. Priming theory suggests that individuals tend to employ intuitive shortcuts and simple rules of thumb by relying on the most accessible information that are ready to use without effort when make a judgment or choice (Krosnick & Kinder 1990). As such, priming methods can provide an empirically grounded and psychologically plausible account of how individuals form and revise their views of certain topics given primed information. The RPI design adopts the strengths of priming experiments

1 Dessler (1991, 340) suggests "an integrative progress" as "a qualitative rather than quantitative improvement of knowledge," and where "it requires not more findings, or better findings, but ***better-connected findings***" [emphasis added].

to direct research participants' attention to the issues that the researcher intends to study through random assignments.

Moreover, the RPI design utilizes the power of in-depth interview methods in providing much deeper understanding of how participants make sense of their experiences and social world. Notably, political scientists have used interview methods for testing or generating falsifiable hypotheses, identifying causal processes, and developing deeper knowledge about a certain community or issues (Mosley 2013), as well as causal mechanisms that may not be available in other forms of data, such as observational and experimental data. Even a single interview can offer information about actions and attitudes held by not only the interviewees but also their neighbors, colleagues, and family members.

An Illustration: Applying the RPI Design to A Qualitatively Oriented Research Project

I apply the RPI design to my own research. I examine whether and to what degree native citizens' differential treatment shapes intergroup attitudes of migrants and their subjective national belonging. I study two co-ethnic migrant groups in South Korea –North Korean migrants and Chinese Korean migrants– who have been socially and politically marginalized and stigmatized due to the ongoing security threat from North Korea and geopolitical conflicts between South Korea and China. I want my findings to accurately reflect views and perceptions of these marginalized migrants based on the intersection of their gender and country of origins. I also want to affirm that whether responses from these migrants would differ depending on knowing or not knowing native South Koreans' differential treatment toward them. In this context, my project is qualitatively oriented with an aim to explore how a cause/factor (host citizens' differential treatment in my research) influences intergroup attitudes of migrants and perceptions of their national belonging.

Using the RPI design, I randomly assign each interview participant to one of two conditions (an experimental condition and a control condition) in which they will be exposed to different information. Each participant will undergo four data collection phases suggested by Robinson and Mendelson (2012). Participants assigned to the experimental condition will be exposed to information about native South Koreans' differential treatment that I constructed based on the Unification Perception Survey[2] data from 2021. Participants in the control condition will be asked of all the interview questions as the treatment group, with the only difference being they are asked questions related to the outcome of interest before giving the native citizens' differential treatment. In this case, the RPI design allows me to achieve my research goals by comparing North Korean migrants' attitudes toward Chinese Korean migrants with and without the "native citizens' differential treatment" information.

Concluding Remarks

Qualitative approaches have been known to offset weakness of quantitatively oriented projects by enhancing internal and external validity (Bell-Martin 2022; Cyr 2017; Encinas 2022; Gonzalez-Ocantos & Masullo 2022) and uncovering the meanings of the behavior measured in experimental contexts (Paluck 2010). Accordingly, the discussion of mixed method designs in political science mostly centers on how to sequentially combine qualitative and quantitative inquiries in multi-phased studies (e.g., Lieberman 2005; Tarrow 1995). In such designs, quantitative components are indispensable, and they are at the heart of the research projects that aim to find law-like propositions across a large number of cases.

In this paper, I have demonstrated one way to conduct mixed method research for qualitatively oriented projects by highlighting two characteristics in the RPI design. First, different from the existing mixed-method designs that often combine qualitative and quantitative inquires, the

2 The "Unification Perception Survey" is an annual survey conducted by the Institute for Peace and Unification Studies at Seoul National University since 2007 to track changes in South Koreans' perceptions of North Korea and their attitudes toward unification.

RPI design integrates qualitative methods (e.g., in-depth interview) with experimental logic but without quantitative data collection and analysis involved. Second, the RPI design allows the simultaneous capture of the differences in the process of meaning construction and the causal relations between groups in a single phase of experimental execution. In other words, unlike existing mixed-method strategies, the RPI design helps researchers to implement mixed-method designs for qualitatively oriented projects by using a combination of different methods concomitantly within a single-phased study.

References

Bell-Martin, R. (2022). Experimentally Testing Ethnographic Insights: Benefits, Challenges, and Considerations. *Qualitative and Multi-Method Research*, *19/20(2/1)*, 2–7.

Cyr, J. (2017). The Unique Utility of Focus Groups for Mixed-Methods Research. *PS: Political Science & Politics*, *50*(04), 1038–1042.

Dessler, D. (1991). Beyond Correlations: Toward a Causal Theory of War. *International Studies Quarterly*, *35*(3), 337.

Encinas, D. (2022). Experiments as case studies: a qualitative approach to external validity. *Qualitative & Multi-Method Research*, *19/20(2/1)*, 20–30.

Gonzalez-Ocantos, E., & Masullo, J. (2022). On the potential complementarities between qualitative work, surveys, and experiments. *Qualitative & Multi-Method Research*.

Krosnick, J. A., & Kinder, D. R. (1990). Altering the Foundations of Support for the President Through Priming. *American Political Science Review*, *84*(2), 497–512.

Levy Paluck, E. (2010). The Promising Integration of Qualitative Methods and Field Experiments. *The ANNALS of the American Academy of Political and Social Science*, *628*(1), 59–71.

Martin, C. J. (2013). CRAFTING INTERVIEWS TO CAPTURE CAUSE AND EFFECT. In L. Mosley (Ed.), *Interview Research in Political Science* (1st ed., Vol. 1, pp. 109–124). Cornell University Press.

Seawright, J. (2016). *Multi-Method Social Science*. Cambridge University Press.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062829

Nearly Realized Cases: A Novel Framework to Select Negative Cases for Comparative Theory Development

Ulaş Erdoğdu
Northwestern University

Introduction

For scholars working with a relatively small number of cases, the selection of negative cases –where the outcome of interest does not occur– has significant consequences for theory development and testing (Mahoney and Goertz 2004). Rare events such as interstate and civil wars, revolutions, coups, genocides, state collapses, and democratic breakdowns are especially important for qualitative scholars since their rarity makes them particularly suitable for qualitative comparison. Yet, their rarity creates problems for negative case selection.

At any given time, most countries are not experiencing revolutions, coups, genocides, armed conflicts, economic crises, or democratic breakdowns. In other words, negative cases far outnumber positive ones. This creates a challenge for researchers, as they must choose nonevents (negative cases) from a disproportionately large number of cases (Mahoney and Goertz 2004). Despite this challenge, there are few articles on negative case selection (Emigh 1997; Mahoney and Goertz 2004; Gray 2018).[1]

Whereas Mahoney and Goertz's (2004) widely read and cited[2] possibility principle provides a very useful framework, it is explicitly developed for theory testing and assumes strong thereoretical expectations. How, then, do we choose our negative cases if our objective is to develop theories? I develop a simple yet powerful framework called "nearly realized cases" to mitigate this problem, where we choose negative cases that are as close as possible to the outcome of interest from a conceptual perspective, where a case has *a negative value on only one constitutive dimension of the outcome*. Below, I briefly explain the possibility principle's weakness, introduce my framework, and discuss its contributions.

The Possibility Principle

Mahoney and Goertz (2004) propose that we should select negative cases where the outcome of interest is possible: "cases are relevant if their value on at least one independent variable is positively related to the outcome of interest." (Mahoney and Goertz 2004, 657) and "a case is considered irrelevant if it possesses a value on a variable that is known from previous research to make the outcome of interest impossible" (Mahoney and Goertz 2004, 658). For example, in Skocpol's theory of social revolutions (1979), where state breakdown and peasant rebellion are identified as causes of her outcome of interest, negative cases would be those that experienced either state breakdown or peasant revolt (Mahoney and Goertz 2004, 659). While these rules make great sense, the possibility principle has weaknesses.

First, the possibility principle relies on an already-developed theory. It uses theoretical expectations and independent variables that predict the outcome of interest and its absence to guide case selection.

1 Emigh discusses how deviant cases drive case selection, whereas Gray discusses literal near misses in daily life such as where an accident almost occured.
2 Cited more than 800 times at the time of this writing.

Second and relatedly, qualitative studies and a small number of cases are particularly effective for building theories, whereas the possibility principle is explicitly developed for theory testing. However, how do we even develop the theories in the first place to test on the negative cases?[3] I prose nearly realized cases to mitigate this issue.

Nearly Realized Cases

All concepts have constitutive dimensions that can be derived from their definitions or are explicitly defined by multiple sets of constitutive dimensions. For instance, democracy is often defined as a system in which there are 1) competitive elections 2) free and fair elections and 3) extensive suffrage; or following Skocpol, social revolutions are defined as instances where there is 1) mass mobilization 2) rapid transformation of the state 3) rapid transformation of the society. Thus, both of these concepts are intersections of these three sets. This gives us set-theoretical relations in which we can move between sets to see which cases were close to the realization from an ontological/conceptual perspective. These relationships can be visualized with Venn diagrams or tables.

I define nearly realized cases as cases that lack only one dimension of the outcome of interest, i.e., have a negative value on only one constitutive dimension of the outcome of interest. The logic here is that cases that are similar in terms of the outcome (positive cases and nearly realized cases) are likely to be similar in terms of independent variables/ conditions and/or processes. Thus, studying and comparing these cases can help us develop a fully-fledged theory of what *causes* and *prevents* the rare case from happening. An example of nearly-realized cases of social revolutions is given below (Table 1).[4] Here, one can easily imagine that these nearly-realized cases are much more likely to have similar conditions and follow similar processes to a positive case that experienced a social revolution compared to a case that experienced none of these conditions.

Table 1

Mass Mobilization	Transformation of the State	Transformation of the Society	Case of
1	1	1	Positive Cases
1	1	0	Nearly Realized Cases
1	0	1	Nearly Realized Cases
0	1	1	Nearly Realized Cases

Contribution

Although this logic of nearly realized cases can be extended to any type of outcome, it is especially likely to be helpful in the study of rare events. First, by definition, most cases will have non-positive values on the dependent variable, meaning there will be too many negative cases to choose from. Second, given their rarity, these events often happen or do not happen through very complex processes where standard statistical techniques may fall short of providing insights.

This framework broadly contributes to the literature on case selection, in particular, to the literature on selecting on the dependent variable and selecting negative cases by developing a new framework to do so. Moreover, most case selection techniques focus on theory testing and/or are driven by theory and independent variables. To the best of my knowledge, this is the first framework that selects cases based on the conceptual structure of the dependent variable/outcome of interest and explicitly aims at theory development. This framework is especially useful for scholars who select their positive cases for logistical reasons, interest, expertise or simple convenience by providing them a tool to choose a negative case for comparison. It is also broadly compatible with both interpretivist and more mainstream approaches to comparison. Thus, it is likely to be beneficial for many scholars.

3 This arguably reflects a bigger problem in case selection literature in that most case selection techniques are -even if implicitly- geared towards theory testing.

4 Although I have shown this using the classic approach to concept formation using binary values, it can easily be extended to the family resemblance approach to concept formation and to continuous set analysis.

References

Emigh, Rebecca Jean. 1997. "The Power of Negative Thinking: The Use of Negative Case Methodology in the Development of Sociological Theory." *Theory and Society* 26 (5): 649–84. https://doi.org/10.1023/A:1006896217647.

Gray, Garry. 2018. "The Sociology of Near Misses." *The Canadian Journal of Sociology / Cahiers Canadiens de Sociologie* 43 (2): 171–90.

Mahoney, James, and Gary Goertz. 2004. "The Possibility Principle: Choosing Negative Cases in Comparative Research." *American Political Science Review* 98 (4): 653–69. https://doi.org/10.1017/S0003055404041401.

Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511815805.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062835

Abductive analysis in qualitative and multimethod political science

Salih O. Noor
University of Chicago

This paper introduces an *abductive framework* for qualitative and multimethod political science. Abduction emphasizes an iterative dialogue between theory and observation, integrating principles of deductive hypothesis formulation with inductive theory evaluation. The paper begins by outlining relevant methodological debates. It then briefly discusses the methodological foundation, techniques for generating explanatory hypotheses, and strategies for hypothesis testing in abductive analysis. The insights and examples are drawn from non-interpretivist, theory-driven qualitative political science and sociology interested in causal explanation.

The inductive and deductive paradigms are widely seen as alternative approaches in social inquiry. Debates over qualitative and quantitative methods are rooted in these paradigms' contrasting epistemological and methodological foundations. However, scholars routinely underscore the superficiality of this divide, highlighting that researchers typically trespass methodological boundaries to leverage both inductive and deductive strategies. For instance, inductively-oriented small-N scholars employ deductive methods for hypothesis formulation and generalization. Similarly, mainstream quantitative scholars subscribing to deductive proceduralism use inductive strategies (e.g. analytic narratives) for hypothesis evaluation and theory refinement through a back-and-forth between theory and data, albeit discreetly (Yom 2016). Finally, mixed-methods research incorporates inductive methods of hypothesis testing into a primarily deductive template to enhance the robustness and validity of findings (Lieberman 2005; Humphreys and Jacobs 2015; Seawright 2016).

While some social scientists show qualitative and quantitative methods constitute different cultures (Goertz and Mahoney 2012), others promote a unified framework emphasizing the significance of discovery, measurement validity, and methodological synthesis (Gerring 2012; Brady and Collier 2004; Adcock and Collier 2001). Iterative analysis is widely recognized as crucial for both qualitative and quantitative analysis (Collier, Seawright, and Munck 2004; George and Bennett 2005; Mahoney 2010). In recent years, *iteration* received growing attention in methodological discussions. For instance, Fairfield and Charman stress the significance of "back and forth between theory development, data collection, and data analysis, rather than a linear sequence from hypothesizing to testing" (2019, 155). Yom (2015) argues that "inductive iteration" contributes to causal analysis in comparative-historical analysis, analytic narratives, and statistical approaches. And Kapiszewski, MacLean, and Read (2022) emphasize dynamic research-design updating to align theory, methods, and data in the course of research.

The abductive approach to inquiry encompasses, among other strategies, iteration, induction-deduction synthesis, and theory adjustment. Abductive logic —ubiquitous in everyday life, underpinning medical diagnostics and criminal detective work— was first elaborated by the pragmatist scientist-philosopher Charles S. Pierce (1839-1914). Peirce distinguished abduction from the dominant logics of deduction and induction, asserting that abduction "is the process of forming an explanatory hypothesis, the only logical operation which introduces any new ideas" (1934, 171). The broader abductive framework consists of three stages of analysis: first, the adoption of an explanatory hypothesis based on surprising evidence —the essence

of Peirce's abduction; second, working out its observable implications using deductive procedures; and finally, inductive evaluation of hypothesis by generating empirical evidence that conforms to or prompts reformulation of hypothesis (Timmermans and Tavory 2012, 171).

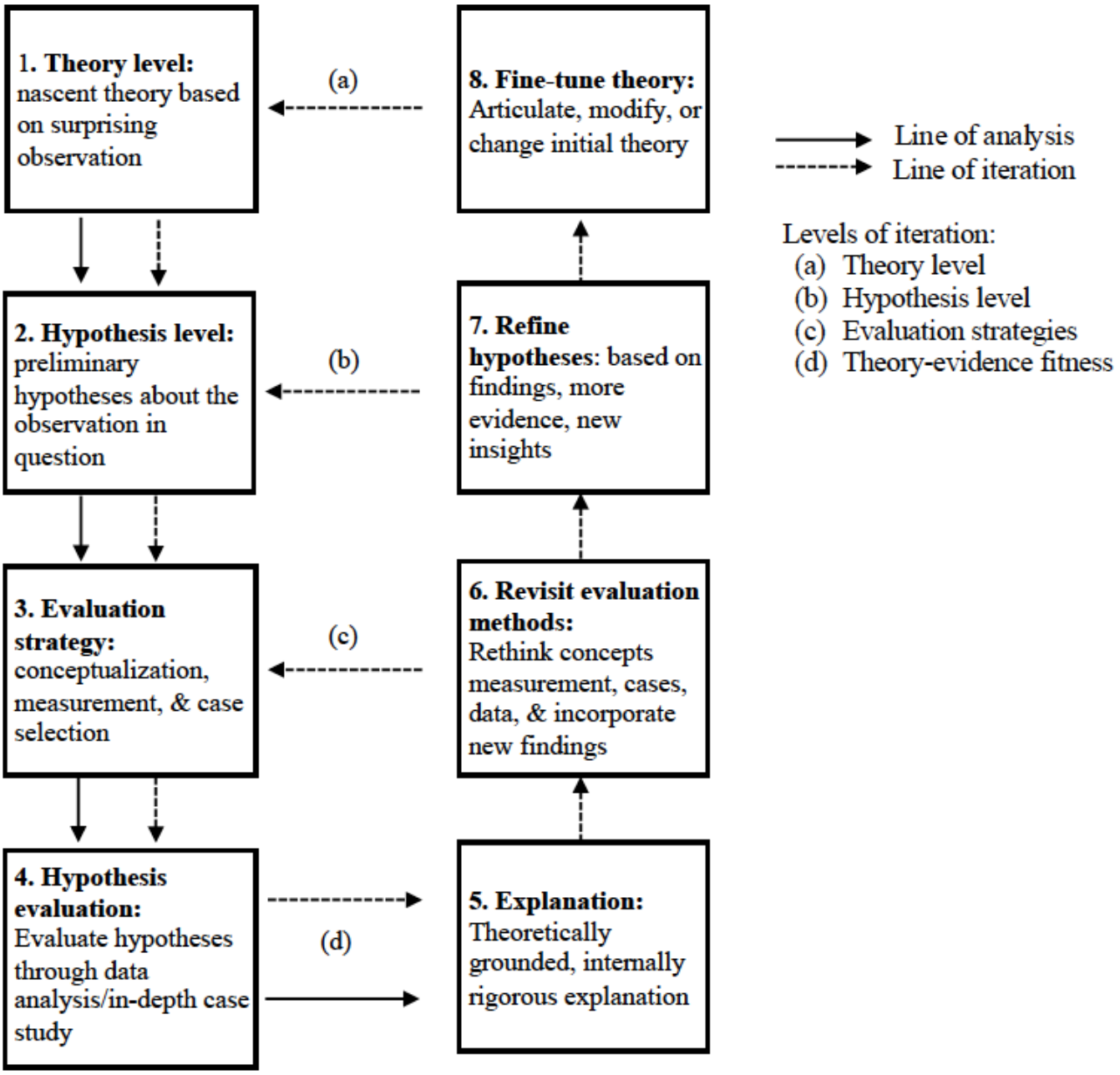
Abductive methods are employed in qualitative theorizing, although scholars rarely identify their approach exactly as abductive. Hence, their work is often mistaken for induction and sometimes criticized as an inductive theory-free wanderlust. Peirce viewed scientific inquiry as a continuous process, where discovery and theory justification are inseparable. This aligns with qualitative analysis in comparative politics and sociology engaged in theory development (and testing). While few rules may exist for uncovering novel hypotheses, four ***strategies of hypothesis generation*** characterize abductive analysis: observation, theoretical familiarity, skepticism and dis-belief, and exploratory research. In essence, abductive analysis begins with a surprising *observation* or puzzle that contradicts expectations of existing theories (Swedberg 2014, 39; Tavory and Timmermans 2014, 36). Innovative explanations in case-study and small-N analysis often emerge from social or political observations that challenge conventional theoretical accounts (see Munck and Snyder 2007).

Meanwhile, researchers must possess *theoretical familiarity* to foster innovation and avoid reinventing the proverbial wheel. While in-depth case-knowledge is crucial, novel theoretical insights (e.g., deviant cases) emerge from intimate engagement with extant scholarship (Mahoney and Rueschemeyer 2003; also, George and Bennett 2005). Empirical observations can be puzzling only to a "theoretically sensitized" observer, and fresh theorization hinges on "the scope and sophistication of the theoretical background a researcher brings along" (Tavory and Timmermans 2014, 41). However, using established theories does not mean commitment to deductive theory-testing. Instead, *skepticism and dis-belief* allow scholars to revisit, "reevaluate and rethink" the phenomenon under investigation (Timmermans and Tavory 2012). "Dis-belief" in dominant concepts, numbers, and conventional accounts contributes to generative hypotheses (Gerring 2012, 64). Finally, pre-analysis *exploration* is central to abductive theorization in qualitative social science. Rather than first drawing a rigid research design, qualitative theorizing involves "early theorization," developing what Mann (1994) called a "theoretical hunch" further developed and rigorously tested during main analysis (Swedberg 2012, 26; Gerring 2012, 52; also, Peirce 1935).

The ***strategies for hypothesis evaluation*** in abductive analysis include iterative conceptualization, measurement, and hypothesis refinement in response to evolving empirical findings. In this process of "analytic iteration," as Imre Lakatos (1978) suggested, theoretical arguments are constantly refined, adjusted, and modified in a recursive movement between theory and empirical data. In qualitative theorizing, hypothesis evaluation contends with diverse cases, some of which fit neatly into the theoretical construct while others challenge it. Instead of outright deductive falsification of the theory, analysts can either exclude non-fitting cases using a "monster-barring" strategy or adjust the theoretical scope to fit it to anomalous or exceptional cases (Lakatos 1976). Observed variation should prompt refinement of both the explanans and the explanandum to allow sharpening one another and achieve a better fit (Katz 2001).

Figure 1 presents a model of analytic iteration with four levels. In this non-linear process of abductive hypothesis generation and evaluation, analysts start on the left-hand side with theoretical hunches and proceed to hypothesis evaluation (solid line). This differs from both barefoot induction that begins with atheoretical data and deductive positivism which proceeds from a fully-fledged theory ready to inflexible top-down hypothesis testing. Instead of aiming to (dis)confirm a theoretical claim in the first go, investigators return to the theory, hypotheses, and evaluation strategies, and so forth (dashed line). Additionally, iteration between theory and data can take place at different levels of the framework, vertically and horizontally, and repeated multiple times. This back-and-forth continues until a theoretically grounded and internally robust explanation satisfactorily accounting for all observed cases is reached.

Figure 1. A model of analytic iteration in abductive analysis



The lower-half of the model focuses on abductive hypothesis evaluation techniques, creating a short feedback-loop between evaluation techniques and the evolving explanation. The primary goal is to refine concepts, measurements, and evidence to enhance internal validity. It addresses conceptual and measurement issues arising from ambiguous or poorly refined initial concepts, poor measurement, and internal contradictions or inconsistencies in the argument (Brady and Collier 2004; Goertz 2006; Goertz and Mahoney 2012). As Adcock and Collier (2001) argue, the iterative dialogue between theoretically defined general concepts and empirically derived observations sharpen preliminary concepts, theoretical claims, and measurement strategies. The preliminary conceptualization and explanation thus undergo "friendly amendments" when prior conceptualizations are found lacking or causal assumptions prove invalid (p. 533). In case-

oriented, theory-driven qualitative research that employ abductive methods, conceptual specificity and equivalence are crucial when studying different contexts or applying a particular concept to diverse contexts. Iteration in the lower boxes facilitates these goals using two sets of strategies found in the literature: pursuing context-specific pieces of observation and context-specific indicators (see Adcock and Collier 2001) as well as "matched comparison" (ibid.) and contextualized comparison (e.g. Locke and Thelen 1995).

In conclusion, abductive logic offers a robust methodological basis that mitigates the pitfalls of over-generalization in theory-laden deduction and crude empiricism of traditionally theory-free induction. It offers methods and strategies for iterative analysis broadly applicable beyond theory-driven qualitative political inquiry, except most experimental research since investigators cannot alter treatments and make other changes amid the study.

References

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95, no 3 (November): 529–46.

Brady, Henry E., and David Collier. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. New York: Rowman & Littlefield Publishers.

Collier, David, Jason Seawright, and Gerardo L. Munck. 2004. "The Quest for Standards: King, Keohane, and Verba's Designing Social Inquiry." *In Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited by Henry E. Brady and David Collier, 21–50. New York: Rowman & Littlefield Publishers.

Fairfield, Tasha, and Andrew Charman. 2019. "A Dialogue with the Data: The Bayesian Foundations of Iterative Research in Qualitative Social Science." *Perspectives on Politics* 17, no. 1 (March): 154–67.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Boston: MIT Press.

Gerring, John. 2012. *Social Science Methodology: A Unified Framework*. 2nd ed. Cambridge: Cambridge University Press.

Goertz, Gary. 2006. *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton: Princeton University Press.

Humphreys, Macartan, and Alan M. Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109, no 4 (November): 653–73.

Kapiszewski, Diana, Lauren M. MacLean, and Benjamin L. Read. 2022. "Dynamic Research Design: Iteration in Field-Based Inquiry." *Comparative Politics* 54, no. 4 (July): 645–70.

Katz, Jack. 2001. "Analytic Induction." In *International Encyclopedia of the Social and Behavioral Sciences*, 1:480–84. Elsevier Oxford, UK.

Lakatos, Imre. 1976. *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge: Cambridge University Press.

———. 1978. *The Methodology of Scientific Research Programs*. Cambridge: Cambridge University Press.

Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99, no. 3 (August): 435–52.

Locke, Richard, and Kathleen Thelen. 1995. "Apples and Oranges Revisited: Contextualized Comparisons and the Study of Comparative Labor Politics." *Politics & Society* 23, 3 (September): 337–67.

Mahoney, James. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62, no. 1 (January): 120–47.

Mahoney, James, and Dietrich Rueschemeyer. 2003. *Comparative Historical Analysis in the Social Sciences*. Cambridge University Press.

Mann, Michael. 1994. "In Praise of Macro-Sociology: A Reply to Goldthorpe." *The British Journal of Sociology* 45, 1 (March): 37–54.

Munck, Gerardo L., and Richard Snyder. 2007. *Passion, Craft, and Method in Comparative Politics*. Baltimore: JHU Press.

Peirce, Charles. 1934. "Collected Papers of Charles Sanders Peirce." In *Pragmatism and Pragmaticism*, edited by C. Hartshorne and P. Weiss. Vol. 5. Cambridge: Harvard University Press.

———. 1935. "Collected Papers of Charles Sanders Peirce." In *Scientific Metaphysics*, edited by C. Hartshorne and P. Weiss. Vol. 6. Cambridge: Harvard University Press.

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

Swedberg, Richard. 2014. *The Art of Social Theory*. Princeton: Princeton University Press.

Tavory, Iddo, and Stefan Timmermans. 2014. *Abductive Analysis: Theorizing Qualitative Research*. Chicago: University of Chicago Press.

Timmermans, Stefan, and Iddo Tavory. 2012. "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis." *Sociological Theory* 30, 3 (Sep.): 167–86.

Yom, Sean. 2015. "From Methodology to Practice: Inductive Iteration in Comparative Research." *Comparative Political Studies* 48, no. 5 (April): 616–44.

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062835

Embracing a Mixed Methods Approach to Text as Data

Irene Morse
University of Michigan

Introduction

Text has long been an important source of information in the social sciences. Historically, using text to measure a particular concept has required practitioners to read (and re-read) documents, making interpretive decisions along the way. In the humanistic social sciences, a great deal of weight is given to expert debate when evaluating the interpretation of texts. This generally means that the research community trusts a particular researcher's interpretation up until another researcher challenges it. Therefore, within this humanistic tradition, using text as a source of information has typically been a highly qualitative method.

Over the past century, researchers have developed ways of quantifying the qualitative information gleaned from texts. Using spreadsheets or other systemic tracking methods, researchers can transform their notes, thoughts, and interpretations into quantitative data. Advances from computer science in the field of Natural Language Processing (NLP) have allowed researchers to quantify texts in new ways and at a much larger scale. Such methods generally utilize algorithms to model text as statistical distributions. Within the social sciences these methods have been termed quantitative text analysis (QTA) or alternatively computational text analysis (CTA).

CTA methods provide an efficient way to analyze large amounts of text and accomplish research tasks that were previously inconceivable without a small army of research assistants. However, the challenge of these methods lies less in their implementation and more in their validation. Many CTA methods are non-deterministic, meaning that they produce slightly different output each time they are used. Additionally, a subset of CTA methods can be described as "black box" approaches, i.e. the rules they use to produce output are not intelligible to humans. Thus the call from Grimmer (2013) to "validate, validate, validate."

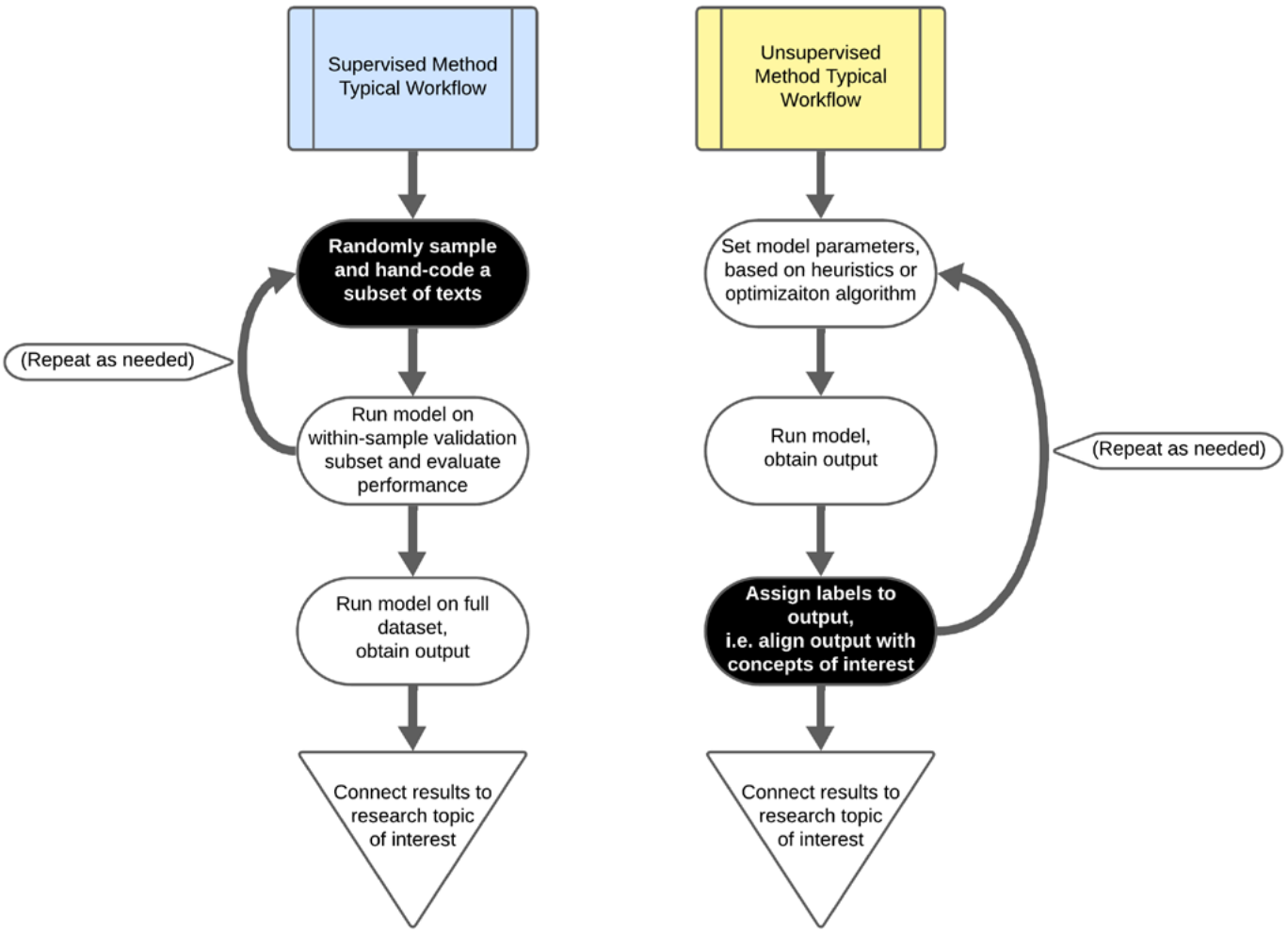
CTA practitioners have developed a number of methods and metrics for validating the work they do. Taking the lead from computer scientists, the vast majority of these validation metrics focus on the performance of the models being used. Others have recognized the need for a validation framework that expands beyond the work done by the computer and incorporates facets of human comprehension and understanding. (See for example Quinn, 2010; Chang, 2009; and Ying, 2022.) This paper expands existing work on validation of CTA methods in two ways. First, I argue that CTA as a method contains qualitative elements, and therefore existing work on validity by qualitative scholars is informative for improving the validity of text-based measurements. Second, I introduce a new conceptual framework for thinking about the validity of text-based measures.

Qualitative Features of CTA

Why turn to the humanistic social sciences when dealing with heavily quantitative methods rooted in computer science and statistics? In spite of the name, I argue that CTA methods are actually better characterized as fundamentally mixed methods, with clear qualitative elements. In their handbook on qualitative research, Corbin and Strauss (2014) define it as "a form of research in which a researcher(s) or designated coresearcher(s) collects and interprets data, making the researcher as much a part of the research process as participants and the data they provide." They

also argue that qualitative research must be characterized by openness and flexibility. Two key elements of this definition deserve emphasis when it comes to CTA. First is the idea of interpretation. Though CTA methods have been celebrated as a "neutral" way to extract meaning from large amounts of text, they still require researchers to make interpretive decisions. These interpretive decisions take place at different stages of the research process depending on the method chosen, but at some point a human researcher must use his or her own expert understanding of a concept to make meaning from the method. Second is the concept of flexibility, or iteration (Yom, 2015). When implementing CTA methods, practitioners must "fine-tune" their models. In some cases this involves an iterative coding and training process in which a subset of text is hand labeled. In other cases the process is more focused on adjusting statistical parameters to achieve a stable or intelligible result. Though the order of the steps may differ slightly, the core features of human interpretation and iteration remain the same. Figure 1 provides on overview of two typical CTA workflows with the interpretive steps highlighted in black and the iterative process shown using arrows.

Figure 1. Distinct Workflows in CTA



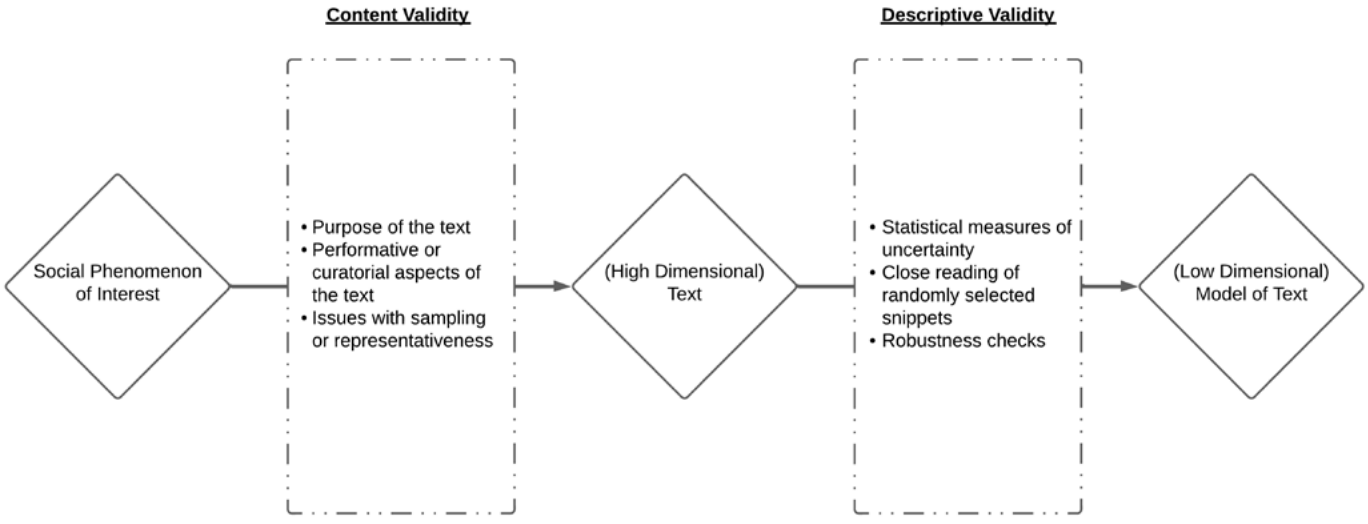
Based on the criteria discussed above, I argue that CTA methods contain qualitative elements and are best characterized as mixed methods, in spite of their use of complex statistical algorithms and need for high computing power. Therefore, the core question that arises is this: How can researchers make valid and replicable inferences from text? More specifically, when researchers make interpretive decisions, what practices can they employ to ensure that such decisions are not influenced by the outcome they seek in their data? Ultimately, this is a question of *validity* in text as data research.

Validity in CTA

CTA practitioners have embraced various methods for validating their measures. Models of text developed by computer scientists and statisticians often come with recommended methods for calculating uncertainty and fine-tuning results. Such statistical techniques have been key to improving CTA results and ensuring that papers using CTA are better able to connect to and build on existing bodies of research that address similar substantive topics but utilize non-text data sources. However, a more robust discussion of validity in CTA papers must recognize the fundamental challenge of using text as data: double abstraction.

Double abstraction occurs when the researcher is doubly removed from the concept of interest. Unless the text itself is the concept of interest, the researcher seeks to use the text to describe or measure a particular social phenomenon. The phenomenon occurs naturally, and the text is produced via some data generating process. This is the first level of abstraction. Subsequently, in order to make the text usable, tractable, and quantifiable, the researchers must model the text. This introduces a second level of abstraction. Validity is a concern at *both* levels of abstraction. (For a visual depiction of this methodological challenge, see Figure 2.)

Figure 2. Double Abstraction in CTA



Existing work on validity for CTA methods has focused almost entirely on the second level of abstraction depicted in Figure 2. I have termed this *descriptive validity* because it defines the degree to which the model accurately describes the raw text. Very few researchers write about the first level of abstraction, which I will refer to as *content validity* because it defines the degree to which a body of text fully represents some social phenomenon of interest. This weakness must be rectified for CTA to achieve its full potential as a social science method and contribute more effectively to existing substantive discussions in the social science literature.

To improve validity in CTA, practitioners should ask two key questions as part of their research practice: (1) Is my text corpus an accurate measure of my concept of interest? (2) Is my text model an accurate measure of my text? In answering these questions, in addition to the host of existing validation methods, I recommend that CTA practitioners consider the robust body of work on validity by qualitative researchers. A number of qualitative techniques are particularly relevant, including concept development, dictionary creation, back translation, and grounded theory.

Conclusion

This paper engages with the literature on validity and, more specifically, how it applies to CTA research. I argue that CTA methods are partially qualitative in nature, due to their interpretive and iterative processes. I then develop a new framework for validity in CTA research that emphasizes the challenge of double abstraction. In conclusion, I argue that CTA research would benefit from employing some qualitative validation methods to address the challenge of double abstraction.

References

Chang, J., S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei (2009). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems 22.

Corbin, J. and A. Strauss (2014). Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage publications.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis 21(3), 267–297.

Kerlinger, F. N. (1966). Foundations of behavioral research. New York.

Kvale, S. (1995). The social construction of validity. Qualitative inquiry 1(1), 19–40.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science 54(1), 209–228.

Ying, L., J. M. Montgomery, and B. M. Stewart (2022). Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. Political Analysis 30(4), 570–589.

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

QMMR Qualitative & Multi-Method Research

Original Article

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062851

Building Historically Oriented Datasets: A Practical Guide

Erica De Bruin
Hamilton College

Kristen A. Harkness
University of St. Andrews

Jun Koga Sudduth
Indiana University Bloomington

Introduction

Political science has been experiencing an upsurge in the construction of original, historically-oriented datasets. Whether cross-national or subnational, these data collection efforts are increasingly recognized as invaluable to the field—providing resources of enormous long-term benefit and enabling richer empirical analysis of a broader range of research questions.[1] There is a large and rich literature on conceptualization and measurement in the social sciences, which provides an important foundation for such dataset construction (e.g., Adcock and Collier 2001; Goertz 2006; Sartori 2009; Schedler 2012). There is also work that identifies best practices or principles scholars should follow in building historically oriented datasets—those that involve "the translation of narrative records of events and processes into numerical scores in the form of indexes, scales, or dummy variables indicating the presence or absence of some trait, such as regime type or degree of conflict" (Lieberman 2010, 39; Salehyan 2015).[2] At a pragmatic level, however, there are few resources on *how* to collect, code, and document historically oriented data.

To aid scholars embarking on efforts to build such historical datasets, this article presents a practical guide for dataset construction. It is geared towards graduate students, early career scholars, and others without extensive financial resources or a large research team. It may also be useful to scholars piloting data collection processes in order to secure funding for subsequent scaling-up. Our goal is to make transparent the choices and trade-offs involved during the data generation process, while providing helpful tips and highlighting pitfalls to avoid.

We provide a sequenced approach for historical data construction. We first discuss how to limit the scope of data collection to make it feasible for researchers with time and budgetary constraints. We then discuss sources and the trickiness of coding variables from collated narrative

1 The limitations and biases of such data are also quite well-known (see e.g., Bagozzi et al. 2019; Dietrich and Eck 2020; Gohdes and Price 2013; Hug 2003; Weidmann 2015).

2 Lieberman (2010), for example, emphasizes the principles of proximity of observations, transparency in citations, certainty of the historical record, and attention to valid comparison.

information, advocating pilot projects and a flexible, iterative process of refinement. Finally, we identify principles that should guide coding documentation and describe how to manage what can be an overwhelming amount of material by creating good workflow processes. Throughout, we draw upon our own experiences, failures, and hard-won lessons learned in constructing cross-national time-series datasets.[3]

Determining the Scope of Your Dataset

Embarking on historical data collection is daunting. Political processes often have no clean beginnings, neat endings, or clearly defined boundaries. Ideally, many of our projects would code hundreds of variables far back into the mists of time, include every conceivable unit of observation, and disaggregate as much as possible. But that task would never end. Thinking sensibly about your own time is important, especially for early career scholars. Job market, tenure clock, and promotion processes create extraordinary pressures to publish quickly. But only the fortunate few have sufficient institutional resources or grant funding to hire large teams of research assistants to gather and code data for a dissertation or first book.

Data collection becomes far more manageable if you can limit its scope. Here, the tricky part is balancing such reductions against the validity and richness of your empirical analyses—you want to ensure that your scoping decisions will still ensure you have leverage over the type of variation that is most relevant for answering your research questions and generating or testing your theories.

Scoping decisions are also fundamentally intertwined with operationalization—the art of turning theoretical concepts into observable and measurable indicators. Operationalization is well-covered in the existing literature, and we would strongly recommend a thorough read (see, for example, Adcock and Collier 2001; Collier, LaPorte, and Seawright 2012; Goertz 2006; Pepinsky 2007; Sartori, 1970). From a practical standpoint, we would also emphasize, before developing your own coding guidelines, first looking at how other scholars have previously operationalized your concepts and closely related ones. This allows you to understand what work has already been done, standard practices in the field, and how other scholars have thought through tricky issues.

In our collective experiences, we have used four strategies to limit the scope of data collection: (1) winnowing variables; (2) limiting temporal or geographical coverage; (3) randomly sampling cases; and (4) reducing granularity. There are both benefits and risks to each.

Winnowing variables. Conceptualizing, measuring, gathering information about, and coding a small handful of variables —as opposed to dozens or hundreds— is intuitively less time-consuming. Such laser focus, however, depends heavily on well-developed theory. You must first identify a narrow set of independent and dependent variables that will enable compelling empirical tests. This includes thinking through how to potentially test causal mechanisms and/or alternative explanations. The greatest risk of this strategy is narrowing too early. It is far easier, from the outset, to track multiple types of information from a set of sources than to realize later that you must add another key variable and need to return to your source material to do so. Pilot projects can help thread this needle, as discussed later, allowing the researcher to start broader and then clarify which variables may be too difficult or time consuming to code.

Limiting temporal or geographic coverage. Every project makes decisions over time and space. Strong theory, and careful consideration of variation, not only strengthen research design but can conserve resources as well. What type of variation will give you the most leverage over your research questions: across cases, or within cases, over time? To the extent that you want to leverage temporal variation, do you need annual, monthly, or daily data? Or would snapshots at specific historical

3 These include datasets on military purges, ethnic stacking in Africa, and features of state security and police forces (e.g., De Bruin 2021, 2022; Harkness 2022; Sudduth 2021).

moments suffice? Is there a narrower time period that would make for an ideal test of the theory or capture the most important variation?

Keep in mind that you can expand coverage after you have demonstrated the value of your data. In her book on building and dismantling ethnic armies, for example, Kristen focused her data collection efforts on the immediate post-colonial period and the third wave of democratization (Harkness 2018). Only later —with more time and more funding— did she build a comprehensive cross-national time-series dataset (Harkness 2022). Similarly, for her pilot study of civilian elite purges, Jun limited her initial data collection to the years 1985-1996. This allowed investigation of possible systematic changes in purge trends since the end of the Cold War, while reducing data collection time.

Also worth considering is the nature of real-world variation. For instance, if there is little temporal variation, the value of collecting time-series data is much less clear. Erica found that coding features of police forces in the first and last years of counterinsurgency campaigns took a fraction of the time than coding annually without much loss in data richness, as very few police forces underwent significant organizational change in between (De Bruin 2022). And it may not be worth coding data before or after the temporal ranges of other datasets containing variables you —or the users of your data— may frequently use.

Similarly, narrowing to a specific geographic region or sub-region can make data collection more manageable. This was the strategy Kristen used in her dataset on ethnic stacking; the dataset focuses on Africa, where the recruitment and promotion of coethnics has been a crucial component of rulers' efforts to control their militaries (Harkness 2022). It has the added benefits of limiting conceptual stretching and honing the researcher's contextual knowledge in a beneficial way. Deep regional or country expertise can also generate public engagement and impact opportunities. Yet, a narrow geographic focus has important implications for generalizability. Other scholars may perceive your theory as only relevant to the region where it was empirically tested. Your data will also have diminished utility to other scholars who either focus their work on a different region or want to conduct global analyses. Finally, some journals discourage regional datasets, limiting —but by no means eliminating— publication outlets.[4]

Randomly sampling cases. Another strategy is to only collect data on a manageable subset of randomly selected cases, as Erica did with the State Security Forces Dataset (De Bruin 2021). One could choose every sixth country listed alphabetically or use a random number generator to select observations. With this approach, geographic and temporal limits are minimized, allowing for better integration with existing (especially cross-national) datasets, and inferences can still be drawn to the wider population. But there are also drawbacks. A random sample may end up excluding important cases within the literature to which your project speaks. It could also require you to develop contextual expertise on perhaps an uncomfortably broad set of cases. Finally, concerns over missing observations may still compel future studies to draw their variables from other datasets —regardless of improvements in reliability or validity that your dataset might be able to make.

Reducing granularity. Finally, coding simple binary variables or categorical variables with a small number of outcomes is usually less time consuming than coding continuous variables. It also enables wider coverage and higher recovery rates, minimizing missing data. For example, Jun found it easier to code whether a leader purged officers from the military at all (0 or 1) than to measure consistently and reliably how many officers were purged (a continuous variable) (Sudduth 2017). Some granularity could be preserved, without too much additional work, by creating categories for different levels of purges (e.g., none, less than ten, 11 to 100, more than 100) (Sudduth 2021). Again, it is worth thinking hard about whether fine-grained distinctions between categories will map onto your theoretical constructs in a meaningful way,

4 For example, the *Journal of Peace Research* has historically tended to publish datasets with global coverage, as well as subnational datasets; others, such as *Conflict Management and Peace Science*, have welcomed those with a regional focus.

or whether you can do without them. The trade-off is information loss —missing details that could be important for later empirical tests and for ensuring other researchers, testing different theories and causal mechanisms, find the data useful.

Regardless of how you ultimately balance this trade-off, we recommend capturing the most granular information that you can find in your source material at the beginning of a project. It is much easier to code at higher levels of abstraction, given a rich information base, than to go back and re-collect more detailed information later in the project. Furthermore, more detailed information can help you understand processes and dynamics better, as well as provide useful anecdotes or serve as the basis for qualitative case studies.

Identifying and Selecting Source Material

Once you know the general scope of the data you need to collect, the next step is to think about sources. Rather than delve immediately into primary sources, we have found it valuable to first gather what information you can from existing academic scholarship (particularly country-specific historiography and ethnography), historical dictionaries, annual surveys, and other such resources. Then you can strategically think through the gaps that need filling with more time-intensive primary sources.

It is important to remember that all sources are subject to data generating processes that create biases in what information is recorded and preserved. News sources are easy to access as most academic institutions subscribe to searchable online databases, including LexisNexis and Keesings, which provide broad global coverage, as well as national and local newspapers. However, they are shaped by audience appeal and may be biased in favor of particular countries, urban areas, and the most "news-worthy" or visible events (Croicu and Eck 2022; Dietrich and Eck 2020; Parkinson 2024; Weidmann 2015). Historical archives can provide astoundingly rich and unique information, enabling greater disaggregation and more direct capture of key concepts. However, access to archives can be tricky and, at a minimum, requires the time and funding to travel to the location of preservation.[5] Archives also contain their own biases, which reflect the incentives of actors recording and releasing the records (Balcells and Sullivan 2018; Lee 2022), issues likely exacerbated by the (often partial) digitization of archival collections (Kim 2022). Whatever the sources you use, we would thus echo Salehyan's (2015) encouragement to think hard about what may be missing from them, and to triangulate contested information from multiple sources wherever possible.[6]

Using Pilot Projects Effectively

Once you have developed initial ideas of how you want to measure your concepts (e.g., coding guidelines) and where you will find relevant historical information (sources), we strongly recommend embarking on a pilot coding project.[7] Pilot projects enable you to iteratively refine your processes of data collection and coding and reflect on measurement validity: do the numerical scores produced map back onto the original concepts and theoretical constructs in a meaningful way? Could revisions to coding practices improve such validity? Carefully crafted pilot projects generate confidence in your approach before sinking extensive time and resources into a particular coding scheme.

We have found it helpful to strategically select a small sample of observations to code that will "stress test" your initial coding rules by making you consider the full range of outcomes on your variables, highlight gray zones between coding categories, and probe the limits of source material.

5 For a thoughtful discussion of the politics of archives, and other ethical issues in archival research on political violence, see Subotić (2021).

6 Both Lee (2022) and Kim (2022) provide excellent guides to how to anticipate and help mitigate biases in different forms of archival research.

7 Budget allowing, it may be helpful to hire one or two research assistants at this stage to help you pilot. You can provide them the guidelines and the same set of sample cases to check the extent to which their codings converge.

Sample cases from different geographical regions or sub-regions and from different historical eras, decades, or years. Pick some difficult cases that do not easily fit your concepts and chosen measures. Include some observations with a wealth of historical materials —an overabundance of sources that will generate differing and sometimes contradictory interpretations. But also include observations with a dearth of sources that will stretch your capacity to find enough material and help highlight which variables it might not be feasible to collect data on across your cases.

Pilot projects usually reveal useful time-saving lessons such as coding rules that need modification and variables that cannot be systematically coded across cases. Keep a flexible mindset and remind yourself that this is the whole point of the pilot — to find the problems and improve measurement validity by iteratively modifying and re-test your approach. You can re-pilot as many times as needed. Then, when you feel confident in your processes and coding guidelines, scale up and tackle the rest of the data collection project.

A final tip: Keep extensive notes for yourself during your pilot. Track which variables were most challenging to find sufficient information on; where you had trouble coding the information you found; sources that were particularly useful (or turned out to be irrelevant); and where interesting and important features of the cases were not yet captured by your coding scheme. Your future self will be very grateful for detailed tracking of your own thoughts and decisions over tricky issues and difficult cases.

Documenting Your Coding Decisions

In documenting your coding decisions, the aim is to ensure transparency and, to the extent possible, replicability. Transparency centers on creating a trail of documentation that allows others to understand each step of how you built the data, the decisions involved, and the sources consulted.[8] It gives meaning to the numbers, enabling good inferences. Replicability sets an even higher bar. Some would argue that after reading your full documentation, other scholars should be able to reach the same coding decisions as you did, in the vast majority most cases. This is a high bar indeed, and one that may overlook the unavoidable messiness of social data.

There will always be debates over data subjectivity, as human judgment is fundamental to translating often incomplete, complex, and sometimes contradictory narrative records into numerical scores. Take the recent debate over whether and to what extent democracy is backsliding globally. Little and Meng (2024, 151) highlight a discrepancy between "objective" indicators of democracy, which they describe as "based in fact," and "subjective" indicators, which depend on a combination of fact and coder judgement. However, even seemingly objective indicators often require multiple judgments by human coders (Schedler 2012). Knutsen et al. (2024) provide the example of coding parliamentary vote share —while theoretically fact-based, coders must still make choices about how to code independents, which electoral rounds to consider, and other issues.

Limitations in source material may also produce ambiguous or conflicting information that must be adjudicated. Jun found this to be the case in coding military purges, which she defined as leaders' actions to dismiss, demote, expel, arrest, or kill individuals within their security apparatus. In some instances, there was evidence that top-ranking officers left their positions unexpectedly, but whether they resigned voluntarily or were forced out remained unclear (Sudduth 2021).

More generally, many of the concepts most integral to the study of politics are deeply normatively important, and thus essentially contested (Gallie 1955-56). As a result, the expectation should not be that our codings will be accepted uncritically. Indeed, debates over coding can be instrumental in advancing our understanding of complex political phenomena. The debate over whether the January 6

8 Many political science journals now subscribe to Data Access and Research Transparency (DA-RT) principles, which aim to increase transparency in social science (https://www.dartstatement.org/).

attack on the U.S. Capitol counted as a coup d'état, example, drew attention to a growing divergence between academic and journalistic uses of the term; the extent to which major coup datasets disagreed about specific cases (Chin, Carter, and Wight 2021); and the need to better map the conceptual terrain of other types of antidemocratic actions, such as "self-coups," which share some features with coups but also diverge from them in important ways (Powell et al. 2022). What good coding and documentation practices do allow us to do is minimize unnecessary disagreement by making explicit the choices we have made and enabling other scholars to evaluate for themselves whether they would come to the same decisions about our cases.

Managing Documentation and Workflow

Handling all the source materials, citation information, qualitative narratives, coding guidelines, and spreadsheets involved in these projects, just for one's own personal use, is a huge task. We cannot stress enough how vital it is to develop good systems, right from the beginning, to track and organize your project. Fixing mistakes can be costly and time consuming. When copy-editing her book, for instance, Kristen had to return to archives she visited as an inexperienced graduate student to fix three footnotes that were missing some small but essential piece of citation information (e.g., part of a document title, the recipient of a letter).

While individualized, the systems we have each developed to manage documentation and workflow revolve around four types of documents that we combine and layer: (1) coding guidelines; (2) source notes; (3) case or coding notes; and (4) data spreadsheets.

First, coding guidelines (the "codebook") contain the final rules for data coding. They should be ultimately published alongside the dataset. For each variable, they should describe the underlying concept and how it has been measured. The best coding guidelines also discuss why you made certain decisions and what reasonable alternatives were considered (but ultimately dismissed). Your thinking will change over the course of the project given the iterative process of flexibility and refinement that we suggest imbues all historic data creation. This makes it crucial to keep "in progress" notes and preserve any iteration of coding guidelines used to produce coded data (linked through version numbers to the associated spreadsheets). Draft coding guidelines can also be shared with advisors, mentors, and colleagues for feedback before the final, cleaned-up version is publicly released.

Second, *source notes* are an important intermediary step between learning relevant information and using it to code your data. This step is tempting to skip. However, for data construction that relies on primary sources or requires more interpretation, it is very useful to preserve nuggets of information in the broader contexts in which they were found. Source notes are especially useful for tracking information from archival documents, interviews, ethnography, newspaper articles, and social media posts.[9] Source notes are also a place to record overly detailed citations, interesting tidbits whose relevance is not immediately apparent, and one's own evolving thinking. The supplementary material included online provides an example of Kristen's source notes on an archive dossier.

Third, *case* or *coding notes* compile information across multiple sources for a particular case or observation. They can be in any format —from polished narratives to bullet points to a series of questions and answers— but should be rationally organized around the original variables that need coding. The idea is to collate, for that observation, all the various relevant pieces of information from the underlying sources. Citation tracking is critical, and we recommend refraining from any clustering: cite at the sentence level and keep sources distinct even where they contain the same basic fact.

9 Ethical and copyright considerations permitting, you may also consider publicly releasing primary source images, notes, or links along with the dataset.

We also endorse tracking "consulted" sources, including those that you do not draw from, so you avoid rereading them. Erica learned this the hard way when she came across her own hand-written post-it note in a book which, it turned out, she had requested via InterLibrary Loan and consulted a year earlier.

Developing a template for case study notes helps guide data collection (and is vital for employing research assistants). Questions to consider when designing a template include: What background information would help understand the context of this case? What do you need to know to code each variable? Is there other information you want to track, even if it cannot be compiled systematically? The extra material often provides a rich resource for later publications, including anecdotally illustrating theoretical points or fleshing out qualitative case studies. For maximum transparency, case study notes can be published alongside the dataset—either in template form or after rewriting into more polished narratives. This renders visible the underpinning information and justifications for why variables obtained the codings they did. The supplementary material includes three examples of different ways to structure case study notes from Erica and Jun.

Finally, *data spreadsheets* record the numeric or categorical codings assigned to each observation. In their final form —you may go through several pilot or interim versions— they provide you and other potential users with the actual quantitative data. You might produce one large spreadsheet, or several data tables linked together through identifiers. For example, datasets are often published in different versions to distinguish between related units of analysis (e.g., conflict actor and conflict-spell, leader-spell and time-series, or event and time-series).

There are many valuable, published resources on good data management practices (see, for example, Weidmann 2023). Just a few pointers to keep in mind from our experience: Use existing coding conventions for identifying countries, conflict actors, etc., to make merging your dataset with others easier. Try to ensure variable names have enough content for you and future users to understand and remember what they capture without constantly referring to the codebook (e.g., v1, v2, v3 are not that helpful, nor are self-invented acronyms). Do some good spot-checking and quality control at the end—it is easy to make errors while inputting loads of data into a spreadsheet. You can also set up some automated quality control checks within your spreadsheets (e.g., if the variable is supposed to be 1-5, and there's a value of 10, you know it's an error). Finally, generate a comma-separated values (CSV) file for sharing purposes that is easy to use with any software program.

Sharing and Updating Your Data

When you are ready to share your data, posting it in an online data repository like the Harvard Dataverse (https://dataverse.harvard.edu/) increases data accessibility beyond your own networks and allows you to track the number of downloads. You can also choose to gather information on who uses your data by requesting names, institutions, and positions from users before they download your datasets. However, keep in mind that requesting this information may create a barrier to access for users uncomfortable with sharing such personal data, so consider how important it is to you.

Finally, while one may be tempted to hang the victory banner at this stage, it is useful to consider how you want to process queries or disputes, as well as how you will revise and release updated versions of your dataset. Periodic updates may be necessary to extend your data collection forward in time. New records may be declassified, or more detailed historiographies published, which change your interpretation of cases. And ideally, lots of other scholars will use and scrutinize your data, which will bring the occasional data entry or coding error to light, even where it does not provoke broader debate over your indicators. As a field, we should think more deeply about how to support periodic cleaning and updating of our data resources —an important task for which little financial support is typically available. In the meantime, developing your own plan for revisions will help ensure your data remains relevant to users.

Conclusions

The construction of original, historical-oriented datasets allows political scientists to document patterns and to develop and test theories about the political world. Our aim, in this article, was to provide a practical guide to building datasets that highlights important tradeoffs and provides advice to assist researchers undertaking such efforts for the first time. We hope that it helps get more data projects off the ground.

Acknowledgements

For helpful feedback, we are grateful to an anonymous reviewer, Carolyn Holmes, Nils Weidmann, Agnes Yu, and other participants in the 2022 American Political Science Association Annual Meeting panel on "Critical Challenges and Solutions Towards (Re)Constructing Conflict Datasets."

References

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95, no. 3 (September): 529–46.

Bagozzi, Benjamin E., Patrick T. Brandt, John R. Freeman, Jennifer S. Holmes, Alisha Kim, Agustin Palao Mendizabal, and Carly Potz-Nielsen. 2019. "The Prevalence and Severity of Underreporting Bias in Machine- and Human-Coded Data." *Political Science Research and Methods* 7, no. 3 (July): 641–49. https://doi.org/10.1017/psrm.2018.11.

Balcells, Laia, and Christopher M. Sullivan. 2018. "New Findings from Conflict Archives: An Introduction and Methodological Framework." *Journal of Peace Research* 55, no. 2 (March): 137–46. https://doi.org/10.1177/0022343317750.

Chin, John J., David B. Carter, and Joseph G. Wight. 2021. "The Varieties of Coups D'état: Introducing the Colpus Dataset." *International Studies Quarterly* 65, no. 4 (December): 1040–51.

Collier, David, Jody LaPorte, and Jason Seawright. 2012. "Putting Typologies to Work: Concept Formation, Measurement, and Analytic Rigor." *Political Research Quarterly* 65, no. 1 (March): 217–32.

Croicu, Mihai, and Kristine Eck. 2022. "Reporting of Non-Fatal Conflict Events." *International Interactions* 48 (3): 450–70.

De Bruin, Erica. 2021. "Mapping Coercive Institutions: The State Security Forces Dataset, 1960-2010." *Journal of Peace Research* 58, no. 2 (March): 315–25.

De Bruin, Erica. 2022. "Policing Insurgency: Are More Militarized Police More Effective?" *Small Wars & Insurgencies* 33(4-5): 742–66.

Dietrich, Nick, and Kristine Eck. 2020. "Known Unknowns: Media Bias in the Reporting of Political Violence." *International Interactions* 46(6): 1043–60.

Gallie, W. Bryce. 1955-1956. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 56: 167–98.

Goertz, Gary. 2006. *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.

Gohdes, Anita, and Megan Price. 2013. "First Things First: Assessing Data Quality before Model Quality." *Journal of Conflict Resolution* 57, no. 6 (December):1090–108.

Harkness, Kristen A. 2018. *When Soldiers Rebel: Ethnic Armies and Political Instability in Africa*. New York: Cambridge University Press.

Harkness, Kristen A. 2022. "The Ethnic Stacking in Africa Dataset: When Leaders Use Ascriptive Identity to Build Military Loyalty." *Conflict Management and Peace Science* 39, no. 5 (September): 609–32.

Hug, Simon. 2003. "Selection Bias in Comparative Research: The Case of Incomplete Data Sets." *Political Analysis* 11, no. 3 (Summer): 255–74.

Kim, Diana S. 2022. "Taming Abundance: Doing Digital Archival Research (As Political Scientists)." *PS: Political Science & Politics* 55, no. 3 (July): 530–8.

Knutsen, Carl Henrik, Kyle L., Marquardt, Brigitte Seim, Michael Coppedge, Amanda B. Edgell, Juraj Medzihorsky, Daniel Pemstein, Jan Teorell, John Gerring, and Staffan I. Lindberg. 2024. "Conceptual and Measurement Issues in Assessing Democratic Backsliding." *PS: Political Science & Politics* 57, no. 2 (April): 162-77.

Lee, Alexander. 2022. "The Library of Babel: How (and How Not) to Use Archival Sources in Political Science." *Journal of Historical Political Economy* 2 (3): 499–526. http://dx.doi.org/10.1561/115.00000038.

Lieberman, Evan S. 2010. "Bridging the Qualitative- Quantitative Divide: Best Practices in the Development of Historically Oriented Replication Databases." *Annual Review of Political Science* 13: 35–59.

Little, Andrew T., and Anne Meng. 2024. "Measuring Democratic Backsliding." *PS: Political Science & Politics* 57, no. 2 (April): 149-61.

Parkinson, Sarah E. 2024. "Unreported Realities: The Political Economy of Media-Source Data." *American Political Science Review* 118, no. 3 (August): 1527-32.

Pepinsky, Thomas B. 2007. "How to Code." Political Concepts Working Paper #18, IPSA Committee on Concepts and Methods, https://www.concepts-methods.org/Files/WorkingPaper/PC%2018%20Pepinsky.pdf.

Powell, Jonathan, Salah Ben Hammou, Amy Erica Smith, Lucas Borba, Drew Holland Kinney, and Mwita Chacha, and Erica De Bruin. 2022. "Forum: A Coup at the Capitol? Conceptualizing Coups and Other Anti-Democratic Actions." *International Studies Review* 24, no. 1 (March): viab062. https://doi.org/10.1093/isr/viab062.

Salehyan, Idean. 2015. "Best Practices in the Collection of Conflict Data." *Journal of Peace Research* 52, no. 1 (January): 105–9.

Sartori, Giovanni. 1970. "Concepts and Misinformation in Comparative Politics." *American Political Science Review* 64, no. 4 (December): 1033–53.

———. 2009. "Guidelines for Concept Analysis." In *Concepts and Methods in Social Science: The Tradition of Giovanni Sartori,* edited by David Collier and John Gerring, 97–150. New York: Routledge.

Schedler, Andreas. 2012. "Judgment and Measurement in Political Science." *Perspectives on Politics* 10, no. 1 (March): 21–36.

Subotić, Jelena. 2021. "Ethics of Archival Research on Political Violence." *Journal of Peace Research* 58, no. 3 (May): 342–54.

Sudduth, Jun Koga. 2017. "Strategic Logic of Elite Purges in Dictatorships." *Comparative Political Studies.* 50, no. 3 (November): 1768–801.

Sudduth, Jun Koga. 2021. "Purging Militaries: Introducing the Military Purges in Dictatorships (MPD) Dataset." *Journal of Peace Research* 58, no. 4 (July): 870–80.

Weidmann, Nils B. 2015. "On the Accuracy of Media-Based Conflict Event Data." *Journal of Conflict Resolution* 59, no. 6 (September): 1129–49.

Weidmann, Nils B. 2023. *Data Management for Social Scientists: From Files to Databases*. New York: Cambridge University Press.

QMMR Qualitative & Multi-Method Research

Original Article

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062861

"An Alternative to Being Shot": Using Archives to Understand Government Decision-Making

Sarah K. Dreier
University of New Mexico

Emily K. Gade
Fors Marsh

Introduction

How do government officials make controversial decisions? In 1971, British Prime Minister Edward Heath and the executive agencies of the British government (HMG[1]) authorized its provincial government in Northern Ireland to adopt internment without trial in an effort to quell the escalating "Troubles" over the region's sovereignty.[2] Researchers often presume that governments' public explanations for due-process violations —in this case (and in many others) declared necessary to safeguard against national security threats— do not fully represent government actors' true motivations, rationales, or attitudes.[3] For example, in contrast to HMG's public pronouncement of internment as abhorrent and regrettable (PREM15/478:179-81), Northern Ireland's Prime Minister, Brian Faulkner, offered an insouciant private discussion of internment as a better alternative to more drastic state measures: "Internment should properly be regarded not so much as an alternative to trial as an alternative to being shot in an act of terrorism" (PREM15/482:6175).[4]

Yet rarely present for those discussions, researchers face considerable obstacles to identifying government actors' true motivations behind controversial policy decisions. As a result, researchers are often left to use incomplete information, generalizable theories of rational behavior or human psychology, or blunt computational text-as-data tools on available data to model, identify, or speculate about these true motivations. Such approaches provide invaluable knowledge about government decision-making but require a level of abstraction that can obfuscate influential factors that drive policy outcomes in a particular case.

This article systematically analyzes declassified government correspondence documents to demonstrate how archive research expands our understanding of government decision-making during Britain's Troubles in Northern Ireland.[5] We examine HMG's decision to implement and

1 Here, Her Majesty Queen Elizabeth's Government (HMG) refers to Prime Minister Heath (1970-74) and his Cabinet.
2 Between 1971 and 1975, almost 2,000 people were interned without trial in Northern Ireland.
3 We examine a liberal democracy's restrictions on civil liberties. Davenport (2007) analyzes state repression among autocracies.
4 The Online Appendix (A2, available at https://www.qmmrpublication.com/issues) lists referenced government officials.
5 All archive data come from The National Archives in the United Kingdom (TNA); see the Online Appendix (A1, available at https://www.qmmrpublication.com/issues) for detailed archive information. See also Dreier and Gade (2024), which provides a detailed discussion of our broader project's methodological approach.

continue internment without trial, documented within 8,430 security-related pages from the British Prime Ministers' correspondence archives (1969-73). This novel collection offers a rare opportunity to observe the unvarnished processes by which government actors debate and arrive at controversial decisions to violate civil liberties.

This article provides real-world observations about how the British government decided to implement internment—observations that neither the state's public record nor generalizable theories alone could expect to sufficiently capture. Using this case, we illustrate how a comprehensive qualitative review of declassified archive records can help researchers: (1) identify important aspects of government decision-making (Lorentzen, Fravel, and Paine 2017; Patty and Penn 2014b); and (2) overcome some obstacles to observing government motivations. Given the proliferation of generalizable or computationally scalable approaches to analyzing government behavior, this article complements previous archive-based research (Blaydes 2018; Kim 2021; Lawrence 2013; Saunders 2017; Weld 2014) to demonstrate how historical records can provide contextual details, fill conspicuous omissions in the public record, and reveal hidden motivations that would otherwise remain unobserved.

Generalized Approaches to Modeling Government Behavior

Formal models explain a tremendous extent of government strategic behavior by deliberately abstracting from contextual circumstances (Cook and Levi 2008; Jervis 2017; Schelling 1980; Weingast et al. 2020). However, formal theory is less equipped to explicitly capture how government decision-making is also shaped by heterogeneous preferences, institutional constraints, limits to human cognition, and individuals' predispositions, normative commitments, and personalities (Hermann and Ozkececi-Taner 2011; Patty and Penn 2014b; Redlawsk and Lau 2013; Stevens 2019). Political psychology research—which identifies the effects of interpersonal relations, emotions, ideology, framing, risk aversion, and other heuristics that shape (arguably sub-optimal) outcomes (Druckman and McDermott 2008; Duelfer and Dyson 2011; Goldstein and Keohane 2019; Mercer 2013, 2014; Pearlman 2013; Tversky and Kahneman 1992; Vis and Kuijpers 2018)—is often conducted in labs or via surveys, removed from real-world contexts. Meanwhile, text-as-data tools often aggregate to coarse metrics to identify patterns across large available text collections (Acree et al. 2020; Gade et al. 2021; Grimmer and Stewart 2013; Huang, Perry, and Spirling 2020; Spirling 2016), sacrificing contextual meaning for large-scale explanation. Augmenting these approaches with comprehensive qualitative analyses of archives can reveal influential case-specific contexts or considerations in the decision-making process that are otherwise unknown, implausible to consider, or deliberately omitted (Kim 2022), provide caveats, nuances, and scope conditions (e.g., information about how real actors achieve equilibrium),[6] and ultimately contribute to —and even help correct—conclusions drawn from generalizable approaches.

Aspects of Government Decision-Making

Reviewing archival records can reinforce to researchers that governments are not unitary rational actors but instead conglomerations of discrete individuals with unique perspectives, emotions, and values. Government decisions result from intricate webs of discussion, debate, and dissent from idiosyncratic actors. We expect the following aspects of government decision-making to be true—at least to some extent—among most governments. These aspects will likely be particularly pronounced among liberal constitutional democracies that protect dissent and maintain multiple decision-makers to check and balance one another.

First, although the international relations cannon (Waltz 2010) and some expected utility models (Karni 2013) treat "the state" as a unitary

6 We thank Jess Sun for providing this point.

actor, and formal theory considers two- or three-actor negotiations (Putnam 1988), in reality government decisions involve numerous shifting and disagreeing actors, each with heterogeneous (and often competing) preferences, obligations, perceptions, constituents, and access to information (Groseclose and McCarty 2001).[7] Even when a single actor is ultimately responsible for a decision, those policy outcomes result from multi-actor domestic and international negotiations,[8] are idiosyncratic to time and context,[9] and can be directly shaped by unique players' exogenous, context-specific preferences—preferences which generalizable theories are not well-poised to intuit.

For example, Northern Ireland PM Chichester-Clark (Faulkner's predecessor) opposed introducing internment (PREM15/475:4846) —partially on constitutional grounds (PREM15/475:4848; PREM15/476:5439)— unless it was deemed necessary by HMG's security forces and military advisors (DEFE24/1214:2298). Chichester-Clark maintained these commitments despite anticipated "political pressure" (PREM15/476:5453), public "demand for internment" (DEFE24/1214:2298), and his possible loss of office (PREM15/476:5453). Chichester-Clark consequently resigned upon receiving an anticipated no confidence vote from his Unionist party, a vote which Chichester-Clark understood he could have circumvented by advocating for internment (PREM15/476:5453). Chichester-Clark was replaced by Faulkner, his co-partisan and former cabinet member. Faulkner initially appeared more open to considering internment (PREM15/477:5578), eventually decided that internment was appropriate and necessary (PREM15/478:255), and successfully convinced British Prime Minister Heath (the ultimate decision-maker) to authorize internment (PREM15/478:211-214). In short, Chichester Clark maintained exogenous preferences (e.g., commitment to constitutional principles) that Faulkner did not appear to share. These idiosyncratic actor preferences are notoriously hard to observe and measure, can have tremendous influence over policy outcomes, and can be discovered in qualitatively reviewed real-world data.

Furthermore, internal decision-making behaviors are routinely shaped by normative preferences, reputational concerns, and other psychological factors that are not easily generalizable or directly attached to beneficial or costly outcomes (Sheffer et al. 2018; Wilson 2011). For example, HMG officials give considerable attention to whether external actors perceived internment as a violation of liberal norms. With colonialism falling out of favor, officials sought to "avoid the impression that Ulster [Northern Ireland] was being treated as little more than a colony" (PREM15/100:4503),[10] even as HMG adopted policies based on "earlier experience in dealing with recalcitrant colonies" (PREM15/480:5636). Officials also demonstrated concern over public critiques from the Roman Catholic Pope, quickly pointing to protective procedures and internee release rates in an effort to convince the Pope that their policies did, indeed, align with standards of liberty and *habeas corpus* (PREM15/1013:8245).[11] Beyond concerns for such ideological disrepute, archives also demonstrate

7 Although government decisions made by groups of individuals with heterogeneous preferences are often sub-optimal, governments are adept at legitimating or justifying those decisions (Patty and Penn 2014a).

8 Negotiations over internment considered "repercussions not only of public opinion in Northern Ireland but internationally" (PREM15/475:4855) and involved parties within and outside Britain, including Faulkner, Ireland's Taoiseach Lynch, the European Commission, the U.S. President, the UN Security Council, competing perspectives within Heath's cabinet, and Heath's constituents.

9 Example: HMG authorized internment (primarily against Catholics) alongside a ban on (*de facto* Protestant) marches as a "propaganda" effort (PREM15/479:5885) to avoid appearing to exclusively target Catholics. This "package deal" (DEFE24/1214:2185) and its anticipated effects on specific populations are not easily generalizable or recognizable beyond this case.

10 "There is. . . a real danger that, unless we step up our propaganda effort, the myth that Northern Ireland is another British Colonial war will get firmly implanted in the American mind" (PREM/151003:6929).

11 Weeks after internment began, the Pope said the situation in Northern Ireland was "aggravated following the adoption of exceptional security measures which were strongly resented by at least part of the citizens" (PREM15/480:5630; also: PREM15/480:5654, PREM15/479:5821). HMG gave considerable attention to the Pope's upcoming Easter 1973 message: "It will be helpful for you to send a further message to the Pope explaining the present position on 'internment' —as he persists in calling it— and expressing the hope that the Pope will be able to make a public and welcoming reference to the [new proposed detention policies]" (PREM15/1692:9316).

priorities shaped by individual beliefs and interpersonal relationships, heated contention, social biases (Patty 2009), emotional concerns[12] or reactions,[13] and dialogical mistakes that may reveal latent attitudes.[14]

Obstacles to Observing Motivations Behind Government Decisions

The archived record of British Prime Minister Edward Heath's correspondences[15] also suggests a disconnect between some British officials' true motivations behind internment (as a political act aimed to appease constituent demands) and Britain's public explanation of those policies (as militarily necessary action against violent threats). This disconnect illustrates three obstacles to researchers' abilities to identify true motivations behind policy decisions: (1) officials can publicly misrepresent their motivations; (2) the misrepresented motivations can appear with greater frequency in the resulting record; and (3) officials can come to repeat, adopt and even believe the misrepresented motivations over time. Using the case of internment, we demonstrate how qualitatively reviewing archive records can help researchers overcome these obstacles.

Motivations Misrepresented in the Public Record

Limited access to internal debates often forces researchers to take public pronouncements at face value or as representative of at least some strategic actors' commitments (Dellis 2007). However, policymakers' explanations to the public cannot be presumed to reflect their actual motivations (Gailmard and Patty 2019; Penn, Patty and Gailmard 2011). In the case of internment, public explanations misrepresented influential internal motivations. Faulkner privately advocated to HMG to authorize internment in order to appease his Unionist party and Protestant constituents,[16] despite British defense officials' assessments that internment could be ineffective or counter-productive to security aims.[17] Days before initiating internment, British Prime Minister Heath and his ministers privately stated that "internment was a major decision, which could not be said. . . to be rationalized by any military necessity. It must therefore be regarded as a political act" (PREM15/478:210-13). Yet in preparing announcements to U.S. President Nixon and Ireland's leader, Taoiseach Lynch, HMG officials edited out Faulkner's political motivations and replaced them with abstract arguments implying that internment was militarily prudent. The archive record depicts these hand-written edits (Figure 1).[18]

12 "I am left with a feeling of dismay about the bareness of the landscape and the absence of any realistic prospect of making progress. . . I have an uncomfortable feeling that . . . we shall be driven to call in question some of [our] political and constitutional assumptions" (PREM/15480:5632).

13 "Few people are concerned any more with the truth, only with enjoying an orgy of self-righteous condemnation" (PREM15/1002:6853).

14 "I am sure that you detected the unfortunate misprint in the letter which I sent you on 2 August…:for "the British gunmen" read "the British Government" (PREM15/1012:7971).

15 Harold Wilson was prime minister during short portions of our observed time period, but Prime Minister Heath oversaw the internment policy decisions.

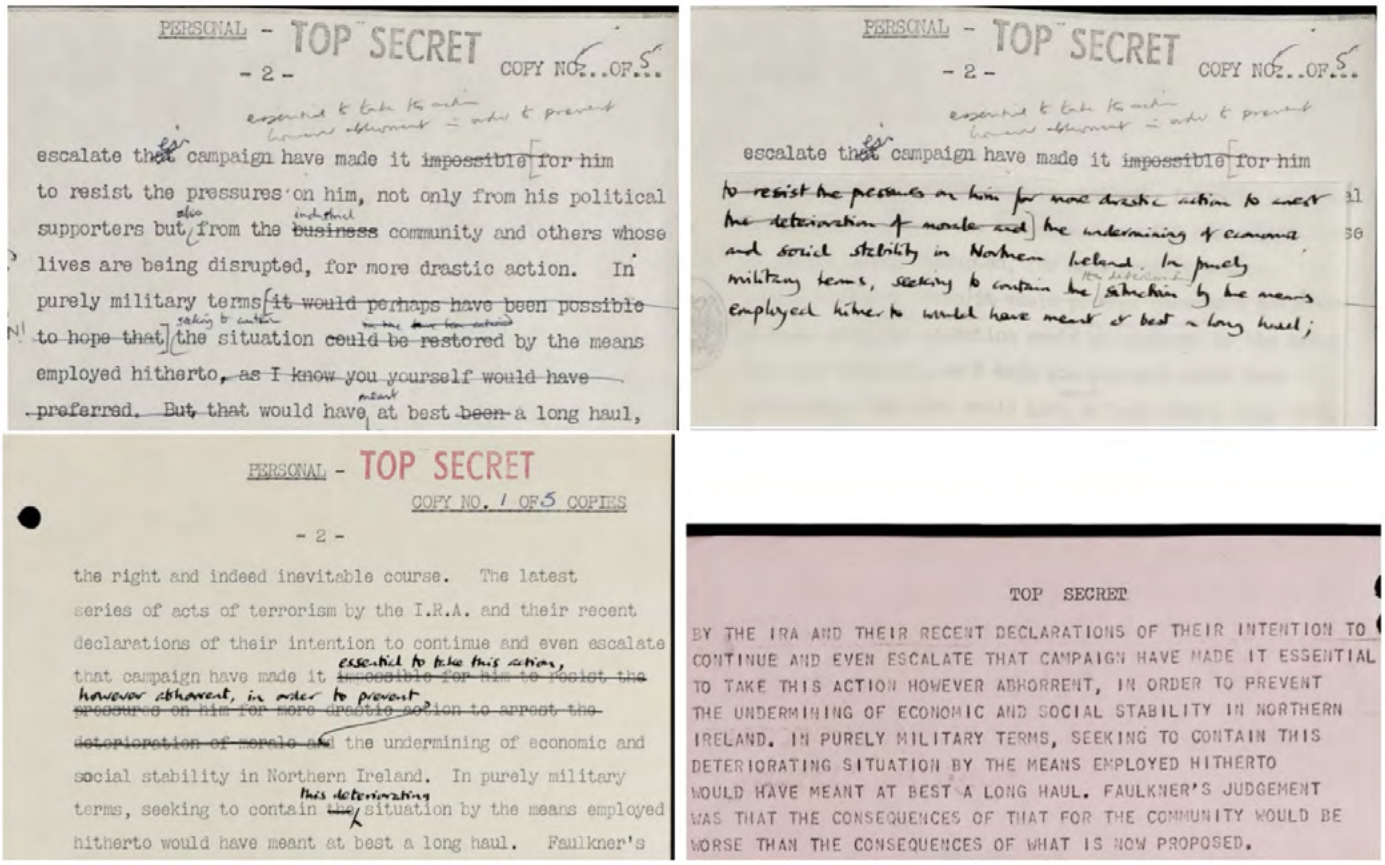
16 "You will see that emphasis has been put on the security arguments for the introduction of internment. Fact is of course that there is very strong pressure within the Unionist Party in the North upon Mr. Faulkner to take this further step" (DEFE/241214:2215; also: PREM15/476:5453, PREM15/475:4854).

17 "General [Officer Commanding the British Army in Northern Ireland] Tuzo still felt that the introduction of internment would have, on balance, a harmful effect on the security situation in Northern Ireland" (PREM15/478:224). HMG understood "it would not be possible to say that [internment] was being taken on the basis of advice from the security forces since. . . the GOC remained of the opinion that internment was not required on strictly military grounds" (PREM15/478:210-13, 175-76; see also PREM15/478:231).

18 Other examples of Britain's misrepresentation or deception include: devising a cover-up story for internment facilities to avoid damaging political and operational consequences ("[T]he real purpose of the accommodation must not be divulged in public"; DEFE24/1214:2262; also PREM15/475:4854) and erroneously claiming that no political adversaries were interned: "[N]o one is being held either because of his political views or because he opposes the Government" (PREM15/480:5589; also: PREM15/482:6174, PREM15/479:5872-73, PREM15/1002:6772). This claim was disputed and subsequently debunked (McCleery 2012, 428).

Although policymakers' public rationales often mirror their internal motivations (Khong 1992), in other cases—including internment in Northern Ireland—official public pronouncements appear to intentionally misrepresent policymakers' true motivations.[19]

Figure 1 Edits striking internment's political motivations and uncertain military merit from public announcements; final telegram: bottom right (PREM15/478:200, 199, 205, 176).



Source: TNA (1971); see also Online Appendix (A1, available at: https://www.qmmrpublication.com/issues).

Without these declassified records, Faulkner's true motivations would likely have gone unsubstantiated. Faulkner's own public testaments explicitly refute the suspicion that internment was politically motivated. In a televised address the morning internment began, Faulkner said, "I have taken this serious step solely for the protection of life and security of property in Northern Ireland. . . [I]t was not a step towards which I would be moved by any political clamour" (PREM15/478:051). Faulkner's subsequent memoir reiterated that "[t]he idea of arresting anyone as an exercise in political cosmetics was repugnant to me" (Faulkner 1978, 119). Three months after internment began, commentary published in the *The Times* criticizing internment (for lacking proper safeguards) nevertheless accepted Faulkner's and HMG's publicly presented rationale that

19 Another likely example of government actors misrepresenting motives: U.S. President Trump unprecedentedly ordered the U.S. Census Bureau to collect citizenship status. Despite its apolitical public presentation, Trump's objective was to gain partisan electoral advantages (Bazelon and Wines 2021).

"detention without trial may be necessary in time[s] of emergency" that pose "threats to national life" (PREM15/484:6311).[20] Even decades later, researchers could only speculate that political pressures motivated internment (McCleery 2015, 36), admitting that: "Faulkner's claim that internment had been introduced purely for security reasons seems somewhat dubious" (McCleery 2015, 45). In short, without observing the declassified archive record, researchers would likely lack a complete understanding of the motivations behind Britain's internment policies.

Misrepresented Rationales Appear with Greater Frequency

Next, although researchers often use frequency counts as evidence of a specific concept's relative importance —an approach amplified by computational text-as-data methods (Gade et al. 2021)— in reality, a concept that appears relatively infrequently can be disproportionately representative of policymakers' true motivations for a policy outcome (Ashworth, de Mesquita, and Friedenberg 2017). Although Faulkner was largely motivated by partisan constituent demands, such political motivations appeared relatively infrequently in the archive data (only 7% of all manually identified motivations or rationalizations),[21] and these motivations effectively disappeared from internal discussions after internment began. Meanwhile, arguments that internment was an appropriate response to a formidable security threat predominated throughout the entire analyzed timeframe. A frequency-focused analysis of human- or computer-coded concepts could obfuscate the influential role that infrequently articulated concepts (here, political motivations) played in dictating policy outcomes and could therefore lead scholars to over-emphasize the importance of more commonly articulated (but arguably less influential) motivations. Thus, the frequency with which a concept is articulated can introduce an obstacle to researcher knowledge. Before making conclusions about the factors that drive government decisions, researchers should augment frequency-based assessments with qualitative evaluations of available evidence (Fairfield and Charman 2019; Tanweer et al. 2021).

Officials Adopt Misrepresented Rationales Over Time

Finally, over time, policymakers could come to adopt the motivations that were initially misrepresented, over-emphasized or invented for public consumption, thus further obscuring a researcher's ability to uncover initial policy motivations.[22] Even though internment lacked initial support from Britain's military officials and had little immediate observable effect on violence,[23] British officials began to reiterate, advance, build policy around, and even adopt the idea that internment was a necessary national security policy.[24] In fact, despite evidence that violence and public support for the Irish Republican Army (whom internment sought to temper) both increased after internment began, HMG ministers asserted that, "[i]n principle internment has been a success. It has done what it was intended to do; and… the chance of gradually bringing terrorism to an end are now reasonably good" (PREM15/482:6207).[25]

20 This commentary, published three months after internment began, was penned by Claire Palley, Professor of Public Law at Queen's University of Belfast (PREM15/484:6311).

21 See the Online Appendix (A3, available at https://www.qmmrpublication.com/issues).

22 Actors benefit from harboring inaccurate beliefs about their own motivations (Little 2019). This behavior could be evidence of self-perception theory (Bem 1972) and motivated bias (Kunda 1990). We thank John Patty for providing this point.

23 The weeks after internment's introduction saw increases in: violence (e.g., explosions and soldier deaths, PREM15/482:6220; PREM15/479:5836), coordinated IRA tactics (PREM15/482:6220), and public support for the IRA: "the strength of support for and tolerance of IRA activities in Northern Ireland. . . has grown enormously, stimulated first by internment" (PREM15/483:6024; see also: PREM15/483:6067).

24 For example, by February 1972 (six months after internment began), HMG defense officials highlighted "important security reasons" for not ending internment (PREM15/1002:6778; see also PREM15/1002:6772).

25 For example, two months after internment began, HMG's Belfast representative said that "intelligence and military reasons argue in the direction of no limit" to the number of people targeted for internment (PREM15/482:6195).

These security-based motivations also shaped subsequent policies dictating when internees were released, when internment would end,[26] and how Britain justified maintaining internment —in light of the emergency security situation (PREM15/1689:159, PREM15/1013:8052)[27]— after adopting direct rule over Northern Ireland in March 1972. Indeed, internment would have been legally dubious under British common law and the European Convention on Human Rights (raising "issues as to the rights and liberty of the subject," PREM15/1002:6812) if internment could not be considered a direct response to an immediate security-related emergency (PREM15/1012:7924). In short, state actors may come to genuinely believe—and act according to—the motivations that their colleagues initially misrepresented for public consumption.[28]

The obstacles we highlighted here emerge when government officials seek to misrepresent the true motivations behind policy outcomes. Government actors are likely to misrepresent motivations when their real objectives appear to violate the norms that their governments purport to uphold and/or when they risk receiving public critique and angering their voters if their true motivations were revealed. We expect such misrepresentation to be particularly evident among liberal constitutional democracies when policymakers in those democracies adopt policies that violate their state's domestic and international commitments to civil liberties.

Conclusion

The government archives we analyzed here demonstrate disagreement among similarly situated actors, specious presentations of motives, and political strategies morphing into prominent narratives of necessity—narratives that influenced policy outcomes and policymaker attitudes (or non-decisions; Bachrach and Baratz 1963). All the while, decisions were shaped by a specific political context; individual actors' strategic and normative commitments; interpersonal relations; and an internal decision-making process replete with discord, deceit, concern, indifference, shifting attitudes, and competing normative priorities. Our observations likely generalize to other (but certainly not all) cases in which government officials consider controversial policies, particularly among officials operating within liberal constitutional democracies. As such, our observations build upon, augment, and potentially even correct conclusions based on generalized political science theories and computational approaches to modeling or speculating about government decision-making.[29]

This article demonstrates the value of augmenting prominent political science theories and computational approaches with comprehensive attention to political context and qualitative review of data generated during real decision-making processes. Doing so can help validate, extend, and challenge researchers' assumptions and conclusions. Without doing so, researchers risk drawing under-specified or misguided conclusions about the true motives and factors that shape policy outcomes. We therefore join other archive-based researchers in encouraging scholars to augment generalizability and computational quantification with context-specific examinations of all available data in order to develop holistic understandings of the real processes by which government officials make controversial decisions.

---

26 Memos by Heath and defense officials: "All internees are. . . either members of the IRA or otherwise involved in terrorism. . . [and therefore likely to] return to their previous activities" (PREM15/1003:6958) as "ring-leaders," "bomb-throwers," and "street shooters" (PREM15/1002:6778-79; also: PREM15/1003:6968, 6813, 6834, 6893; CJ4/458:1814-15; PREM15/1013:8245).

27 "Detention of Terrorists" (Northern Ireland) Order (Nov 1972) (https://discovery.nationalarchives.gov.uk/details/r/C11110954); "Northern Ireland (Emergency Provisions) Act 1973" (https://www.legislation.gov.uk/ukpga/1973/53).

28 We thank Jon Mercer for providing this point. Another likely example: U.S. commitments to providing financial assistance or promoting democracy in strategically important countries —policies motivated by fiscal or strategic interests— developed into long-term policy objectives and institutions (e.g., the Truman Doctrine, the U.S. Agency for International Development, or the U.S. military presence in Afghanistan).

29 This research did not assess whether the decision-making processes analyzed here lead to sub-optimal, normatively problematic or biased outcomes (Khalil 2009; McDermott 2004; Tversky and Kahneman 1992).

Competing Interests The authors declare none.

Funding and Acknowledgements This work was funded by the National Science Foundation Grant #1823547, "Civil Rights Violations and the Democratic Rule of Law." We thank John Patty, Jon Mercer, Jessica Sun, Jack Paine, Dan Reiter, John Wilkerson, Rachel Cichowski, the Oppression/ Resistance Lab at Emory University, and our anonymous reviewers for the thoughtful feedback on this article. We thank Michael McCann and Noah Smith for their dedicated mentorship across the course of this project. Any mistakes are our own.

References

Acree, Brice, Justin Gross, Noah Smith, Yanchuan Sim, and Amber Boydstun. 2020. "Etch A-Sketching: Evaluating the Post-primary Rhetorical Moderation Hypothesis." *American Politics Research* 48, no. 1 (January): 99–131.

Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg. 2017. "Accountability and Information in Elections." *American Economic Journal: Microeconomics* 9, no. 2 (May): 95–138.

Bachrach, Peter, and Morton Baratz. 1963. "Decisions and Nondecisions: An Analytical Framework." *American Political Science Review* 57, no. 3 (September): 632–42.

Bazelon, Emily, and Michael Wines. 2021. "How the Census Bureau Stood Up to Donald Trump's Meddling." *New York Times*, August 12, 2021. https://www.nytimes.com/2021/08/12/sunday-review/census-redistricting-trump immigrants.html.

Bem, Daryl. 1972. "Self-Perception Theory." *Advances in Experimental Social Psychology* 6: 1–62.

Blaydes, Lisa. 2018. *State of Repression: Iraq Under Saddam Hussein*. Princeton: Princeton University Press.

Cook, Karen Schweers, and Margaret Levi. 2008. *The Limits of Rationality*. Chicago: University of Chicago Press.

Davenport, Christian. 2007. "State Repression and Political Order." *Annual Review of Political Science* 10:1–23.

Dellis, Arnaud. 2007. "Blame-Game Politics in a Coalition Government." *Journal of Public Economics* 91, no. 1-2 (February): 77–96.

Dreier, Sarah, and Emily Gade. 2024. "Analyzing Archived Government Correspondence Files (1969-73): Internment in Northern Ireland" *Harvard Dataverse* https://dataverse.harvard.edu/dataverse/dreier-gade-internment-ni.

Druckman, James, and Rose McDermott. 2008. "Emotion and the Framing of Risky Choice." *Political Behavior* 30 (3): 297–321.

Duelfer, Charles, and Stephen Benedict Dyson. 2011. "Chronic Misperception and International Conflict: The Us-Iraq Experience." *International Security* 36, no. 1 (Summer): 73–100.

Fairfield, Tasha, and Andrew Charman. 2019. "A Dialogue with the Data: The Bayesian Foundations of Iterative Research in Qualitative Social Science." *Perspectives on Politics* 17, no. 1 (March): 154–67.

Faulkner, Brian. 1978. *Memoirs of a Statesman*. London: Weidenfeld and Nicolson.

Gade, Emily, Sarah Dreier, John Wilkerson, and Anne Washington. 2021. "Archived Attributes: An Internet-Text Approach to Measuring Legislator Attitudes and Behavior." *British Journal of Political Science* 51, no. 4 (October): 1734–41.

Gailmard, Sean, and John Patty. 2019. "Preventing Prevention." *American Journal of Political Science* 63, no. 2 (April): 342–52.

Goldstein, Judith, and Robert Keohane. 2019. "Ideas and Foreign Policy: An Analytical Framework." In *Ideas and Foreign Policy*, edited by Judith Goldstein and Robert Keohane, 3-30. Ithaca: Cornell University Press.

Grimmer, Justin, and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21, no. 3 (Summer): 267–97.

Groseclose, Tim, and Nolan McCarty. 2001. "The Politics of Blame: Bargaining Before an Audience." *American Journal of Political Science* 45, no. 1 (January): 100–19.

Hermann, Margaret, and Binnur Ozkececi-Taner. 2011. "The Experiment and Foreign Policy Decision Making." In *Cambridge Handbook of Experimental Political Science*, edited by James Druckman, Donald Greene, James Kuklinski, and Arthur Lupia, 430-42. Cambridge: Cambridge University Press.

Huang, Leslie, Patrick Perry, and Arthur Spirling. 2020. "A General Model of Author "Style" with Application to the UK House of Commons, 1935–2018." *Political Analysis* 28, no. 3 (July): 412-34.

Jervis, Robert. 2017. *Perception and Misperception in International Politics*. Princeton: Princeton University Press.

Karni, Edi. 2013. *Decision Making Under Uncertainty*. Cambridge: Harvard University Press.

Khalil, Elias. 2009. *The New Behavioral Economics*. Cheltenham: Edward Elgar Publishing.

Khong, Yuen Foong. 1992. *Analogies at War*. Princeton: Princeton University Press.

Kim, Diana S. 2021. *Empires of Vice: The Rise of Opium Prohibition across Southeast Asia*. Princeton: Princeton University Press.

Kim, Diana S. 2022. "Taming Abundance: Doing Digital Archival Research (As Political Scientists)." *PS: Political Science & Politics* 55, no. 3 (July): 530–8.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108, no. 3 (November): 480-98.

Lawrence, Adria. 2013. *Imperial Rule and the Politics of Nationalism: Anti-colonial Protest in the French Empire*. Cambridge: Cambridge University Press.

Little, Andrew T. 2019. "The Distortion of Related Beliefs." *American Journal of Political Science* 63, no. 3 (July): 675–89.

Lorentzen, Peter, M. Taylor Fravel, and Jack Paine. 2017. "Qualitative Investigation of Theoretical Models: The Value of Process Tracing." *Journal of Theoretical Politics* 29, no. 3 (July): 467– 91.

McCleery, Martin. 2015. *Operation Demetrius and Its Aftermath: A New History of the Use of Internment Without Trial in Northern Ireland, 1971-75*. Manchester: Manchester University Press.

McCleery, Martin J. 2012. "Debunking the Myths of Operation Demetrius: The Introduction of Internment in Northern Ireland in 1971." *Irish Political Studies* 27 (3): 411–30.

McDermott, Rose. 2004. "Prospect Theory in Political Science: Gains and Losses from the First Decade." *Political Psychology* 25, no. 2 (April): 289–312.

Mercer, Jonathan. 2013. "Emotion and Strategy in the Korean War." *International Organization* 67, no. 2 (Spring): 221–52.

Mercer, Jonathan. 2014. "Feeling like a State: Social Emotion and Identity." *International Theory* 6, no. 3 (September): 515-35.

Patty, John. 2009. "The Politics of Biased Information." *The Journal of Politics* 71, no. 2 (April): 385–97.

Patty, John, and Elizabeth Maggie Penn. 2014a. "Sequential Decision Making and Information Aggregation in Small Networks." *Political Science Research and Methods* 2, no. 2 (October): 243–71.

Patty, John, and Elizabeth Maggie Penn. 2014b. *Social Choice and Legitimacy: The Possibilities of Impossibility*. Cambridge: Cambridge University Press.

Pearlman, Wendy. 2013. "Emotions and the Microfoundations of the Arab Uprisings." *Perspectives on Politics* 11, no. 2 (June): 387–409.

Penn, Elizabeth Maggie, John Patty, and Sean Gailmard. 2011. "Manipulation and Single Peakedness: A General Result." *American Journal of Political Science* 55, no. 2 (April): 436–49.

Putnam, Robert. 1988. "Diplomacy and Domestic Politics: The Logic of Two-Level Games." *International Organization* 42, no. 3 (Summer): 427–60.

Redlawsk, David, and Richard Lau. 2013. "Behavioral Decision-Making." In *The Oxford Handbook of Political Psychology*, 2nd ed., edited by Leonie Huddy, David O. Sears, and Jack S. Levy, 130-64. Oxford: Oxford University Press.

Saunders, Elizabeth N. 2017. *Leaders at War: How Presidents Shape Military Interventions*. Ithaca: Cornell University Press.

Schelling, Thomas. 1980. *The Strategy of Conflict: With a New Preface by The Author*. Cambridge: Harvard University Press.

Sheffer, Lior, Peter John Loewen, Stuart Soroka, Stefaan Walgrave, and Tamir Sheafer. 2018. "Nonrepresentative Representatives: An Experimental Study of the Decision Making of Elected Politicians." *American Political Science Review* 112, no. 2 (May): 302–21.

Spirling, Arthur. 2016. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *The Journal of Politics* 78, no. 1 (January): 120–36.

Stevens, Daniel. 2019. "Satisficing in Political Decision Making." *Oxford Research Encyclopedia: Politics*.

Tanweer, Anissa, Emily Gade, PM Krafft, and Sarah Dreier. 2021. "Why the Data Revolution Needs Qualitative Thinking." *Harvard Data Science Review* 3.3.

The National Archives (TNA). 1969-73. Records from the Prime Minister's Office (PREM), the Ministry of Defense (DEFE), and the Northern Ireland office (CJ). Record opening date: January 2002. Photographed in September 2015. https://www.nationalarchives.gov.uk/.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5(4): 297–323.

Vis, Barbara, and Dieuwertje Kuijpers. 2018. "Prospect Theory and Foreign Policy Decision Making: Underexposed Issues, Advancements, and Ways Forward." *Contemporary Security Policy* 39 (4): 575–89.

Waltz, Kenneth. 2010. *Theory of International Politics*. Long Grove: Waveland Press.

Weingast, Barry, Jean-Laurent Rosenthal, Margaret Levi, Avner Greif, and Robert Bates. 2020. *Analytic Narratives*. Princeton: Princeton University Press.

Weld, Kirsten. 2014. *Paper Cadavers: The Archives of Dictatorship in Guatemala*. Durham: Duke University Press.

Wilson, Rick. 2011. "The Contribution of Behavioral Economics to Political Science." *Annual Review of Political Science* 14 (June): 201–23.

QMMR Qualitative & Multi-Method Research

Notes from the Field

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062870

Reflections from Field Surveys and Ethnography in Northern Rural India[1]

Ajita Srivastava
Bremen International Graduate School of Social Sciences (BIGSSS)

Introduction

"Think like your participant," advised a professor when I was preparing surveys for fieldwork in the rural and urban parts of Northern India. At that time, I did not fully understand the meaning of that advice because discerning how your participant thinks is something many social scientists, especially psychologists, take for granted.[2] After all, being able to understand the thoughts and feelings of your participant with considerable accuracy is the key "occupational skill" required of a social scientist. Moreover, I am Indian. I speak the local language, understand the customs, and internalize the norms. What could possibly go wrong? I wish I had known better…

Original Plan

The goal of my doctoral research was to compare the decision-making strategies of people in rural and metropolitan areas in India. I speculated that people in urban areas would reveal maximizing decision tendencies (i.e., continuously striving for better options), whereas people from rural areas would be more likely to show satisficing (i.e., choosing to "make do" or choosing satisfactory options). To explore this empirically, I initially planned to conduct surveys with rural farmers and metropolitan corporate employees in Northern India. Given that developing countries (i.e., the Global South) have larger agricultural populations (Salim 2015) and the corporate sector is more recent but fast-growing (Edwards 1997), it seemed plausible to anticipate rural–urban differences in cultural, market, and economic values. I expected these recent but far-reaching differences between India's rural-agricultural and urban-corporate regions to have solidified into the above-mentioned differences in decision making. To test this idea, I planned to collect only quantitative data, with the

1 This project was funded by EU-Cofund program Horizon 2020, Marie Skłodowska Curie Actions Program project number 713639 through the Bremen International Graduate School of Social Sciences (BIGSSS), Germany. The present research would not have been possible without Shariq Khan, Vinay Krishna, Satyendra Kumar, Ayushi Goyal, Shalinee Tripathi, Swapnil Gupta, Pooja Rani Pandey, Abhay Singh, and Vaibhav Dwivedi, who assisted me in the field studies. I am also grateful to Dr. Dora Simunovic, Dr. Regina Arandt and Prof. Ulrich Kühnen for their insightful comments on this essay.

2 I received training in psychology from an interdisciplinary social sciences institution along with sociologists and political scientists.

help of my research assistants (RAs), from 893 farmers and corporate employees. The 454 rural farmers we interviewed serve as the focus of the present reflection.[3]

I first compared rural–urban decision-making preferences by collecting data through quantitative survey measures. Using survey scales translated into the native language of the region (Hindi) seemed to be a simple and feasible way of testing my hypotheses. I consulted friends, colleagues, and acquaintances from rural regions in India about the questionnaires and pre-tested my instruments with Indian students to get an overall assessment of the "cultural fit" of the measures in the Indian context. I also ran pilot tests with my RAs and asked them to answer the survey questions as if they were participants, since most had relatives and families in nearby rural regions that they visited occasionally. After considering all suggestions and modifying and compiling the questionnaires, I thought I was ready for fieldwork in a rural part of a Hindi-speaking region.

What Standardized Methods Texts and Training Cannot Anticipate

Little did I know that my preparations were insufficient. This became evident in two ways that I will discuss here. My RAs and I collected data in the villages of Bhadohi, a rural district in Uttar Pradesh, which is a state in Northern India. The district has a large number of medium to small-scale farmers. As we reached one of the villages in the afternoon, I saw people who were either running small shops, coming back from the fields, or sitting in front of their houses basking in the winter sun, chatting with their friends or relatives. I heard them talking in a Hindi dialect I could not fully understand. With that, some of my assumptions were already being challenged.

Still hopeful and optimistic, we conducted a few pretests with the local population to explore how the participants understood the survey measures and reacted to them. We approached an old man sitting in the sun on a cot in front of his house. We formally introduced ourselves as university students and asked if he could possibly spare us a little time. He seemed welcoming and agreed to do a survey. I opened the copy of the survey questionnaire to start asking questions —then it all struck me! I saw him and the questionnaire, and I thought, "This was a very bad idea!" I had a printout with scale items in formal Hindi in front of me, with questions like: *My personal identity, independent of others, is very important to me*, with 7 points from disagree to agree. I was sitting in front of a man in his seventies, a rural farmer wearing a kurta and dhoti, looking at me eagerly and waiting for me to say something meaningful. Did he ever think about his identity at all or in "personal" terms beyond being a husband, father, grandfather or any other relational roles? Had he ever rated himself on a numbered "disagree to agree" scale? Even more fundamental doubts came to mind, such as, "Does he speak formal Hindi at all?" In the space of one breath, I realized that all the reading and training on data collection I had completed throughout my education had not prepared me for this moment! University student participants and RAs hardly give you an insight into the lives of people outside of universities, even if they belong to the same culture or similar community. They are especially different from semi-literate populations far removed from urban contexts. Now what do I do? I am supposed to demonstrate to the RAs how this works, but I am stumped! The carefully planned method did not seem to work, but it was impossible to change the design of the whole study then and there.

With some hesitation, I read the first question aloud and then asked him how much he agreed with it on a scale from 1 to 7. He seemed confused for quite some time and then said, "Yes, of course I do." I asked again how much on a scale of 1 to 7 did he agree. Again, he seemed confused and said he did not understand. During an awkward pause, I kept thinking about how to reformulate the question and had almost decided to give up, when one RA said, "Let me try something!" He

3 Data from 377 farmers were used in the final quantitative analysis. The reflections I share here come from all the participants, including those who were not included in the quantitative analyses.

started talking to the man in an informal way (referring to him as *Chacha* or father's brother) and casually talked about the topics the questions addressed, occasionally asking, "What do you think?" or "Did this ever happen?" I recognized that I would need to *talk* to participants beyond what was written in the questionnaire to determine whether the phenomenon I intended to measure existed in their reality. Merely handing out the questionnaire or reading aloud from my printouts would not suffice. The questions were not alien to the participants, and certainly this older man was more than capable of understanding their meaning, but it would require some probing and explanation. Standardized survey procedures are quite formal in the sense of "sticking to the script" to assure uniformity in the administration of the questionnaire and reduce any bias on the part of the interviewer. Yet a lack of informal interaction beyond the questionnaire can prevent the interviewer from ascertaining whether the construct is mutually understood in the same way.

The second issue I observed was that the participants were rarely ever alone. They were often surrounded by family members, neighbors, or curious passersby. Some of these bystanders would interfere with the interview —either directly by engaging in the conversation or indirectly by just being there to monitor the situation. Again, from the perspective of standardized methodology, there should be no bystanders or audience in an interview situation whatsoever. At the same time, it was important for me to accept that one cannot avoid bystanders or guardians in a field study in a rural context. Moreover, privacy has a different meaning in traditional, collectivistic contexts. Asking for a private space in which to conduct a survey or interview can arouse suspicion, as it might suggest that one wants to discuss something questionable or inappropriate. This can make participants uncomfortable and cause them to distrust the interviewer. Previous researchers in the Indian rural context have tried to keep bystanders engaged in other activities at a distance to avoid interfering with interviews (Sinha 1983). However, this strategy can easily backfire since the participant can either get distracted by feeling left out of something interesting or feel anxious in the presence of a complete stranger.

In summary, I found myself confronted with numerous problems. First, my participants were unfamiliar with the measures and the standardized survey language. Second, I could not conduct the study in the formal, sterile style of a one-to-one survey as suggested by the methods texts and the training I received. The solution to these issues became clear as I went on with the study: I needed to complement my survey study with a conversational, narrative approach that, at the same time, did not create "noise" in the data.

Exploring Solutions

I soon discovered that asking how much the participants agreed or disagreed with some statement on a scale was highly unnatural. For remote populations such as the older man's village in rural Northern India, eliciting genuine responses required asking questions in a natural, conversational way. To adapt the survey to this more conversational and understandable approach, we broke the survey questions down into two steps. First, we would read out the item and ascertain whether the participant understood it. If not, we explained the question further. Following this, we simplified the response options on the Likert scale. We divided the broad response option for each item into two parts: We first asked participants to indicate whether they agreed with the statement and then asked to what extent they agreed or disagreed. Through this technique, we were able to ascertain how well the statements made sense to the participants. This also relieved them of having to immediately "force fit" their opinion on a continuum of numbers and, instead, enabled a transition from their natural familiar way of conveying opinions to a classical survey response format.

I decided to combine ethnographic research with the quantitative survey research. Ethnographic research aims to understand participants in their own environment (i.e., their field), *ethno* refering to "the common-sense knowledge of one's environment" (Garfinkel 1974, 16). In ethnographic research, one observes the social

environment of the participants and relies on observation and interaction to make sense of the participants' interpretations of their social world. In my study, I included ethnographic observations as informal and unstructured conversations about the questions during and after the survey. The participants would try to associate, and pinpoint certain beliefs or experiences related to the survey statements, which I will elaborate on in the later sections of this reflection. This helped me link their reasons and understanding to the numerical responses and at the same time assess if they understood the questions. A "yes/no" response was frequently followed by "because what happens is…," leading to an incident or example from their experience and surroundings or by the participant expressing a belief. The participants' confidence and emphasis could be reflected in extreme responses, and their doubt and hesitation were also reflected in the in-between options (e.g., somewhat agree, a little, etc.). This enabled me to confirm participants' degree of agreement or disagreement, especially among participants who were semi-literate, for whom the genuineness of numeric responses was hard to ascertain. Thus, this method allowed me to ensure that the interviewer and the participant understood the question and response format in the same manner.

Another way in which an ethnographic approach helped me was by enabling me to maintain the accuracy of the individual data even after the bystanders' interference. As already mentioned, participants were often surrounded by bystanders who could not be removed from the situation. Bystanders would either leave the interview shortly after knowing what was happening or stay and try to give their input in two ways: agreeing with the participant or refuting the participant's opinions.

We usually carried on with the survey if the bystander(s) only silently observed. If a bystander just agreed and nodded at the participant's response, we recorded the response as it was. If they provided additional explanations or comments, we noted them as well, but separately. We had to tread carefully if the bystanders disagreed and tried to exert their opinions on the participants to avoid any kind of discord, as it would also affect the participant. Disagreements like these could make the participant change their responses to appease the bystander, or if they did not change their responses for that question, the disagreement could emotionally charge them and affect the subsequent responses. We would tell the bystanders that we could not record responses for more than one person on one form and asked them to wait so that we could either arrange for another appointment or do a separate survey right after the one we were conducting. Most of the time they would leave before the survey with the actual participant was over. If there was too much interference on a response, we marked it on the form and removed it from the final analysis.

By the end of the study, I had quantitative data as well as many survey-related comments, explanations, and conversation snippets gathered from participants and bystanders, which served as qualitative data that later helped me to explain the findings of my study. I discuss these methodological contributions below.

Rapport Establishment and Depth of Responses

Rapport constitutes the interactive practices that support the information exchange process between the interviewer and interviewee and foster the views of the speaker (Prior 2018). Establishing rapport is essential to obtaining authentic responses from the participants in a qualitative or mixed method study, which is somewhat challenging with a special population like the one I was working with. An issue that emerges concerning rapport in such field studies is the cultural effect of being observed by the interviewer. The interviewers themselves can also affect the responses of the participants. Participants tend to respond according to what they feel the interviewer might like or approve of. This phenomenon, also known as a "courtesy effect," is quite prevalent in the Indian rural context (Sinha 1983). Traditional collectivistic cultures practice low self-expression to maintain harmony and avoid conflict (Lalwani, Shavitt, and Johnson

2006).[4] Thus, social desirability,[5] which is giving responses acceptable to social standards, is likely to play a major role during surveys making it difficult to get genuine answers from participants. Here again, talking more in-depth and in a conversational manner to the participants proved helpful in getting around this potential issue, improving the accuracy of responses. It also allowed participants to follow up their responses with further reasoning, which helped the interviewer capture their opinions about the statement in question rather than only recording the degree of agreement or disagreement.

Construct Validity

The more conversational approach also revealed the difference between how certain constructs are measured and how they are reflected in local communities (i.e., the construct validity of the measure). A scale has good construct validity if it measures the theoretical construct it is supposed to measure (Anastasi 1976, 151). I quickly recognized that the construct entailed far more that could potentially be measured than was allowed by the scale. For example, one of the constructs measured was individualism–collectivism. Individualism is the extent to which people feel independent as opposed to being interdependent as members of larger groups or communities in collectivistic contexts (Hofstede, Hofstede, and Minkov 2010). Previous studies show that urbanization and economic commercialization lead to the adoption of individualistic practices, whereas traditional rural people are more collectivistic and community-oriented (Greenfield 2009).

Contrary to expectations, a couple of rural participants scored high on individualism as measured by the standard scale. However, the ethnographic observations indicated clear collectivistic tendencies. For example, an older farmer who scored high on individualism lamented that his sons were not taking care of him, which is a typical collectivistic expectation. Similarly, a newly married young woman who had just finished school also scored high on individualism and emphasized how self-reliant she was (e.g., "I do all my chores on my own"). However, more in-depth conversations revealed that decisions about her education or marriage weren't her own. Therefore, my initial assumptions regarding individualism and collectivism were not supported by the data. However, I did find an interesting relationship between both constructs: the preliminary analysis showed that both dimensions were positively correlated with each other, instead of being orthogonal dimensions as given in the conventional literature (Markus and Kitayama 1991). Thanks to the additional room for a more ethnographic approach, I was able to gather an informed idea of why this was the case. Individualism as a construct did not exist in the rural context and hence was not understood by the participants. This was also confirmed by the low consistency of individualism, and it being a relatively small factor compared to collectivism in the exploratory factor analysis. Since individualism was not a part of the social reality of rural participants, it could neither be measured nor compared with the urban sample. As a result, the construct had to be removed from the final analysis of the overall quantitative data. This made me realize that although my prediction was plausible, the low construct validity of the measure led to my null results. The qualitative data provided insights with which to corroborate the null findings with the existing literature on individualism–collectivism.

Ecological Validity

The ethnographic approach also helped me understand the participants' context and enhance the measures' ecological validity. Ecological validity refers to the implications of a study for real-world settings (Andrade 2018). Ethnography helped me determine whether the constructs and measures had real, tangible consequences for the

4 Self-expression is the act of conveying one's own thoughts and feelings.

5 Social desirability denotes the tendency of giving a socially acceptable response, even if it is not in alignment with one's personal opinion. While the courtesy effect is the desirability of response directed only towards the interviewer, social desirability is governed by general social and community standards.

participants. For example, one of my concepts was neoliberalism, measured as an attitude towards affirmative action, competitiveness, and personal wherewithal. A related construct was "discrimination." The latter concept had an interesting meaning in this context since the caste boundaries are considerably firm in rural India. Even settlements are divided according to castes or surnames. Rarely do the upper caste and lower caste people share the same well for water. People —especially those from Brahmin communities (the highest caste in Hinduism) —asked for the researchers' full names to determine their caste before serving tea or snacks since they keep separate eating and drinking utensils for the people who are in a lower caste. However, since such hierarchical differences are a norm in the region, most people, regardless of caste, understood them as natural rather than discriminatory. While talking about the diminishing caste differences, some villagers from the Brahmin communities said that they had started to share a well with people of lower castes and, therefore, they were becoming more progressive. Some of them also recounted that since the newest village chief was from a lower caste, they were the ones who faced discrimination. The above understanding of discrimination was different not only from the original measure, but also from that of the urban participants in the study. This new understanding that emerged from ethnographic insights affected the factor structure of neoliberal values, leading to adjustments to accommodate and compare the neoliberal values of rural people with those of people living in the city. This insight also helped add "local meaning" while discussing the findings about neoliberalism and discrimination.

The concerns regarding construct and ecological validity of quantitative measures in non-academic and non-Western populations and pertinent ethnographic solutions are not new. Mixed-method studies in Mexico on violence victimization and collective action have used ethnographic field research to construct surveys specific to the pertinent population (Bell-Martin 2022). More local to my research, ethnographic field observations have been used for sampling and vignette designing for migrant urban workers in informal sector[6] in India (Thachil 2018). In these studies, ethnographic field work was done prior to quantitative studies. This ethnographic research yielded the exact local meanings and regional terms reflecting the theoretical constructs in the participants' social reality. These were used in the quantitative measures of surveys and vignettes. This indeed makes the measures more sound in terms of construct and ecological validity. In my study, I faced the constraint of maintaining a balance between making the measures localized for the rural participants, and at the same time keeping a common ground to obtain comparable data from the urban participants. Hence, I could not construct regionally-embedded measures for the rural participants, and instead had to collect and integrate the qualitative and quantitative data simultaneously.

Conclusions

Combining survey and ethnographic research allowed me to benefit from the best of both the quantitative and the qualitative approaches. Quantitative data helped ensure the generalizability of the findings, while the qualitative data allowed me to tap into the sensitive ecological characteristics of the context. From a methodological point of view, undertaking ethnographic research while collecting survey data allowed me to establish rapport with the unconventional population and assess and improve the construct and ecological validity of my measures.

Conventional survey methods lack the conversational element, which sometimes makes it hard for the researcher to trace the thought processes behind the responses and the nature of data one has obtained. Since "numbers don't know where they come from" (Lord 1956, 751), it is usually left to the researchers' experience and good sense to interpret the participants' understanding of the questions. This can be especially challenging

6 Undocumented workers, mostly daily wage laborers like construction workers, street vendors, and rickshaw pullers.

when transporting constructs and measures from one culture to another. Ethnographic research ensures knowledge of the social reality and lived experiences that quantitative research usually lacks. That said, surveys can also complement ethnographic research by testing and trying to replicate the qualitative findings on larger sample in a more systematic way, which ethnographic research cannot do on its own (Bell-Martin 2022).

Another crucial way in which the qualitative approach helped was by enabling the testing of the *external validity* of the existing constructs. Most of the existing psychological measures have been constructed in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) cultures and tested on university students. The measures are usually adapted to other cultures through basic translation without tapping into the cultural and ecological semantics of the construct. This has implications, especially for Global South samples, which comprise 83% of the world's population (UNCTAD, 2022) and the non-student and non-academic population worldwide. While the issues related to directly transferring psychological methods from Western academic contexts to non-Western non-academic contexts have previously been raised (Sinha 1983), concrete steps towards including the ecological sensibilities of these populations are scarce. Combining qualitative, open-ended research methods, such as ethnography, with quantitative methods provides a shared ground to understand and adapt cultural concepts between the Western world and the Global South, improving operationalization and enhancing the validity of the constructs. In my own field research, complementing the survey with an ethnographic approach facilitated the collection of higher-quality data by allowing me to establish rapport with the participants, convey the true meaning of various concepts that researchers take for granted, and limit the effects of social desirability. It also allowed participants to self-express in a way that enriched the data collected via the survey instrument. More importantly, this mixed-method approach made it possible to conduct interviews, which were essential to gather rich data from the special sample I was working with without losing the survey mode, which was key for the quantitative approach planned while the study was being designed. Recent studies recommend further steps of pre-registering qualitative components of a mixed method design. This is a vital step in terms of formalization, transparency, and replicability of social science research (Peréz Bentacur and Tiscornia 2022). A clear documentation of ethnographic analyses in terms of fieldwork and how did each information contribute to the quantitative measurement development would increase the external validity through replication in new contexts and new populations.

Returning to the advice of "thinking like your participant," stepping into the participants' shoes is indeed necessary to designing sound research in the social sciences. The challenge, however, is finding the correct shoe size, especially if you are designing a study with an unconventional population. The mixed method approach allows the researcher to test existing theories as well as refine them for special populations in an unstructured and uncertain environment. When embarking on research in a new context, an investigator should understand the novel context as a space in which to test and expand theories, invest in pretesting but be aware that the real thing may require some trial and error—and, of course, leave any assumptions at the door!

References

Anastasi, Anne. 1976. "Validity: Basic Concepts." In Psy*chological Testing* 4th ed., 134-61. New York: Macmillan Press.

Andrade, Chittaranjan. 2018. "Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation," *Indian Journal of Psychological Medicine* 40, no. 5 (Sept-Oct): 498–9. https://doi.org/10.4103/IJPSYM.IJPSYM_334_18

Bell-Martin, Rebecca. 2022. "Experimentally Testing Ethnographic Insights: Benefits, Challenges, and Considerations." *Qualitative and Multi-Method Research* 19/20, no. 2/1(Fall-Spring): 2-7. https://doi.org/10.5281/zenodo.6457914.

Edwards, Sebastian. 1997. "Trade Liberalization Reforms and the World Bank." *The American Economic Review* 87, no. 2 (May): 43–8.

Garfinkel, Harold. 1974. "The Origins of the Term '*Ethnomethodology*.'" In Ethnomethodology, edited by Turner Roy, 15-8. Middlesex: Penguin Press.

Greenfield, Patricia M. 2009. "Linking Social Change and Developmental Change: Shifting Pathways of Human Development." *Developmental Psychology* 45 (2): 401–18. https://doi.org/10.1037/a0014726.

Hofstede, Geert, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind*, 3rd ed. London: McGrawHill.

Lalwani, Ashok K., Sharon Shavitt, and Timothy Johnson. 2006. "What Is the Relation Between Cultural Orientation and Socially Desirable Responding?"*Journal of Personality and Social Psychology* 90, no. 1 (January): 165–78. https://doi.org/10.1037/0022-3514.90.1.165.

Lord, Frederic M. 1953. "On the Statistical Treatment of Football Numbers." *American Psychologist* 8 (12): 750–1. https://doi.org/10.1037/h0063675.

Markus, Hazel R., and Shinobu Kitayama. 1991. "Culture and the Self: Implications for Cognition, Emotion, and Motivation." *Psychological Review* 98 (2): 224–53. https://doi.org/10.1037/0033-295X.98.2.224.

Peréz Bentacur, Verónica, and Lucía Tiscornia. 2022. "Transparency and Replicability in Mixed-Methods Designs Using Experiments." *Qualitative and Multi-Method Research* 19/20, no. 2/1(Fall-Spring): 14-20. https://doi.org/10.5281/zenodo.6448122.

Prior, Matthew T. 2018. "Accomplishing "Rapport" in Qualitative Research Interviews: Empathic Moments in Interaction." *Applied Linguistics Review* 9 (4): 487-511. https://doi.org/10.1515/applirev-2017-0029.

Salim, Zamroni. 2015. "*Globalization and Agriculture: Optimizing Trade Policies for Small Farmers." In Globalization and Development Volume One: Leading Issues In Development with Globalization* edited by Shigeru T. Otsubo, 176–99. https://doi.org/10.4324/9781315687100.

Sinha, Durganand. 1983. "Human Assessment in the Indian Context." In *Human Assessment and Cultural Factors*, edited by S.H. Irvine and John W. Berry, 17-34. London: Springer https://doi.org/10.1007/978-1-4899-2151-2_2.

Thachil, Tariq. 2018. "Improving Surveys Through Ethnography: Insights from India's Urban Periphery" *Studies in Comparative International Development* 53, no. 3 (July): 281-99. https://doi.org/10.1007/s12116-018-9272-3

United Nations Trade and Development (UNCTAD). 2022. "Now 8 Billion and Counting: Where the World's Population Has Grown Most and Why That Matters." *UNCTAD*, November 15, 2022. https://unctad.org/data-visualization/now-8-billion-and-counting-where-worlds-population-has-grown-most-and-why.

QMMR Qualitative & Multi-Method Research

Notes from the Classroom

Qualitative and Multi-Method Research Fall 2024, Volume 22.2 https://doi.org/10.5281/zenodo.14062872

Controlled Comparison and Process Tracing: A "Behind the Scenes" Look

Trevor E. Brown

Isabel M. Perera
Cornell University

Introduction

Case-based research remains a staple of political science. Yet when embarking on research projects, many users, including advanced undergraduates, graduate students, and even seasoned researchers, remain confused about its various types. Casual references to "in-depth case studies," "process tracing," and "comparative analysis" can overlook (or misuse) the multiple ways to analyze a case. In both the classroom and in the research review process, we have encountered students and peers who are unclear about these different approaches, as well as the advantages they offer and the challenges they pose.

This confusion in part reflects the lack of many transparent, first-hand accounts of how researchers select their research design and, in turn, their cases. Students, teachers, and researchers alike would benefit from greater explicit reflection on which methods and cases are chosen and why. In this note from the classroom, we therefore offer a "behind the scenes" look at the methodological choices made in our forthcoming article in *World Politics,* "Why States Do or Do Not Privatize: Cross-Class Coalitions in the Public Sector." We lay out the objectives of our study, why we opted for case analysis, and importantly, how we deliberated between two major (if too-often conflated) case-analytic approaches: controlled comparisons and process tracing. We then discuss which design we chose, why we selected the cases we did, and the challenges we faced in doing so. As our reflection suggests, research design and case selection often hinge on a variety of considerations, ranging from factors that are more theoretical to those that are more pragmatic and rooted in the researcher's substantive knowledge. A list of questions for review and discussion is included at the end.

Which Design to Use? Staying on Track when Dealing with Complex Research Material

Our article investigates the following research question: Why do some states privatize public services, while others do not? Over the last five decades, wealthy democracies have increasingly outsourced public services to private actors. Scholars have studied this process in great depth, yet they have tended to focus on so-called "positive cases" —or instances in which countries have successfully privatized a given service. The lack of attention to "negative" cases— or instances where countries do not privatize a particular service —leave us without the necessary variation on the dependent variable to understand which conditions lead to privatization. In other words, without examining instances of non-privatization,

we simply cannot know what causes privatization to begin with (Mahoney and Goertz 2004). This empirical gap served as a good opportunity to test a theory developed by one author in the area of mental health provision (Perera, forthcoming). In brief, the theory posits that where public sector rank-and-file workers form coalitions with managers, governments can maintain or expand public services. When no such coalition emerges, services are vulnerable to cutbacks.

Because we were interested in testing and potentially establishing the relationship between two variables (i.e., worker-management coalitions and privatization), we needed a research design that offered us variation on the dependent variable. Process tracing, which is often used for inductive theory-building or identifying causal mechanisms, would not suffice, at least on its own (Bennett 2009; Collier 2011). This left us with a few design options, such as a controlled qualitative comparison of multiple cases, or some type of regression analysis.

As we reviewed our design options, a few other complications came into play. Because privatization is especially prevalent and studied in wealthy capitalist democracies (whose long histories of independent state formation help to generated large, relatively institutionalized public service infrastructures; but see MacLean 2011), and because both of us focus on such countries in our other research, our intervention would be most fruitful there. We were thus left with only 20 or so potential cases —too few units to generate the power for large-N statistical inference, at least on the subject of national privatization initiatives. Perhaps more importantly, the complex, macro-structural, and cross-temporal nature of several variables that are often associated with privatization required a deep case knowledge that is difficult to capture and analyze in regression techniques. Quantitative evidence can certainly be deployed in qualitative research; indeed, we ultimately used some ourselves. Yet attempts to summarize privatization patterns in a complex industry with only quantitative indicators would be incomplete, to conduct statistical analysis on them would be infeasible, and to draw conclusions from them would be erroneous. As a result, a controlled comparison was necessary, turning us first to Mill's methods.

Mill's Methods and the Controlled Comparison: A Small-N Design to Identify Causal Variables

Although John Stuart Mill may be best known in political science for his contributions to nineteenth-century liberal thought, his contributions to the scientific method are also the logical underpinnings of contemporary controlled comparisons (Przeworski and Teune 1970, 32). In his classic treatise, *A System of Logic* (2012, originally published in 1843), Mill identified five patterns of inductive inference that ground causal empiricism: 1) the method of agreement, 2) the method of difference, 3) the joint method of agreement and difference, 4) the method of residues, and 5) the method of concomitant variation. Importantly, the patterns developed the logic of experimental design —now often celebrated as the gold standard for causal inference— before the advent of randomization in the late 1880s (Copi, Cohen, and Rodych 2019, 525). In this way, Mill's methods offer a unitary logic of causation for both observational and experimental research, rendering their tools "permanently useful" to the natural and social sciences (Copi, Cohen, and Rodych 2019, 525). These five patterns adopt a variable-based approach to inferring causation.

What Mill's methods can do, then, is guide research that aims to identify necessary or sufficient causal *variables* by comparing two or more cases.[1] Perhaps the most prominently used in political science is the "method of difference," also known as the classic "most similar systems design" (Przeworski and Teune 1970). In Mill's own words:

> If an instance in which the phenomenon under investigation occurs and an instance in which it does not occur, have every circumstance in

1 That these methods can do so does not meant that they always do. For guidance on how best to apply these methods, see for example Braumoeller and Goertz (2000); Dul, Vis, and Goertz (2019); Goertz (2006); Goertz and Starr (2002).

> common save one, that one occurring only in the former, the circumstance in which alone the two instances differs, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon (2012, 455).

Put another way, across cases with different outcomes, the presence or absence of a factor in one case but not others can explain the variation under consideration. If said factor exists or does not exist in all cases, it cannot explain the observed difference in outcomes. This method therefore matches cause and effect by "controlling for" shared circumstances, or ruling out potentially confounding factors. That the cause, effect, and shared circumstances are observable is a fundamental presupposition of this approach. Otherwise, the researcher cannot confidently claim to control their presence or absence. Contemporary empiricists typically view these observable circumstances as variables: measurable factors whose value (categorical or continuous) might change across cases.

To be sure, Mill's methods have limitations (see for example, Przeworski and Teune 1970; Seawright 2021; Lieberson 1994). Each demands a high degree of control over case variation. As Lieberson (1994) outlines, the method of difference presupposes a deterministic relationship between cause and effect; assume that just one cause is present; and make interactive relationships difficult to identify. Moreover, as Seawright (2021, 34) argues, unlike those using experimental designs or even large-N statistical analyses, small-N researchers cannot claim that unobserved variation on potentially important covariates "balance out" either within or across units.

Nevertheless, such imperfections have not deterred the ample use of Mill's logic in social science methodology, even in more "relaxed" forms (see Brady and Collier 2010, 337fn8; Przeworski and Teune 1970, Chapt. 2; Slater and Ziblatt 2013). Of particular note is Mill's lasting importance to small-N research, or research that relies on just a few cases. Such studies address phenomena that occur in too few cases to gain sufficient statistical power for quantitative techniques (such as cross-sectional regressions), or at highly-aggregated levels where deep case knowledge is required (such as restructuring patterns in a complex industry, an example we discuss below). When well-selected and carefully designed, case comparisons therefore can yield insights of significant generalizability. As Slater and Ziblatt (2013) argue, particularly when supplemented with some within-case analysis (as we include in our article), controlled comparisons remain "indispensable."

The sheer necessity of small-N research in the social sciences requires that qualitative methodologists follow several guidelines for overcoming the limitations of Mill's methods. Aware that analysts cannot command full control over their cases, for example, methodologists recommend focusing one's efforts on controlling for the major alternative explanations (Slater and Ziblatt 2013). A challenge along the same vein is that of historical regress: At what point in time does the "cause" originate and, by extension, at what point in time should the analyst attempt to "control" the comparison? Scholars can attempt to address this problem by setting a "critical juncture" as the point of departure (Pierson 2004). Crucially, as Slater and Simmons (2010) have argued, establishing such a juncture requires that there are no "critical antecedents" —or pre-existing conditions— up the chain that help explain the outcome.

Both small-N controlled comparisons and much large-N research adopt positivist and deductive approaches to social inquiry, privilege identifying causal variables before mechanisms, and do so by manipulating or otherwise controlling for alternative variables and confounders across comparable units. This shared framework, though, does not render the former lesser or even redundant. On the contrary, and as previously noted, scholars have emphasized the "enduring indispensability" of the controlled comparison (Slater and Ziblatt 2013).

Process Tracing: A Small-N Within-Case Analysis to Identify Causal Mechanisms

Yet there is another prominent method of case research that we also considered: process tracing (Brady and Collier 2010; Bennett 2010). In this approach, the researcher selects one case and tracks how a particular process unfolds

within it to explain a given outcome. In contrast to Mill's method of difference, process tracing can identify, document, and assess the impact of the causal *mechanisms* that connect the variables in a particular sequence.[2] For many, this approach strives to do so by (1) building a strong case for the proposed hypotheses under investigation while (2) eliminating competing explanations with varied pieces of evidence.[3]

One of the core methodological advantages of process tracing is its ability to pin down causal mechanisms, or, in our view (again, definitions vary), how X leads to Y. Mechanisms are nonetheless notoriously difficult to observe; researchers instead rely on the traces they leave behind (Elster 1989, 3–10). Consider policy feedbacks. This mechanism shows how two measurable variables —the structure of a given public policy and the public's support for that policy— can be mutually reinforcing. But how does an investigator measure the feedback itself? The answer, according to process tracing methodologists, is to document the "fingerprints" or "causal process observations" —that is, the empirical observable implications of feedback (Beach and Pedersen 2013; Brady and Collier 2010). In other words, scholars must ask themselves: If policy feedback is present (and operated as hypothesized), what empirical residue would be left over to demonstrate that it occurred? For instance, in-depth interviews with policy beneficiaries, survey data, or analyses of group position statements could generate observable evidence demonstrating that recipients' views are subconsciously conditional on the policy structure. Meanwhile, "reading between the lines" of elites' meeting deliberations or analyzing legislators' vote choices could help an investigator determine the intentions driving policy design (see e.g., Thurston 2018). Varying pieces of evidence would satisfy different evidentiary standards. For example, some methodologists have analogized different types of process tracing "tests" for alternative pieces of evidence, popularizing the use of the following:

- "doubly decisive," necessary and sufficient to confirm a particular explanation;
- "smoking gun," sufficient but not necessary for a given hypothesis;
- "hoop tests," necessary but not sufficient;
- "straw-in-the-wind," suggestive evidence in favor or against a hypothesis, but neither sufficient nor necessary.[4]

Although these are static tests, the methodologists who use them deploy them synthetically over the full length of the causal process.[5] Done in this way, this research strategy can be well-suited to recover causal mechanisms and demonstrate over-time processes.

How We Decided Which Method to Use

Put simply, Mill's structured comparisons and process tracing are different types of qualitative case analysis, each with their own logic and objectives. The choice of which approach to pursue —a controlled comparison of two or more

2 Here we are employing definition of causal mechanism proposed by Falleti and Lynch (2009), though see that article, as well as Hedström and Ylikoski (2010), Beach and Pedersen (2018), and especially Mahoney (2001), for a collection of alternative definitions. For a more detailed discussion of mechanisms, particularly in the context of process tracing, see the "Symposium on Causal Mechanisms and Process Tracing" from the 2016 Spring/Fall issue of *Qualitative and Multi-Method Research*, edited by Alan Jacobs and Tim Büthe.

3 Methodologists writing for this journal and others debate whether and under what conditions process tracing succeeds at these aims, especially the second. See Jacobs and Büthe (2016); Gonzalez-Ocantos and Masullo (2024). Growing attention to Bayesian approaches, for example, has highlighted that scholars can increase their confidence in some hypothesis over others, but never completely eliminate an alternative explanation.

4 See, for example, Van Evera (1997) and Collier (2011, Table 1) for further elaboration on these tests. For alternative perspectives, see Beach and Pederson (2013, Chapt. 8) and Mahoney (2012).

5 Otherwise, scholars may fail to capture the temporal dynamism embedded in the very process they are attempting to observe. As Bateman and Teele (2020, 268–9) point out, a piece of evidence that fails to satisfy a "hoop test" at t^2, for example, might have been pivotal at t^1 in facilitating the conditions that ultimately produce the outcome at t^2.

units or within-case process tracing of a single unit— therefore depends on the researcher's analytic objectives. Our objective was to identify the variables that produced an outcome, so a Mill's-inspired controlled comparison is useful. As such, we relied on the frequently-used "most similar systems" approach. That approach would help us to establish that the relationship between X and Y is consistent across multiple cases.

Process tracing was less appropriate. Our analytic focus was not on mechanisms or theory building. Moreover, we were writing an article, not a book. Space constraints would prevent us from explicating the full path that produced Y, or attempt to recover the mechanisms that lead from X. We did still use some process tracing techniques, as these two methods are not mutually exclusive. Once a relationship between variables is established via Mill's method of difference, researchers often select a single case in which the causal variable is present to trace the mechanism that links those variables. Although we structured our research article as a controlled comparison, where possible we illustrated the "fingerprints" of the mechanisms that linked the variables of interest (Beach and Pedersen 2013).

Case Selection: Which Railway to Ride, and Why?

Once landing on our design, we next needed a procedure for selecting cases. At the outset, we reasoned that we needed at least three. While controlled comparison designs have successfully relied on just two cases, we worried that only examining two could leave us with spurious results. Examining three cases helped us increase the confidence that our findings were reliable. Given the practical space constraints of an article, we felt that three cases offered the right balance between mastery (that is, our ability to collect in-depth, variable-based evidence) and generalizability (bolstering confidence in our findings for other countries), relative to just two cases. While ideally we might include more, three cases also enabled us to stay in bounds of most journals' word limitations, an important consideration when conducting qualitative research.

As discussed above, we were first driven by the need for variation on our dependent variable. This meant we needed an industry that some countries had privatized and others had not. After some digging, passenger rail appeared to fit the bill. Throughout the 1980s and 1990s, a number of countries privatized their passenger rail systems, while others kept it in public ownership (Kopicki and Thompson 1995). Passenger rail was a strong fit for other reasons. Not only had it been privatized in several countries, but the nature of the industry itself also helped us control for a potential confounding factor —patterns of consumer mobilization. While the service is popular in many countries, consumers are rarely a mobilized constituency for passenger rail. As an interest group, they thus likely did not condition the outcome.

Once landing on an industry, we proceeded to select a "negative" case —or a case where privatization was attempted but failed. We were attracted to the United States for a few reasons. First, both of us have strong knowledge of the country. (Indeed, both of us have used Amtrak, America's passenger rail, several times!) Moreover, we felt this case helped deepen our puzzle. Conventional wisdom holds that the United States is home to one of the most extremely "hands off" political economies in the world (Thelen 2014). Of all the countries that have privatized their railways, how had the seemingly anti-statist United States not done so?

Once we had a negative case, we proceeded to find those with "most similar" qualities. Here we needed to be careful to "match" our cases on otherwise potentially influential variables, or variables that could otherwise be present (or not) across several cases and explain our outcome. We started by holding our sample within the cluster of so-called liberal market economies (LMEs), the group of economies famously identified by Hall and Soskice's (2011) "varieties of capitalism" framework. Keeping our case selection within this group allowed us to select cases while confidently ruling out other potentially important factors (major alternative explanations, per Slater and Ziblatt 2013) that might otherwise be related to privatization. These include strong coordinating mechanisms that tend to facilitate labor-

management cooperation throughout the entire economy, not just the public sector.

Timing was also important. As several scholars have documented, the late 1970s through the 1990s mark a time when neoliberalism —or market fundamentalism— was prevalent (Mudge 2008). Consistent with this characterization, the Reagan Administration tried multiple times to privatize Amtrak. Ideally, then, we would find cases that tried to privatize when neoliberalism was prevalent in a given country.

Other political factors mattered. Scholars have demonstrated that whether the political Left or Right is in power can be influential (Obinger, Schmitt, and Zohlnhöfer 2014). In general, Right-leaning parties have tended to favor privatization efforts, while Left-leaning parties have favored national ownership. We thus sought cases where the Right was in power while passenger railway was in jeopardy. Political institutions are also important. Scholarship in the comparative political economy (CPE) tradition suggests that countries with proportional representation systems tend to empower the Left and ultimately result in more progressive redistributive policies, in contrast to majoritarian first-past-the-post systems, where Right-leaning parties tend to enjoy greater power due to their more efficient geographic distribution (Iversen and Soskice 2006).

We ultimately landed on two additional cases: the United Kingdom and New Zealand. Both countries are liberal market economies and, at the time of privatization, featured majoritarian, first-past-the-post electoral rules. Initiated under the Thatcher Government and pursued in earnest under the conservative Major Government, the United Kingdom privatized their passenger railway system in the early 1993. Similarly, the right-leaning New Zealand governments restructured their passenger system several times during the 1980s, culminating in a sale to a private purchaser in 1993. All three of our cases were marked by periods of neoliberal governance —known as "Reaganomics" in the United States, "Thatcherism" in the UK, and "Rogernomics" in New Zealand. Since politicians in several cases attempted but (crucially) did not all succeed at privatization, we were confident that there were no major "critical antecedents" further up the causal chain that condition the outcome, following Slater and Simmons (2010).

Few, if any, observational designs can claim to "control for" or rule out all potential confounding factors simply through our research design. Ours was no different in this regard. Consider that the United States is marked by a uniquely high number of veto points, including an especially strong upper house (i.e., the Senate), relative to most other affluent democracies, including our comparison cases (Stepan and Linz 2011). While traditionally pointed to as a tool used by the wealthy to thwart progressive reforms aimed at arresting inequality (Enns et al. 2014), such veto points can also be used to gum up or stop conservative reforms.

As such, we proceeded by collecting evidence on such potential confounders as well as our key variables of interest. For example, we showed that US veto points played less of a role in in passenger rail politics in this era than scholars might expect. A review of the U.S. Congressional record and spatial railway patterns offered strong evidence that the Senate's comparatively unique power in policymaking could not fully explain the failure of the Reagan Administration to privatize Amtrak. Other potential confounders included the financial standing of the railway sectors; in each case, all were in financial disarray at the time of privatization attempt. We also examined the level of institutionalization of each railway. Theories rooted in policy feedback and path dependency might predict that more institutionalized railways are more likely to survive attacks (Pierson 1993; 2000). Yet as it turned out, the American Amtrak system was the youngest, least institutionalized, and as a result perhaps the most vulnerable to retrenchment. Lastly, we considered the level of trade union organization in all cases. Each railway boasted a high level of union organization —in each case, at least 75 percent of the workforce— suggesting that rank-and-file labor power alone could not account for our outcome.

Throughout the research process, we reviewed all relevant published accounts of rail politics and policy from just before, during, and after the efforts to privatize, including both peer-reviewed academic scholarship and gray literature (e.g., government and think tank reports). We also identified sources

to interview virtually, as the pandemic prevented us from doing so in person. For each case, we aimed to speak to all pertinent industry experts and government officials, as well as the primary representatives of public rail managers and public rail workers, respectively, when privatization efforts were underway. In total, we interviewed 22 individuals. Finally, we supplemented information gathered in steps one and two by reviewing newspaper coverage of rail restructuring during the dates of interest, including from the *Financial Times*, *The Guardian*, *New York Times*, the *New Zealand Herald*, and the *Washington Post*. We ultimately found strong evidence that the presence (and lack thereof) of a coalition between managers and workers shaped whether the railway industry was privatized. Our analysis also required some within-case analysis, not totally unlike process tracing, suggesting these two methods, while distinct, can serve as important compliments to one another (Slater and Ziblatt 2013).

Discussion Questions

1. Have you, or do you plan to, use either of these methods in your on-going research? Why or why not? What makes them appropriate, or not so useful?
2. The authors view their study as largely "variable-based." What does this mean and how does it make it different from within-case process tracing?
3. According to the authors, what factors are similar across each case? What varies? Why is this important?
4. Could you imagine a study of railway privatization set up as a statistical analysis? If yes, how would you code each variable? If no, why?
5. Would a regression-based study be closer to Mill's method of difference or process tracing? What advantages would it offer? What disadvantages would be present?
6. Imagine you wanted to use within-case process tracing to study the privatization of railways. What would be the analytic goal of doing so? Based on the authors' suggestion, which case might you select and for what types of evidence would you search?

Acknowledgments:

The authors are grateful to the participants of the Historical Qualitative Methods Working Group at Cornell University for their valuable feedback on this piece, as well to Jeffrey Checkel and the editors of *QMMR* for enabling its publication in this journal.

References

Bateman, David A., and Dawne L. Teele. 2020. "A Developmental Approach to Historical Causal Inference." *Public Choice* 185, no. 3-4 (December): 253–79.

Beach, Derek, and Rasmus Brun Pedersen. 2013. *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.

Beach, Derek, and Rasmus Brun Pedersen. 2018. "Selecting Appropriate Cases When Tracing Causal Mechanisms." *Sociological Methods & Research* 47, no. 4 (November): 837–71.

Bennett, Andrew. 2009. "Process Tracing: a Bayesian Perspective." In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, 702–21. Oxford: Oxford University Press.

———. 2010. "Process Tracing and Causal Inference." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, *2nd Edition*, edited by Henry E. Brady and David Collier, 207–20. Lanham: Rowman & Littlefield Publishers.

Brady, Henry E., and David Collier, eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. 2nd ed. Lanham: Rowman & Littlefield Publishers.

Braumoeller, Bear F., and Gary Goertz. 2000. "The Methodology of Necessary Conditions." *American Journal of Political Science* 44, no. 4 (October): 844–58.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44, no. 4 (October): 823–30.

Copi, Irving M., Carl Cohen, and Victor Rodych, eds. 2019. *Introduction to Logic*. 15th ed. New York: Routledge.

Dul, Jan, Barbara Vis, and Gary Goertz. 2021. "Necessary Condition Analysis (NCA) Does Exactly What It Should Do When Applied Properly: A Reply to a Comment on NCA*." *Sociological Methods & Research* 50, no. 2 (May): 926-36.

Elster, John. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Enns, Peter K., Nathan J. Kelly, Jana Morgan, Thomas Volscho, and Christopher Witko. 2014. "Conditional Status Quo Bias and Top Income Shares: How US Political Institutions Have Benefited the Rich." *The Journal of Politics* 76, no. 2 (April): 289–303.

Falleti, Tulia G., and Julia F. Lynch. 2009. "Context and Causal Mechanisms in Political Analysis." *Comparative Political Studies* 42, no. 9 (September): 1143–66.

Goertz, Gary, and Harvey Starr. 2002. *Necessary Conditions: Theory, Methodology, and Applications*. Lanham: Rowman & Littlefield.

Goertz, Gary. 2006. "Assessing the Trivialness, Relevance, and Relative Importance of Necessary or Sufficient Conditions in Social Science." *Studies in Comparative International Development* 41, no. 2 (June): 88–109.

Gonzalez-Ocantos, Ezequiel, and Juan Masullo. 2024. "Aligning Interviewing with Process Tracing." *Sociological Methods & Research* 1–49.

Hall, Peter A., and David Soskice. 2001. "An Introduction to Varieties of Capitalism." In *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage*, edited by Peter A. Hall and David Soskice, 1–68. Oxford: Oxford University Press.

Hedström, Peter, and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36 (1): 49–67.

Iversen, Torben, and David Soskice. 2006. "Electoral Institutions and the Politics of Coalitions: Why Some Democracies Redistribute More Than Others." *American Political Science Review* 100, no. 2 (May): 165–81.

Jacobs, Alan M., and Tim Büthe, eds. 2016. "Symposium: Causal Mechanisms and Process Tracing." *Qualitative & Multi-Method Research* 14, no. 1/2 (Spring/Fall): 13-39.

Kopicki, Ron, and Louis S. Thompson. 1995. "Best Methods of Railway Restructuring and Privatization." Discussion Paper No. 111. World Bank, Washington D.C., August 31. http://documents.worldbank.org/curated/en/761811468761126588/Best-methods-of-railway-restructuring-and-privatization.

Lieberson, Stanley. 1994. "More on the Uneasy Case for Using Mill-Type Methods in Small-N Comparative Studies." *Social Forces* 72, no. 4 (June): 1225–37.

MacLean, Lauren M. 2011. "State Retrenchment and the Exercise of Citizenship in Africa." *Comparative Political Studies* 44, no. 9 (September): 1238–66.

Mahoney, James, and Gary Goertz. 2004. "The Possibility Principle: Choosing Negative Cases in Comparative Research." *American Political Science Review* 98, no. 4 (November): 653–69.

Mahoney, James. 2001. "Review Essay: Beyond Correlational Analysis: Recent Innovations in Theory and Method." *Sociological Forum* 16, no. 3 (September): 575–93.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods & Research* 41, no. 4 (November): 570–97.

Mill, John Stuart. 2012 [1843]. *A System of Logic, Ratiocinative and Inductive*. New York: Cambridge University Press.

Mudge, Stephanie Lee. 2008. "What is Neo-liberalism?" *Socio-Economic Review* 6, no. 4 (October): 703–31.

Obinger, Herbert, Carina Schmitt, and Reimut Zohlnhöfer. 2014. "Partisan Politics and Privatization in OECD Countries." *Comparative Political Studies* 47, no. 9 (August): 1294–323.

Perera, Isabel M. Forthcoming. *The Welfare Workforce: Why Mental Health Care Varies Across Affluent Democracies*. New York: Cambridge University Press (Studies in Comparative Politics Series).

Perera, Isabel M., and Trevor E. Brown. Forthcoming. "Why States Do or Do Not Privatize: Cross-Class Coalitions in the Public Sector." *World Politics*.

Pierson, Paul. 1993. "When Effect Becomes Cause: Policy Feedback and Political Change." *World Politics* 45 (4): 595–628.

———. 2000. "Increasing Returns, Path Dependence, and the Study of Politics."*The American Political Science Review* 94 (2): 251–267.

———. 2004. *Politics in Time: History, Institutions, and Social Analysis*. Princeton, N.J.: Princeton University Press.

Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley-Interscience.

Seawright, Jason. 2021. "Beyond Mill: Why Cross-Case Qualitative Causal Inference Is Weak, and Why We Should Still Compare." In *Rethinking Comparison: Innovative Methods for Qualitative Political Inquiry*, edited by Erica S. Simmons and Nicholas Rush Smith, 31–46. Cambridge: Cambridge University Press.

Slater, Dan, and Daniel Ziblatt. 2013. "The Enduring Indispensability of the Controlled Comparison." *Comparative Political Studies* 46, no. 10 (October): 1301–27.

Slater, Dan, and Erica Simmons. 2010. "Informative Regress: Critical Antecedents in Comparative Politics." *Comparative Political Studies* 43, no. 7 (July): 886–917.

Stepan, Alfred, and Juan J. Linz. 2011. "Comparative Perspectives on Inequality and the Quality of Democracy in the United States." *Perspectives on Politics* 9, no. 4 (December): 841–56.

Thelen, Kathleen. 2014. *Varieties of Liberalization and the New Politics of Social Solidarity*. New York: Cambridge University Press.

Thurston, Chloe N. 2018. *At the Boundaries of Homeownership Credit, Discrimination, and the American State*. New York: Cambridge University Press.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca: Cornell University Press.

Longform APSA Awards (2024)

Giovanni Sartori Book Award

Committee: Jana Krause (Chair) (University of Oslo), Keesha Middlemass (Howard University), John Yasuda (Johns Hopkins)

Recipients: Adam Auerbach (American University) and Tariq Thachil (University of Pennsylvania), *Migrants and Machine Politics*, Princeton University Press.

"Migrants and Machine Politics" by Adam Michael Auerbach and Tariq Thachil is an outstanding book that brings together deep ethnographic fieldwork, interviewing, and survey research to analyze political agency among Indian slum dwellers. Party 'machines' are almost always studied from an instrumental perspective of elites. Three years of combined ethnographic fieldwork and hundreds of interviews allow Auerbach and Thachil to challenge the literature on brokerage, urban politics and party machines and identify how political network are constructed at the grassroots level through intense political competition. This methodological perspective restores agency to slum dwellers who do not serve as passive targets of elite machinations but rather build ties to governing authorities through selecting their local community leaders. The book identifies how slum dwellers chose community leaders who best represent their interests and secure development; how community leaders act as brokers and pick projects that advance their political ambitions; and how political parties pick brokers who are competent in delivering the vote. The rich analysis reveals that in India, slum leaders frequently eschew exploiting ethnic divisions within their neighborhoods, instead favoring inclusive strategies for mobilizing broad swathes of support to help launch their own political careers outside slums. Auerbach and Thachil's multimethod work is crucial reading on local-level democracy, party politics, and urban development.

Honorable Mention: Janice K. Gallagher (Rutgers), *Bootstrap Justice,* Oxford University Press.

Bootstrap Justice by Janice K. Gallagher is a deeply researched book that centers the personal experience and political agency of the family members of forcefully disappeared persons in Mexico. Based on ten years of ethnographic immersive fieldwork, Gallagher offers new and enriching conceptual analysis of how trauma reconfigures personal and political agency and enables sustained collective mobilization. The book unpacks the relationship between mobilization and impunity by focusing on the evolving legal consciousness of victims and traces collective claim-making on state institutions to combat impunity. It presents an important analysis of how bottom-up legal activism challenges and partly erodes impunity – while never losing sight of the enormous effort and initiative required of citizens who had to 'pull themselves up by their bootstraps' to pursue justice.

David Collier Mid-Career Achievement Award

Committee: Raúl Madrid (Chair) (UT Austin), Gerardo Munck (University of Southern California), Ben Read (UC, Santa Cruz)

Recipient: Tasha Fairfield (LSE)

The David Collier Mid-Career Achievement Award Selection Committee voted to confer the 2024 prize on Professor Tasha Fairfield of the London School of Economics. In a 2022 Cambridge University Press book and a number of articles and book chapters, Professor Fairfield and her co-author Andrew Charman of the University of California, Berkeley, have developed a formal approach to qualitative inference based on Bayesian theory.

Contributors to *QMMR*'s Fall 2023 symposium on the book, *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*, praised it as a major landmark. It provides a rigorous methodology for weighing evidence from competing theories and offers guidance on a host of related issues, including case selection. It has the potential to have an important impact not just on qualitative methods but on quantitative methods as well. Indeed, in his nomination letter, Professor Andrew Eggers of the University of Chicago writes that he views the approach as "the correct way forward for thinking about how to synthesize evidence in any problem where we want to assess distinct explanations, including when the evidence consists of results from randomized control trials or other quantitative studies." In recognition of the importance of her work, Professor Fairfield has frequently been invited to participate in edited volumes and deliver talks at universities both in Europe and the United States. In addition, her 2017 article in *Political Analysis* won the Sage Best Paper Award from the Qualitative and Multi-Method Section of the American Political Science Association.

Professor Fairfield has also made important institutional contributions to the study of qualitative methods, another criterion for this award. She has served as an Executive Committee Member and as Secretary/Treasurer of the Qualitative and Multi-Method Section of the American Political Science Association and she participated in APSA's Qualitative Transparency Deliberations Working Group on Comparative Methods and Process Tracing. She has also convened a qualitative Bayesian reasoning network, and she has taught various courses and workshops at the Institute for Qualitative and Multi-method Research (IQMR) as well as at the annual meeting of the American Political Science Association.

Alexander L. George Article Award

Committee: Rodrigo Barrenechea (Universidad Católica del Uruguay), Ora Szekely (Clarke University), Stephanie Ternullo (Harvard)

Co-Recipient: Killian Clarke (Georgetown), "Ambivalent allies: How inconsistent foreign support dooms new democracies," *Journal of Peace Research*

"Ambivalent Allies" by Killian Clarke exemplifies the spirit of the Alexander George Award. Clarke tackles an important topic – the success or failure of transitions to democracy – by zeroing in on the role of foreign support in shaping outcomes. Based on a deep analysis of the failed Egyptian transition between 2011 and 2013, he convincingly argues that it is ambivalence and mixed messages on the part of allies abroad that ultimately most endangers a democratic transition. Clarke's article is based on excellent qualitative methods. His analysis of the Egyptian case is based on an impressive number of interviews with Egyptian political figures who can be hard to access, allowing for a finely-grained presentation of the case based on careful process tracing. His controlled comparison with two other cases which vary on the dependent variable further strengthens the argument. Overall, this compelling and impressive use of qualitative methods makes an important contribution to our understanding of both the domestic and international dimensions of the success or failure of democratic transitions.

Co-Recipient: Tahlia Gerzso, "Judicial resistance during electoral disputes: Evidence from Kenya," *Electoral Studies*

"Judicial Resistance During Electoral Disputes" by Tahlia Gerzso is an exemplary piece of case study research, in line with the spirit of the Alexander George Award. Gerzso investigates how the Supreme Court in Kenya could nullify the reelection of the incumbent president, a surprising result that challenges conventional wisdom about the power of the courts in hybrid regimes. As such, this carefully executed case study makes an important contribution to the literature on

resistance to autocracy and autocratization, arguing that judicial reforms that grant judges with more independence was key to this outcome. The article combines quantitative and qualitative analysis of data collected by Gerzso, which are then used to test different observable implications of the causal theory. It uses process tracing to identify theorized mechanisms at work and tests for alternative explanations. Overall, this article's methodological rigor exemplifies what a case study should be. By meticulously tracing the causal pathways and ruling out alternative explanations, Gerzso's work not only deepens our understanding of judicial independence in autocratic contexts but also exemplifies how robust methodological approaches can yield significant theoretical insights.

Kendra Koivu Paper Award

Committee: Susanna Campbell (American University), Sherry Zaks (USC), Janet Lewis (George Washington)

Recipient: Shelley Liu (Duke), "Coercive Legacies of Rebel Governance: Evidence from Zimbabwe"

This paper demonstrates an excellent marshalling of qualitative evidence in service of mechanism development as part of a broader multi-method research design. Specifically, Liu used interviews from archives in Zimbabwe and South Africa that enabled her to trace the continuity of institutions from wartime to post-war governance, giving rise to a holistic understanding of the wartime roots of post-war governance. She also used these same sources to develop and code a new fine-grained measure of local-level rebel control during war, which enabled her to measure degrees of rebel control in rural Zimbabwe. This is an excellent example of an integrated analysis where the quantitative work is contingent on the qualitative data collection and analysis.

2024

QMMR

Qualitative and Multi-Method Research

QMMR Qualitative & Multi-Method Research