Generalization, case studies, and within-case causal inference: Large-N Qualitative Analysis (LNQA)

Version 18, 2020

Gary Goertz

and

Stephan Haggard

Note: The order of the authors is alphabetical.

Prepared for the *The Oxford handbook on the philosophy of political science*.

Thanks to Sharon Crasnow, Harold Kinkaid and Julian Riess for comments on an earlier draft and Terence Teo for help with the data analysis of democratization and inequality. Goertz thanks James Copestake and the University of Bath for providing a wonderful environment for writing this chapter. Thanks the the participants in the *Handbook* workshop at Washington State University as well for valuable feedback.

Abstract

Experiments no less than case studies always raise the question of external validity. This chapter discusses this problem and reviews a developing qualitative research practice that we call "large-N qualitative analysis" (LNQA). The core of the methodology lies in exploring postulated causal mechanisms within individual cases, but for a relatively large number of cases or even the entire population. The approach not only raises a number of interesting methodological issues, but wider epistemological questions about how to generalize and the relationship between type and token causal inference.

> Singular causal claims are primary. This is true in two senses. First, they are a necessary ingredient in the methods we use to establish generic causal claims. Even the methods that test causal laws by looking for regularities will not work unless some singular causal information is filled in first. Second, the regularities themselves play a secondary role in establishing a causal law. They are just evidence – and only one kind of evidence at that – that certain kinds of singular causal facts? have happened.
>
> *Nancy Cartwright*

> The particular and productive character of mechanisms in fact implies that we should think of causation as fundamentally a singular and intrinsic relation between events, rather than as something mediated by laws or universals.
>
> *Stuart Glennan*

Introduction

Philosophers of social science and causation have a long tradition of distinguishing between type versus token causal inference. Types are abstract and general; tokens are concrete particulars. As illustrated in the epigraphs to this chapter, we think that all causal regularities or generalizations ultimately rest on the effects that operate in individual cases. If an experiment shows that there is a significant average treatment effect, that must mean that there are individual cases in which the treatment influenced the outcome for that case. Although cast in probabilistic terms, the average treatment effect is ultimately a kind of summing up of individual level causal influences. If there was no causal effect at the case level, there could be no treatment effect at the population level.

In this paper we pursue the notion that type causation is a generalization of token causal inference. In the social sciences, serious interest in the role of token causal inference received little methodological attention until qualitative and multimethod research took off in political science and sociology over the course of the 1990s. Process tracing and counterfactuals have been focal points in that literature. Process tracing explores the mechanism by which X produced or caused Y in an individual case. Similarly, counterfactual analysis focuses on the determinants of outcomes in individual cases by posing and then confirming or dismissing alternative explanations. This is typically counterfactual dependence *with mechanisms*.[1] We focus on what in the causal mechanism literature is often known at the "trigger" or which in the causal chain metaphor is the initial factor in the causal mechanism.[2] The logic here is of sufficient conditions: the initial

[1] There is not agreement at all when the process tracing literature about counterfactuals in process tracing. Many discuss process tracing without a discussion of counterfactual dependence or against the idea.

[2] We shall not deal with the more complex situation where there may be multiple factors, e.g., interactions, at the beginning of the chain.

factor is sufficient to set the causal mechanism in motion. In this paper we leave the causal mechanism in general as a black box and focus on the generalizability of the mechanism.

A central contention of this paper is that both experiments and case studies face the problem of external validity or, what we prefer, the problem of generalization. How generalizable is the randomized experiment or case study? Experimentalists in political science have started to tackle this problem. The Metakata project has repeated experiments across different countries to see how generalizable findings are (Dunning et al. 2019). We are seeing a similar effort to think about generalization from case studies as well. A recet example is Kaplan's excellent book on civil society organizations in civil war (2017; see the Appendix for a discussion). It starts with what we call a causal mechanism case study. The remainder of the book is preoccupied with how generalizable that mechanism is in Columbia as well as in other settings such as Syria and Afghanistan. Ziblatt's (2017) analysis of the role of conservative parties in European democratization rests on an extensive causal mechanism case study of the UK as well as a comparison with Germany. In his last chapter, however, he provides additional case studies on other European countries and briefer discussions of transitions in Latin America, the Middle East, and Asia (see below for more discussion).

A new research practice has emerged in recent years both among multimethod and qualitative researchers: to multiply the number of qualitative case studies in order to strengthen causal inference that we call Large-N Qualitative Analysis (LNQA). Early examples of the work took a particular form: they sought to challenge prominent statistical or game-theoretical findings by showing that postulated causal relationships did not in fact hold when subjected to closer scrutiny via within-case causal inference; among the targets of this work were prominent accounts on inequality and democratization (Haggard and Kaufman 2016), democratization and war (Narang and Nelson 2009), the effect of audience costs on conflict (Trachtenberg 2012) and the role that rebel victory plays in civil war settlements (Wallensteen 2015; see Goertz 2017, chapter 7 for a discussion). However, the approach has subsequently expanded, as we will show, to define a wider research methodology aimed not only at disconfirming existing analysis but supporting multi-method and in-depth case study work as well.

LNQA is clearly most conducive to the analysis of rare events, or those in which the N is small, such as famines, wars and civil wars, regime changes or the acquisition of nuclear weapons. The approach sometimes starts with statistical analysis, and thus takes a multi-method approach. Other examples, such as the work by Kaplan just cited, start with a single in-depth case study and then augments it with others. But the core of the approach is the use of a (relatively) large number of individual case studies, and even a whole population, in order to strengthen causal inference and generalizability.

To date, these practices have not been justified by reference to methodological works or even by methodological discussions (see however Haggard and Kaufman 2016; Goertz 2017). In the spirit of what Goertz calls "methodological ethnography," this chapter outlines this approach and seeks to ground it theoretically. Based on practice both among experimentalists and those using case studies we argue that the logic of generalization at work is what we will call "absolute generalization" versus the statistical logic of comparison and relative generalization. Causal inference is strengthened via multiple within-case causal inferences rather than comparisons between control and treatment groups or other comparative approaches.

Towards the end of the chapter we explore some concrete examples of this methodology in action. We provide an extended discussion of two prominent books that have effectively employed the research methodology that we outline here. One is an international relations example, Sechser and Fuhrmann's *Nuclear weapons and coercive diplomacy* the other is from comparative politics, Ziblatt's *Conservative parties and the birth of democracy*. They illustrate virtually all of the key features of LNQA. In addition, in the Appendix we provide some additional, shorter examples of this methodology in action. This methodology now forms a standard research design for dealing with generalization within case study research.

Our analysis of case studies and generalization links very naturally to the philosophical literature on causal mechanisms. As we move through the methodological issues in a political science context we connect with familiar authors and works in the causal mechanism literature in philosophy. For example, our emphasis on within-case causal inference and generalization fits quite naturally with the requirement that causal explanation involve the analysis of mechanisms. As we shall see, "regularities" – be they observational or experimental – require more detailed analysis of mechanisms within cases.

Generalization and (external validity, extrapolation, regularities, transportability, analytic generalization, etc.)

The concepts of external validity, along with its partner internal validity, were introduced into the methodological literature in the classic Campbell and Stanley volume. Campbell and Stanley (1963) define external validity in terms of "generalizability: To what populations, settings, treatment variables, and measurement variables can [an] effect be generalized?"[3]

[3] "The internal-external validity distinction is discussed primarily in social science books on research methods. See also Cook and Campbell (1979). A finer set of categories is sometimes used to distinguish different dimensions of external validity (see e.g., Christensen 2001, Ch. 14): population validity (generalizing to a different population of subjects), ecological validity (generalizing to the behavior of the same subjects in different circumstances, Brunsvik 1955) and temporal validity (generalizing to the same population, in the same circumstances, but at a different time). On external validity in psychology, see also Kruglanski (1975), Henshel (1980), Berkovitz and Donnerstein (1982). The same problem arises in different guises in other sciences. In biochemistry it is known as the in vitro–in vivo problem; see for instance, Strand, Fjelland, and Flatmark (1996)" Guala (2005, 142).

As Shadish, Cook, and Campbell note: "Although internal validity may be the sine qua non of experiments, most researchers use experiments to make *generalizable* causal inferences" (2002, 18–20). However, experiments are generally seen as weak on external validity, in part because sample populations in lab experiments are not seen as representative (Druckman and Kam 2011; Bardsley et al. 2010). However, the problem goes deeper and extends to field experiments as well. What is to assure that an experiment in one setting will necessarily yield the same result in a different one where context is fundamentally different?

In her discussion of exzternal validity, McDermott provides the standard solution: "external validity results primarily from replication of particular experiments across diverse populations and different settings, using a variety of methods and measures" (McDermott 2011, 34). Literature reviews and meta-analysis attempt syntheses of these findings, and implicitly reach judgments of the extent to which diverse experimental findings should be considered robust. Recently, a major research project – the Metaketa project – has attempted to test some core propositions in political science through highly-structured replications. While meta-analysis and replication have gotten more sophisticated, however, there is surprisingly little guidance on how such replications might produce higher or lower levels of generalization.

In their very nice review of the experimental literature on behavioral economics Bardsley et al. call these experiments aimed at increasing external validity "exhibits": "Experimental economics began to expand beyond theory testing as it accumulated a stock of what, following Sugden (2005), we will call 'exhibits,' using as examples the Ellsberg paradox, ultimatum game, and the trust game, among others. An exhibit is a replicable experimental design that reliably produces some interesting result" (Bardsley et al. 2010, epub 409).

Among experimenters in political science the term "transportability" seems to have gained in popularity. We prefer the term generalization because external validity has other dimensions as well, such as how realistic the lab experiment might be. We think it is also the preferred language of those who do case studies: the question most often posed to such studies is exactly their generalizability.

For experimentalists, the definition of generalizability then becomes something like:

> The extent to which the same treatment $X = 1$ produces a similarly significant average treatment effect $Y = 1$ under some scope conditions S.

Our working definition of generalization in the case study context, by contrast, underlines the importance of token causal inference to the process of achieving external validity:

> The same causal mechanism produces the same outcome based on valid within-case causal analysis in all or a high percentage of cases within some scope conditions S.

In the causal mechanism literature in philosophy this is the "regularity" condition that typically appears in conceptualizations of causal mechanisms.

In short, experiments and case studies have problems of generalizability. To be sure, these problems are subtly different. Experiments generate findings in the form of an average treatment effect, which may or may not extend to other settings. Within-case causal inference offers an explanation for a particular case but the mechanisms may or may not yield the same outcome in a different setting. When well done, however, both have high degrees of internal validity. But case studies are not alone in being are vulnerable to the question of generalizability; experiments face this challenge too.

Absolute generalizations

A core claim of this paper is one about methodological ethnography. Scholars doing large-N qualitative analysis, working either with the entire population of relevant cases or a relatively large sample of them (roughly 10+ case studies), often perform what we call in this section absolute tests. A claim is made about a causal relationship or the operation of a causal mechanism in law-like, sufficient-condition, or even necessary-and-sufficient condition, form: if X then Y. Conversely, a number of prominent studies have gone after law-like statements, showing that they in fact fail the sufficient or necessary-and-sufficient condition test when mechanisms are examined more carefully. Yet these practices are rarely if ever justified methodologically or with reference to a corresponding methodological literature. In this and the following section we take up the logic of absolute and relative generalizations, starting with a basic reduced-form example of the former, and then introducing relative as well as absolute tests and the crucial role of causal mechanisms in the method.

Table 1 presents our basic set up in a 2×2 table. A distinctive feature of the approach is both its consideration of the distribution of cases and the particular emphasis it places on the $X = 1$ column. $X = 1$ means the treatment has been given in an experimental context or that X has occurred in an observational setting. Two outcomes are then possible: the treatment has an effect (the (1,1) cell) or it doesn't (the (1,0) cell). We call the (1,1) cell the causal mechanism cell, as consideration of cases from the cell are designed to test for the operation of the postulated causal mechanism. As we shall see below, the $X = 0$ column plays a role when we deal with equifinality, but not in the basic generalization of the causal mechanism.

Many multi-method and qualitative books in recent years include a core case study that illustrates the basic theory. In multi-method books these will sometimes follow the statistical analysis; in qualitative books they are more likely to lead and are often intensive, multi-chapter analyses. These cases inevitably come from the (1,1) cell; they are designed not just to illustrate but to test for the effect of the postulated causal mechanism.

In purely qualitative books, these central causal mechanism case studies generate the question about generalization which then occupies the latter

Table 1: Basic setup

	$X = 0$	$X = 1$
$Y = 1$	Equifinality	Causal mechanism
$Y = 0$		Falsification

Table 2: Relative versus absolute generalizations: balance of power theory

	Non-Hegemon	Hegemon
Balancing	.30	.55
No Balancing	.70	.45

χ^2=28, p = .000, N=445.

Source: Levy and Thompson 2005, table 2.

part of the book. However, just looking at the (1,1) cell ignores the situation where the causal mechanism may not be working, which is critical to the generalizability question. This is the (1,0) cell of table 1, and we return to it in more detail below.

An example from Levy and Thompson illustrates the basics of the generalization logic and the utility of focusing on the $X = 1$ column with a classic hypothesis from realism in international politics. Levy and Thompson test one of the most influential theories in the history of international relations, balance of power theory. As they note in another article "The proposition that near-hegemonic concentrations of power in the system nearly always trigger a counter-balancing coalition of the other great powers has long been regarded as an 'iron law' by balance of power theorists" (Levy and Thompson 2010). This "iron law" is a generalization in the terms of this chapter, a type-level causal claim, and one that is made in strong law-like or sufficient-condition form. A core version of the balance of power hypothesis thus involves balancing against hegemons: *if* there is a hegemon *then* other states will form an alliance to balance it. The short version of the hypothesis is "if hegemon, then balancing."

The logic of "if treatment then outcome" suggests where we need to go to see how generalizable a causal mechanism case study might be. The "if" defines what we call an absolute generalization: if $X = 1$ then the outcome $Y = 1$ occurs.

Table 2 shows balancing 55 percent of the time if there is a hegemon. If the iron law with respect to balancing held, then the probability of balancing would be near 1.0, which is the common-sense meaning of an "iron law." So this proposition is rejected because .55 is not near 1.0.[4]

[4] In the philosophical literature on causal mechanisms one often reads about the initial conditions or a trigger which sets the causal mechanism in motion, for example:

If this were an experimental test, we would be asking whether the hegemon "treatment" were adequate to generate a statistically-significant population-level effect or an average treatment effect. The comparative generalization test compares the percentages in the $X = 1$ versus the $X = 0$ column. This then generates well-known 2×2 statistics of association as well as average treatment effects.

In the relative test in table 2 this becomes the bar of 30 percent in the non-hegemon column. This is of course why it is a relative test; it is the comparison of the percentages in the two columns as opposed to the absolute percentage in one column. Thus hegemonic balancing passes the relative test, i.e., significant χ^2, but not the absolute one. This is because the χ^2 test, like most tests of twoway tables – is comparing percentages across columns, i.e., 30 percent is significantly different from 55 percent.

Note that Levy and Thompson are not posing the question in relative terms, however; they are postulating a law-like regularity. Literature on scientific laws, e.g., (Armstrong 1983) almost inevitably discusses them in terms of how many Y's are also an X, an absolute as opposed to comparative framing. The famous democratic peace example is posed as follows: joint democracy triggers a mechanism (or mechanisms) for not-war 100 percent of the time. The hegemonic balancing hypothesis has form of a sufficient condition: *if* hegemon *then* balancing.

Despite its failure to meet the conditions of an absolute test, does it nonetheless constitute a modestly important generalization? And can we draw a judgment to that effect without relying on statistical, comparative analyses? Those interested in necessary conditions or Qualitative Comparative Analysis (QCA) have thought about this question, and about standards for absolute generalizations, in this case a sufficient condition generalization.

Within QCA there are some common standards, e.g., like p values, for saying there is significant support for a sufficient condition hypothesis. These tend to have a minimum bar of around 75–80 percent (often higher for necessary conditions; see Schneider and Wagemann 2012). Within QCA this constitutes the criterion for passing the sufficient condition test. Since the balancing hypothesis is a sufficient condition one, percentages above this or some other stipulated bar constitute passing the test. We call it an absolute test because it *it only uses information in the $X = 1$ column*.

This example illustrates two key points. First, generalization from case studies is typically framed in terms of absolute not relative generalizations. Second, the absolute and relative criteria for judging generalization do not have to agree because they are *different* criteria. It is possible to have comparative effects that are significant and to also see strong absolute effects. However, two other outcomes are also possible. First, it is possible

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (Machamer et al. 2000, 3; displayed definition)

The key point is that there is some initial triggering condition which then generates the mechanism. This always has logical form *if* triggering condition *then* mechanism *then* outcome. The "regular changes" means a generally reliable causal mechanism.

to have a high absolute score and still conclude that the relative evidence is weak. Conversely, the relative test might appear strong but there are too many falsifying cases (i.e., (1,0) cases) to satisfy an absolute criterion.

The hegemony example is a hypothesis that does not pass the absolute test. One might wonder whether this is "too hard" a test. We do not think it is an issue that the test is hard, but it could be problematic if it were virtually impossible to pass. Below we consider a prominent book by Ziblatt in some detail and pursue this question. The basic hypotheses is "if strong conservative party before mass democratization then stable democracy." Ziblatt does not do an explicit test such as those presented in our tables, and it is not clear exactly what population is fulfilling the "if" which defines $X = 1$ column, ie. the scope conditions. Nonetheless, it appears from his discussion that the proposition would pass the absolute test for both pre-war European democratic experiences and a sample of post-war cases. Ziblatt discusses a number cases in varying degrees of detail but mentions no clear falsifying example.

Until this point, causal mechanisms and within-case causal inference have not made an appearance. Whether we are performing an absolute or relative test we are still looking at patterns across data and are not looking at causal inferences within any of the given cases. However, the set-up is critical because it tells us where to go to do the within-case causal inference. Again, this is the critical role of the $X = 1$ column in defining the population for generalization and thus for case selection.

Relative tests and within-case causal inference

We shall treat the notion of a token cause to be roughly equivalent to within-case causal inference, which means making causal claims about individual cases. The modern philosophical literature on token causal inference, starting with Anscombe (1971) and Lewis (1973), rests basically on the possibility of doing counterfactuals as a way of generating causal inference in individual cases. However, as Holland has famously discussed this is "impossible":

> *Fundamental Problem of Causal Inference.* It is impossible to *observe* the value of $Y_t(i)$ and $Y_c(i)$ on the same unit and, therefore, it is impossible to *observe* the effect of t on i. (Holland 1986, 947)
>
> The important point is that the statistical solution replaces the impossible-to-observe causal effect of t on a specific unit with the possible-to-estimate *average* causal effect of t over a population of units. (Holland 1986, 947)

One natural consequence of this statement of the problem of inference is that it is "impossible" to do within-case causal inference because one cannot construct a real counterfactual case for the comparison. As Holland notes, the best we can do is to compare control groups with treatment groups, hopefully with randomization of treatments, and to derive average treatment effects in populations. Causal inference is based on cross-case comparisons.[5]

The literature on process-tracing and causal process observation, which adopts a mechanism approach to causation, rejects these assumptions. Rather, it assumes that token causal inference is possible. Overwhelming evidence from our everyday lives as well as natural science supports this claim. We assume that individual events can be explained "internally" without reference to population patterns. One of the most famous examples is the space shuttle Challenger explosion. Scientists devoted tremendous energy into why this single event happened, and in less than three years another Space Shuttle Discovery lifted off with a crew of five from Kennedy Space Center. Clearly, the teams investigating the initial failure believed it was possible to find the cause of a singular event.

Glennan makes this point in his discussion of the new mechanical philosophy, which is closely related to the move toward within-case causal inference. Based on the distribution of conforming and non-conforming cases in a statistical analysis such as table 2, we only have what he calls "a bare causal explanation": "Bare causal explanations show what depends upon what without showing why or how this dependence obtains. The causal claims required are established, broadly speaking, by observational and experimental methods like Mill's methods of agreement and difference or controlled experiments. Ontologically speaking, causal dependencies require the existence of mechanisms, but bare causal explanations are silent on what those mechanisms are" (Glennan 2017, 224).

He then discusses at some length an example from the history of medicine where Semmelweis linked the failure to wash hands and instruments to sepsis in Vienna hospitals: "Semmelweis sought to explain the epidemic of puerperal (or childbed) fever among mothers giving birth at the Vienna General Hospital during the 1840s. His first observation was that the division of the hospital to which the women were admitted appeared to be causally relevant, since the death rate from puerperal fever for women in the First Division was three to four times that of women admitted to the Second Division (6.8–11.4% versus 2.0–2.7%)" (Glennan 2017, 224). These statistics imply basically something like table 2.

In his specific example, Clara – a mother with puerpral fever contracted in the non-hygenic division of the hospital – clearly belongs in the (1,1) causal mechanism cell. She is definitely in the treatment group and perhaps even receives more specifically the treatment of unwashed hands and instruments. As Glennan notes this does not necessarily mean that she got the disease via those treatments:

> Let us start with the single case. Suppose Clara contracted puerperal fever (call this event e). What caused her to contract it? A first explanation might simply be that Clara contracted puerperal fever because she delivered her baby in the First Division (call this c). If

[5] An increasing popular approach to within-case causal inference is to create a "synthethic" counterfactual case based on the combination of real cases and then compare that with the actual case. Abadie et al. (2015) construct a counterfactual Germany to explore the impact of German unification on economic growth. This counterfactual Germany is amalgam of "similar" countries such as France, Canada, etc.

> the claim that c caused e is true, that is, if there exists a mechanism by which c contributes to the production e, then that claim provides a bare causal explanation of e. Note that the mere fact that there is a higher incidence of puerperal fever in the First Division is not sufficient to guarantee there is such a mechanism, because it might be the case that that mechanism did not depend upon Clara's being in the First Division. (Glennan 2017, 225)

The kicker comes in the final statement that he makes in discussing this example: "I would argue that until this generalization is attached to particular cases, there is no explanation" (Glennan 2017, 225).

The within-case causal inferences for cell (1,1) cases are important because, as in table 2, correlation is not causation, neither in observational nor experimental research: it is always an *inference* more or less well founded. In an experimental setting those individual cases in the (1,1) cell all count as evidence for the impact of the treatment. This is exactly the point Cartwright is making in her epigraph to this chapter. The average treatment effect must be built upon individual cases where the treatment caused the outcome at least in part.

The within-case causal analysis is thus an examination of cases in the (1,1) cell to ascertain whether the postulated causal mechanism is active; within-case causal inference as used here implies a focus on causal mechanisms.

With that framing, we can now turn to some examples from the increasing body of mixed-method and qualitative research that seeks to strengthen causal inference and generalization by conducting within-case causal inference on a large number of cases. We start with the famous theory of Acemoglu and Robinson that inequality affects the likelihood of democratization. We draw on the multi-method analysis of Haggard and Kaufman (2016), which supplements statistical analysis with consideration of causal mechanisms in individual cases to see how this methodology plays out in one important substantive domain. We also use the case to again underline the difference between absolute and relative tests.

Acemoglu and Johnson's theory is presented in formal terms, through a series of game theoretic models. They do not explicitly state their core arguments in absolute terms, but arguably their models do show crisp equilibria that should rule out certain transition paths. Put in relative or possibilistic terms, however, their claims involve both inequality and how transitions occur. First, they argue that transitions to democratic rule are more likely at moderate levels of inequality than in highly-unequal or highly-equal countries; at high levels of inequality, elites will resist the attendant distributional outcomes; at low-levels of inequality, demands for redistribution via regime change are muted. As stated, however, levels of inequality constitute only a permissive condition for democratization. Acemoglu and Robinson also argue that inequality is ultimately related to democratization via the mechanism of mass mobilization; it is through mass mobilization or the exercise of what Acemoglu and Robinson call "de facto power," that authoritarian rulers are ultimately dislodged. In the absence of such pressure, why would autocrats forego the advantages of incumbency?

Table 3: Inequality and democratization: absolute test

	Low inequality $X = 0$	Medium inequality $X = 1$	High inequality $X = 2$
Democratization	19	21	29
No Democratization	34	41	29
Absolute test	.35	.34	.50
Total	53	62	58

χ^2=3.8, p = .15, N=173.

Source: Haggard and Kaufman 2016, inequality data Solt 2019.

We can now replicate the analysis on hegemony and balancing with data on inequality and regime change, but now based on both absolute and relative tests and within-case causal analysis of the theory linking inequality and regime change via mass mobilization. This exercise is given in table 3, with data from Haggard and Kaufman 2016.[6] again, both claims derived from the Acemoglu and Robinson model are subject to scrutiny here: those having to do with the greater likelihood of transitions in medium-inequality countries; and that they should occur via mass mobilization.

We can pose their claims in both absolute and relative forms. The absolute hypotheses would be that all democratic transitions if there is medium-levels of inequality then there is democratization through the mechanism of mass mobilization. The relative hypotheses would be that they are more likely to occur in moderately unequal authoritarian regimes and via the mechanism of mass mobilization. Note that "relative" means relative to other paths to the outcome, and thus implicitly raises the issue of equifinality or other paths to the outcome.

Table 3 gives the same basic analysis as table 2 above for the hegemony and balancing hypothesis. Now can begin to consider a stipulated causal mechanism. If a country is democratic for the entire inequality period, it is deleted; we are only considering the population of authoritarian regimes, seeing the conditions under which they might transition. We use the terciles of the inequality data for all authoritarian governments to constitute the inequality categories. Given that inequality changes quite slowly we have treated each country as one observation; we consider each of the three inequality categories as a "treatment." If a country never changes inequality category and never has a transition, it counts as one observation of $Y = 0$. If there is a transition, then that constitutes a positive value on Y

[6] Inequality data are the Gini market data from Solt 2019. Transitions based on Haggard and Kaufman 2016. Division into the three inequality categories is based on the terciles of the Gini inequality data for authoritarian regimes, 1980–2008, the terciles are divided at less than 42.6, and greater than 47.6. In the polity data we consider all –66, –77, and –88 observations as authoritarian.

for the whole inequality period. If a country's level of inequality changes to another category (tercile), however, that constitutes a new treatment. It is thus possible that a given country constitutes one observation if it's inequality category does not change or potentially three or four observations if it changes inequality categories. Thus the number of years per observation, i.e., country-inequality category, can vary significantly depending on how long the country stays within a given inequality category. However, the number of years in each inequality category overall is based on equal treatment: because we use the terciles of authoritarian regime years, the basic categories have basically the same number of years.

The $X = 1$ column focuses on the core Acemoglu and Robinson inequality hypothesis in its absolute form. The $X = 0$ as well as the $X = 2$ columns present the incidence of transitions which do not occur at intermediate levels of inequality. As with the hegemony example above the key thing is the percentage of democratizing cases in the $X = 1$ column. The medium inequality column does not pass the absolute test, with the proportion of cases being about one-third. For the relative test the χ^2 statistic for the table is not significant either indicating that the proportions for the other columns are not radically different. It is higher for the high inequality category, but there is no difference between the low inequality category and the medium at all.

The game theory model in the book might be read to make an absolute claim regarding the high and low inequality columns: there should be no transitions in these situations. This would be an absolute test with probability of 0.0. One can do a probabilistic test of these absolute hypotheses saying that if the probability is less than, say, .25 (symmetric to the .75 used for the QCA sufficiency bar), then it passes the absolute test for these columns. As is clear from the table these columns do not pass this absolute test.

However the mere incidence of cases across different levels of inequality does not test for the presence of the stipulated causal mechanism, namely, mass mobilization. The theory rests on the presumption – quite reasonable – that such transitions are not simply granted from above, but ultimately reflect the exercise of "de facto power" on the part of mass publics; in Przeworski's (2009) terms, democracy is "conquered" not "granted." The (1,1) cell for Acemoglu and Robinson (2006) is thus ultimately not only a cell in which there is an intermediate level of inequality and a transition, but one in which there is an intermediate level of inequality and in which the transition occurs through mass mobilization. Using the usual symbols for tracing a causal mechanism, this can be indicated as $X \rightarrow M \rightarrow Y$ where X is moderate inequality and M is mobilization.

The within-case question is therefore whether the observed cases in the (1,1) cell were caused by inequality via mobilization; answering this question requires consideration of each observation individually, in short, token causal inference analysis. This could be done via process tracing, counterfactuals, or other within-case causal inference strategies; Haggard and Kaufman do it through construction of a qualitative data set that interrogates each case for the presence or absence of mass mobilization.

Table 4: Inequality, mobilization, democratization: token causal inference and causal mechanism tests

	Low inequality $X = 0$	Medium inequality $X = 1$	High inequality $X = 2$
Democratization via mobilization	**13** (19)	**8** (21)	**14** (29)
No Democratization	34	41	29

χ^2=1.96, p = .38, N=65.

Source: From Haggard and Kaufman 2016, using Gini coefficient as measure of inequality.

Note: Total transitions cases from table 3 in parentheses.

When using the within-case generalization strategy several things can happen. The first is that the cases in the (1,1) cell were generated via the mechanism of mobilization. These token analyses support the basic theory. But it could also be that an examination of the (1,1) cases reveals that moderate inequality leads to democratization but not via their proposed mechanism. We shall deal with this important issue in the next section, where we note that mobilization may appear as a mechanism in the $X = 0$ column as well as the $X = 2$ column. This could support the mobilization mechanism, but not the connection between democratization and a particular level of inequality and mass mobilization.

The causal mechanism test for generalization generally involves the simultaneous analysis of both X as well as the mechanism, M. In table 4 we include those cases in the (1,1) cell that were generated by mobilization. We also include those in the other columns that were also generated by mass mobilization. In parentheses we have included the original Ns from table 3.

While table 4 looks like a regular two-way table it is fundamentally different from table 3 above. To emphasize this we have put the number of cases in which democratization occurs via mobilization in boldface. We do this to stress that these are counts of token, within-case causal inferences: the number of cases that exhibit the postulated mechanism. Unlike table 3 which might be used to make a causal inference, *table 4 is a summary of token causal inferences*. The boldface numbers would be like summarizing the results of a number of experiments. The question then becomes does this summary of token causal inferences permit us to make a type causal inference about moderate inequality operating through the mechanism of mass mobilization?

Table 3 looks at the basic hypothesis with the mechanism still black boxed: $(X = 1) \rightarrow (Y = 1)$. This happened 21 times. We now include the mobilization mechanism into the mix. This is asking if in these 21 cases we saw this: $(X = 1) \rightarrow (M = 1) \rightarrow (Y = 1)$, where $M = 1$ means that the mobilization mechanism was part of the reason why Y occurred. Generating these counts requires process tracing and even counterfactual dependence Analysis, in short, token causal inference with respect to each case.

If we include mobilization token causal inferences into the mix, the data in table 4 fails to show support either for the reduced-form inequality hypothesis nor for the mobilization mechanism. In other words, Table 4 shows that of the 21 $(X = 1) \rightarrow (Y = 1)$ cases only 8 had $(X = 1) \rightarrow (M = 1) \rightarrow (Y = 1)$. Only 38 percent of the transitions in the medium-inequality category come via their postulated causal mechanism. Hence some other mechanism generated the outcome in the other 13 cases.

It is also worth noting that the percentage of mobilization mechanism cases is higher in the high inequality category than it is in the middle category. If one did a χ^2 statistic just comparing the middle category to extreme inequality separately the χ^2 statistic begins to look quite significant, but not in the direction that Acemoglu and Robinson suggest; more transitions are likely to take place via mass mobilization in the high-inequality cases, even though they were not expected to take place there at all! This suggests that mass mobilization may play a role in some democratic transitions, but again, that inequality does not appear to play a significant causal role.

Recall that we have set the bar very high – at .75 – for having an empirically supported absolute-type causal-mechanism generalization. One might have a different bar for saying there is "no evidence" for their theory. Some large-N qualitative analysis has in fact taken this form, seeking to reject the nature of a particular causal claim altogether once causal mechanisms are analyzed in more detail. Trachtenberg (2012) for example claimed exactly this in his analysis of the audience costs model of major power crises. Examining a population of major power interstate, militarized crises, he claims to find no evidence whatsoever for the presence of audience costs; rather than the probability being near 1.0 he finds that it is 0.0. One could ask if the absolute test provides evidence for being significantly above, say .10. This would then be simply a proportion test of whether the actual value of .34 is less than or equal to .10. With a .10 bar we cannot claim that the generalization of "no support at all" ($p < .10$) is confirmed.

The analysis opens doors onto other lines of research. Following the logic of large-N qualitative analysis, we might investigate in much more detail the eight cases that support their mechanism. These are the confirming cases and it could be worth exploring how the details of these five match the details and discussion of their mechanism. Haggard and Kaufman ultimately use the information they collect on so-called "distributive conflict transitions" to theorize about other causal factors that might influence this class of transitions; we return to this feature of the approach in more detail below.

Nonetheless, the findings are damning: Haggard and Kaumfan – using a combination of statistical and within-case causal analysis – find at best modest support for the Acemoglu and Robinson theory, and neither with respect to their claims about inequality nor their claims about the role of mass mobilization. How damning depends the extent to which Acemoglu and Robinson were claiming to providing *the* mechanism for democratization or *a* mechanism, i.e., 8/21 for the Acemoglu and Robinson mechanism and 13/21 for some other unspecified mechanism. There is certainly evidence for the latter but not the former.

In addition, the analysis showed quite a few instances of $((X = 2)\, OR\, (X = 0)) \rightarrow (M = 1) \rightarrow (Y = 1)$. There in fact more instances of other inequality levels generating the mobilization mechanism. So Acemoglu and Robinson get the mechanism right but the triggering conditions wrong. These simple analyses – from which not too much should be concluded – indicate that the most at-risk authoritarian countries are in fact the high inequality ones, directly contrary to the theory. Also, the high inequality countries have the highest percentage of transition via mobilization of the three inequality categories, again opening interesting research questions about why democratization occurs in such settings.

To summarize, the LNQA methodology often involves investigation of absolute tests to see if there is any prima facie evidence for a strong generalization. However, it can also be used to support relative generalizations, particularly in mixed method designs with complementary statistical analysis. The next move, however, is crucial: one should do the within-case causal inference approach to see if the hypothesized mechanisms are present or not in the cases. The absolute test focuses on X while the causal mechanism tests focus on the M. As can be seen, these are related but separate analyses; a reduced form finding may or may not be supported when we turn to evidence of the presence of the mechanism.

Equifinality and multiple pathways to the outcome

Our analysis above only focused on the connection between a particular theoretical hunch linking moderate inequality, mass mobilization, and democratization. However, it is critical to understand how multiple causal paths can lead to the same outcome. How does the ever-present possibility of equifiality figure into the methodology?[7] Note that mass mobilization was seen to operate in only about half of the cases across all levels of inequality, implying that some other causal mechanism or mechanisms were at work in the democratization process. For example, a number of scholars have argued that international pressure is another cause of democratization.

The question of equifinality arises when there are cases in the $X = 0$ and $Y = 1$ cell. That means something other than X is causing Y. In general most scholars do not claim there is only one path to the outcome; in various ways they assume equifinality. For example, international relations scholars would probably find it objectionable to claim that hegemony is the only circumstance under which balancing could occur. Similarly, it would

[7] As discussed in some detail in Goertz and Mahoney (2012), equifinality is just assumed in many contexts. For example, it lies at the core of the INUS model. A non-INUS model is one like $Y = X_1$ AND X_2, where X_i are individually necessary and jointly sufficient for Y, there is only one path to Y. Why it is central in qualitative methods in general, and QCA in particular, is the idea that the number of mechanisms generating the outcome is limited to a few like in QCA and not a huge number like in general statistical models.

be odd to claim that democratization only occurs in moderately unequal countries – and through mass mobilization – when there are plenty of instances where this is manifestly is not the case. That is exactly what table 3 shows.

In the discussion of balancing against hegemony, it is quite clear that the basic hypothesis was in fact posed as an absolute one. However, it might be the case that the argument is rather a relative one. In the democracy case, moderately unequal authoritarian countries are more likely to democratize than very equal or unequal countries. If posed in these terms, the hypothesis demands a relative test. Haggard and Kaufman provide an extensive set of relative tests using standard statistical techniques and reject the relationship between inequality and democratization in this form as well.

The question here is how causal mechanism and token analysis might fit into a comparative, relative test? We have already suggested an answer: that while their particular theory linking moderate inequality to democratization through mass mobilization is rejected, it is the case that mass mobilization is the causal mechanism at work in over half of all transitions. Haggard and Kaufman go on to argue that mass mobilization may not be due not to inequality but to the robustness of social organization: mass mobilization is more likely where unions and other civil society organizations are present. In effect, Haggard and Kaufman also use the distribution of cases not simply to cast doubt on the Acemoglu and Robinson model, but also to identify alternative causal pathways to democracy: from social organization, through mass mobilization to democracy.

Let's call this additional path Z, perhaps international pressure, social organization or some other mechanism. So now instead of a two-way table we would have a three-dimensional table with the third dimension being the alternative path to Y. This is in fact what Haggard and Kaufman do: they theorize that there are two causal pathways to democracy, one involving mass mobilization and the other not. They then go on to explore some of the underlying causal factors at work in cases not characterized by pressure from below, including international pressures and the calculations of incumbent elites.

Opening another theoretical front necessarily raises questions of overdetermination. Overdetermination means there are cases where $X = 1$ and $Z = 1$ are producing $Y = 1$. The key point is that a particular case might lie on two or more pathways or have two mechanisms present at the same time. This can be thought of as $(X = 1) \rightarrow ((M_1 = 1) \text{ AND } (M_2 = 1)) \rightarrow (Y = 1)$. Counterfactually this means that Y still occurs if we counterfactually make one but not both mechanisms, M_i, equal to zero. In QCA analyses, for example, this is quite common.

As discussed in all the process tracing literature, a key role of this methodology is to evaluate and adjudicate between competing explanations and the individual case. That is exactly the problem here. If possible it would be useful to determine the extent to which one pathway is really the dominant explanation or the other. Of course one might conclude that there is a mixed mechanism involving them both.

These overdetermined cases would be thrown out of statistical analyses because they are not informative. But within-case causal analysis allows the researcher to take a more nuanced approach. For example, within case causal analysis could conclude that the other factor was present (say international pressure), but had no causal influence over the outcome. For example, Schenoni and his colleagues (2019) looking at the resolution of territorial conflicts in Latin America argue that they result from the conjunction of militarization of the territorial dispute, regime change toward democracy, and international mediation. One common critique is that they have omitted the core explanatory factor of US hegemony. US hegemony is a Z variable in addition to their three X variables. In a series of within-case analyses they argue that this was not the case for the individual settlements that they study: US actions were not a cause of territorial conflict resolution. In the case of democratization, moderate inequality might be present, but within-case analyses might argue that it had no causal impact on the outcome; it was the other path – e.g., international pressure – which had causal effect. Of course, it is possible that both are part of the explanation of the outcome.

As this short discussion stresses, the approach outlined here leads quite naturally to discussion of equifinality in a way that standard statistical tests do not. Balancing may be causally related to rising hegemons, but not only. Similarly, democracy may be caused by mass mobilization, but not only. Considerations of equifinality lead the researcher to think about how to theorize alternative pathways and to establish the scope conditions under which one or another pathway emerges. To deal with issues of equifinality empirically also requires within-case causal analysis to disentangle the impact of confounders.

The causal mechanism approach can also be brought to bear in the context of relative tests, typically statistical tests, to look at the overall hypothesis regarding X. It can easily be the case that the statistical analysis suggests a strong relationship while the within-case causal analysis of the (1,1) cases does not support the basic hypothesis in strong form. For example, even if there had been a statistical relationship between inequality and regime change, Haggard and Kaufmans would still show that it did not occur through mass mobilization. Causal mechanism tests can easily serve on their own to check on experimental or observational analyses.

Case selection strategies when there are too many cases for intensive within-case inference

Absolutely core to the within-case generalization strategy is establishing a list of cases where one should see the mechanism in action. This is the critical role of the *if* discussed above. Case selection establishes the universe of cases where one should see the mechanism in action based on the triggering conditions or whatever the scope of the mechanism might be. As practiced, a common feature of LNQA is the focus on rare events: the panel, typically a country-year panel – may include thousands of cells, but instances of the outcome are relatively rare. It might seem that the study of rare events would – virtually by definition – constitute an area of niche concern. In fact, nearly the opposite is the case. Many phenomena that are central to the disciplines of economics, political science and sociology are in fact rare events. In economics, examples include financial crises, episodes of unusually high growth, famines, or – rarer still – the emergence of international financial centers. In political science, transitions to and from democratic rule have been relatively infrequent, to which one could add coups, civil wars and – again, rarer still – social revolutions. International relations is similarly preoccupied with events that are fairly uncommon, most notably wars, but also phenomenon such as the acquisition of nuclear weapons or – again, rarer still – the rise or decline of hegemonic powers and possible reactions to those power shifts.

The precise definition of a rare event is of course relatively elastic, and practical considerations necessarily come into play. For example, the number of social revolutions in world history is relatively small, arguably less than ten. Considering such events permits much more complex causal arguments. Other events might be more common; for example, Haggard and Kaufman consider 78 discrete transitions, but focusing in on the presence or absence of a very particular causal mechanism. In the examples we discuss below, the number of cases considered falls in the 10-30 range. Ziblatt similarly looks at democratic transitions. Sechser and Fuhrmann consider an even smaller population of coercive nuclear threats. When the total number of cases in the $X = 1$ column is relatively small it becomes possible to examine them all in some detail via within-case causal inference. Sechser and Fuhrmann illustrate this very nicely by having a whole chapter devoted to the (1,0) cases and another chapter devoted to all the (1,1) cases. In a slightly different set up Ziblatt does more or less the same thing.

If one moves to phenomena in which the $X = 1$ cases are large, replicating within-case causal inference across all cases becomes impractical. There are two related approaches for addressing this problem, and they can be outlined by considering the democratic peace literature. The democratic peace illustrates a setting where the $X = 1$ column has many cases, all democratic dyads, and the outcome variable not-war (or peace), i.e., $Y = 1$, is quite common. This means virtually by definition that democracy will pass the absolute test as described above. In this relatively common scenario both in the $X = 0$ and $X = 1$ columns you have very high percentages. That means both $X = 1$ and $X = 0$ pass the absolute test. In the QCA framework this would raise the question of potentially trivial sufficient conditions. So after passing the absolute test one would move to the trivialness test which involves the other column (see any QCA textbook for procedures for dealing with this issue, e.g., Schneider and Wagemann 2012)

Another way to deal with these cases is to use the same basic logic with the $Y = 1$ row instead of the $X = 1$ column, where $Y = 1$ is now war. This makes it a necessary condition test, but by definition one with relatively few cases because Y, war, itself is rare. If $Y = 1$ is common then by definition $Y = 0$ (war in the democratic peace) is rare and the visibility of falsifying examples, i.e., (1,0) cell cases, goes up dramatically up. This can be seen in the democratic peace literature where there was a tremendous amount of attention given, by critics in particular, to potential falsifying cases of democracies fighting each other (e.g., Ray 1993).

When there are "a lot" of cases in the (1,1) as well as the falsifying (1,0) cell, making repeated within-case causal inference impractical, one could randomly select among the population of these relevant cases. Fearon and Laitin (2008) have argued for random case selection for case studies. There are few who have found their argument convincing, at least in practice. When they describe random selection it is among all of the cases in the 2×2 table. This has meant choosing cases that are not directly relevant to the causal mechanism because they are not in the $X = 1$ column or the causal mechanism (1,1) cell. However, if one restricts the analysis to say, the causal mechanism cell, then random selection makes much more sense. One is randomly selecting among all the cases which are used to support the causal generalization in the experiment or statistical analysis.

In short, although LNQA has emerged largely to address rare events, it is possible that the method could be extended to larger populations. If there are too many cases to examine individually in the $X = 1$ column or in the causal mechanism cell we think the first good response is to think about randomly selecting cases for intensive within-case analysis. Of course, this is subject to practical constraints, but nevertheless is a very good starting point.

Large-N Qualitative Analysis: some examples

In this section we give two more extended examples of large-N qualitative analysis in practice. Our purposes are several. First, from the standpoint of our anthropological approach to the method, they show how wide-ranging the applications of this approach have been, showing up in fields as diverse as the study of nuclear weapons and historical analyses of democratization. In addition to reiterating our analysis of the method, these cases also show how it is used both in a multi-method context and where the central approach is rooted in case studies, in this case an historical analysis of democratization in the UK and Germany.

Sechser and Fuhrmann (2017): Nuclear Weapons and Coercive Diplomacy

Sechser and Furhmann provide an example of a multi-method approach to large-N qualitative analysis. In the first half of the book, they undertake a large-N statistical analysis. They report on detailed statistical tests of the effect of possessing nuclear weapons on two outcomes: whether nuclear states make more effective compellent threats; and whether they achieve better outcomes in territorial disputes. Their statistical tests fail to find a nuclear advantage.

We ignore these chapters, however, and focus on the two main case study chapters which form at least half of the volume and which are structured along the lines we have outlined here. These chapters are self-consciously addressed to questions of the postulated causal mechanisms behind their "nuclear skepticism" argument, including particular questions about the credibility of nuclear threats and possible backlash effects of using them. They carefully delimit the scope of cases to those in which countries attempted nuclear coercion. They explicitly adopt the LNQA method we have outlined: "[W]e delve deeply into history's most serious coercive nuclear crises. Coercive nuclear threats are rare: nuclear weapons have been invoked to achieve coercive goals less than two dozen times since 1945. We study each of these episodes, drawing on declassified documents when possible" (Sechser and Fuhrmann 2017, 20).

Thus the $X = 1$ cases are those of attempted nuclear coercion and brinksmanship. The outcome is whether the state was able to coerce the target into changing its behavior. It should be emphasized that their theory explains nuclear coercion threat failure. This makes failure the $Y = 1$ cases (we do this to remain consistent with what we have done throughout this paper). It then makes complete sense that their first case study chapter involves the causal mechanism cases where a coercive threat was made but failed. The next chapter by contrast takes up potentially falsifying cases: cases where the threat was made and appeared to succeed, making it the (1,0) set of cases (threat, but success instead of failure). They use this chapter to show that – following detailed within-case causal inference – these nominally falsifying cases may not be falsifying after all.

Their justification for the approach fits the large-N qualitative analysis that we have outlined: "The purpose of quantitative analysis was to identify general trends – not to explain any single case. Even if the quantitative analysis suggests that our argument is generally correct, nuclear skepticism theory may fail to explain some cases. Why does this matter? The unexplained cases – often referred to as outliers – may be particularly salient" (p. 130). Put differently, Sechser and Fuhrmann underline that we do not simply want relative comparisons; we want convincing explanations of cases that are deemed important on substantive grounds.

They pay particular attention to case selection, and seek to consider the entire universe of cases in which states attempt nuclear coercion. They identify 13 cases that are clear examples and another six "borderline" cases; the two groups are pooled. These include the (1,1) cases which confirm their theory of nuclear coercion failure: "In this chapter, we discuss nine nuclear coercion failures. These are cases in which countries openly brandished nuclear weapons but still failed to achieve their coercive objectives" (p. 132). Case studies of these failure cases show the operation of the causal mechanisms postulated in their theory of coercion failure, such as the fact that the threats were not credible and were resisted by presumably weaker parties.

In the next chapter – appropriately called "Think Again: Reassessing Nuclear Victories," they turn to the (1,0) cases: "cases in which nuclear blackmail seemingly worked" (p. 173)." In these cases, causal mechanism analysis becomes central: the cases are designed to see if the mechanism proposed by nuclear coercion theorists really explains the outcome (ie., that nuclear coercion "worked"). They could have probed for some conditional factors suggested by their theory that might have operated in these success cases, thus establishing scope conditions on their skepticism. But their within-case causal analysis concludes regarding the (1,0) cases that "in each instance, at least one of three factors mitigates the conclusion that a nuclear threat resulted in a coercive victory. First, factors other than nuclear weapons often played a significant role in states' decisions to back down. Second, on close inspection, some crisis outcomes were not truly 'victories' for the coercer. Third, when nuclear weapons have helped countries in crises, they have aided in deterrence rather than coercion" (p. 174).

By looking at the cases closely they in effect also identify measurement error. When they say that these were not cases of coercion success that means that that instead of being $Y = 0$ cases they are in fact $Y = 1$ cases in which nuclear coercion failed. Hence they are removed from the falsifying cases population. A related critique is that the outcome is not successful coercion but rather successful deterrence. This is more nuanced, but is arguing again that when Y is coercion success one should not count deterrence success as its equivalent. The treatment (attempted nuclear coercion) might produce other positive outcomes, but that is not what is being tested; the test is of the hypothesis that nuclear weapons can compel, not deter.

If one includes the clear 13 failure cases in the previous chapter, the absolute test would score at best 10 of 23 cases (44 percent). Once one takes mis-measurement into account, however, Sechser and Fuhrmann claim that not a single one of the apparently successful cases in fact constitutes a success. Their within-case analysis brings the total down from a potential 10 to virtually zero. In other words, they find no clear-cut case of nuclear coercion success. This illustrates the potentially dramatic impact of doing causal mechanism, within-case causal analysis.

They end their book with this reflection: "It is worth noting that the cases that provide the strongest support for the nuclear coercionist view – including the Cuban missile crisis – happened in the early days of the Cold War. . . . There is scant evidence that nuclear blackmail has worked since the collapse of the Soviet Union. The nuclear crises of the last quarter-century illustrate the coercive limits, rather than the virtues, of nuclear weapons." They are clearly thinking about how generalizable their findings are over time. They think that they are generalizable: it is hard to find any success case in the last 30 years of international politics.

Ziblatt (2017): Conservative Parties and the Birth of Democracy

Ziblatt's influential, prize-winning book on the role of conservative parties and democratization is an example of a study that is built from the start around intensive examination of two cases, and then broadened out to consider other cases using the LNQA generalization strategy.

Ziblatt is interested in the role of conservative parties as countries move from authoritarian regimes to democracy. He argues that democracy was not the result of underlying structural factors, such as socioeconomic change, nor class pressures, whether from the middle or working classes. Rather he argues that democracy depended on how successfully conservative political parties were able to recast themselves to adapt to electoral pressure while holding off authoritarian tendencies in their own right wings. His two core case studies occupy most of the book and include the UK, where the conservative party recast itself, and Germany which saw failures in that regard. UK is discussed in chapters 3–5, Germany appears in chapters 6–9.

In the final chapter (appropriately titled "How Countries Democratize: Europe and Beyond"), he considers generalization case studies. The generalization goal of the final chapter is stated clearly in the introduction:

> Strictly speaking this book's argument has made sense of political developments within Britain and Germany between the middle of the nineteenth and the middle of the twentieth centuries. But a second purpose, as promised in the introduction, was that the interpretation of these specific historical experiences has general implications for how to think about the enduring impact of old-regime forces on democratization in other places and times. Is our understanding of the world in fact deepened when we widen our scope beyond the main cases we have studied? What more general implications can we draw? (Ziblatt 2017, 334–35)

In that final chapter he outlines the scope of these potential generalization case studies for Europe: "Table 10.1 provides a list of the major parties of the electoral right after 1918, noting which electoral right party had the greatest number of votes in the first postwar democratic elections, a score for the fragmentation of the right camp of parties in this period, and whether or not democracy survived the interwar years" (p. 336). He then proceeds to choose four or five of these for case studies lasting a few pages. Then he considers the Latin American cases. Here the analysis is very short and arguably more superficial; however the purpose – as in Haggard and Kaufman – is focused: to test for the operation of the favored causal mechanism related to conservative parties.

He chooses causal mechanism cases for further generalization (aka "on-line" cases) from the (1,1) cell:

> We begin by analyzing two "well-predicted" cases that appear to fit the framework: one where the right was relatively cohesive and democracy survived (Sweden), and one where it remained organizationally fractious and democracy ultimately collapsed (Spain).

That is, for his generalization case studies he starts by choosing a case that is close to UK – Sweden – and one that is seen as similar to Germany, Spain. The Sweden case study takes up four pages, while the Spanish one eight.

Ziblatt then moves to consider non-European cases. He summarizes the patterns briefly: "In the four countries where conservative political parties emerged *before* mass suffrage – Chile, Colombia, Costa Rica, and Uruguay – democratization even if predominately oligarchic at first, was on average more stable than in the rest of the region. By contrast, in the remaining twelve countries – Argentina, Brazil, Ecuador, Peru, and so on – where no conservative political party existed until *after* mass democratization, democracy was, on average, less durable" (Ziblatt 2017, 358–59). He then proceeds to spend a couple of pages on Argentina as the Latin American example. He then spends one paragraph on some Asian cases like South Korea and Taiwan and then two paragraphs on the Arab Spring, in each case testing for the presence of absence of conservative parties and the presence or absence of the outcome, democratization.

This example illustrates nicely that the case studies need not be treated equally. The amount of space devoted to each case study can be skewed as long as adequate information is provided to reach a reasonable conclusion with respect to the presence or absence of the postulated causal mechanism. The book is framed around two core causal mechanism case studies, followed by a series of generalization case studies that are unequal in length but nonetheless focused on the core causal relationship, a strategy that reflects standard practice over the last few years in case study books. In each generalization case study, the purpose is to focus on the postulated causal mechanism and see if it works in that case; generalization is enhanced by effectively "summing" these well-explained cases.

Conclusion

In this chapter we explore a research paradigm involving within-case causal inference and a generalization strategy using case studies to support generalization claims about causal mechanisms. Probably the most distinctive feature of the approach are two. The first is the effort to use a large number of cases, and even the entire population of the $X = 1$ cases. The second is the systematic use of within-case causal inference as opposed to experimental designs – in which treatment is contrasted with control – or cross-case observational designs, such as those deployed in many studies in comparative politics and international relations.

This research design typically takes advantage of the fact that there are often relatively few cases in the causal mechanism cell. Going back to Ziblatt the universe for half of the book is mass democratization in Europe after 1918 and he is thus able to consider the causal effect of his chosen theory involving the timing of the emergence of conservative parties and their relative strength. Similarly with Sechser and Fuhrmann, the panels they use for their statistical analysis have as many as 6500 observations in some models. But the number of cases in which nuclear states unambiguously made coercive nuclear threats is not more than a couple dozen. This makes the design described here a plausible alternative (for Ziblatt) or complement (for Sechser and Fuhrmann) to standard statistical analysis.

One might argue that absolute tests with just one variable are very unrealistic, particularly with a relatively high bar of 75 percent. An perhaps widely-held intuition is that it is unlikely that many factors will pass this sort of law-like test; note that it is much stronger than claims that a given causal variable has at least a statistically significant effect on a population when potential confounds are either randomized away or controlled for. Not surprisingly, some of the early examples of this work were aimed at taking down expansive law-like claims with respect to a diverse array of outcomes, from the role of economic interests in the design of electoral systems to the role of audience costs in war.

A natural response is to say there must be other factors that are important in producing the outcome, and that one is not enough. QCA deals with this by focusing on interaction terms: it is only when there is an interaction of three or four causal factors that the outcome is very likely to occur and will pass the absolute test. The core point here is that the logic of absolute tests is the same whether it is one variable or the interaction of four or five variables.

As we have noted, this design has become quite standard practice in both case study–only books in recent years as well as those engaged in mixed-methods approaches. But the practice has yet to appear in the methodological literature (again, see Goertz 2017 chapter 7). In almost all cases authors just "do it." The fact that it seems not to have provoked any backlash on the part of reviewers, commissioning editors, or others seems to indicate that the logic convinces most.

We have sought to extract some of the main features of this approach and link them to wider discussions about generalization in philosophy and the social sciences. First, and most obviously, it is best suited to analysis of rare events, which in fact figure quite prominently in the political science canon. Second, it rests on focused tests of postulated causal mechanisms; theory matters. And finally, it requires use of within-case causal inference techniques, including process-tracing predominantly.

We can see several ways in which this work might be pushed forward. One interesting link is to discussions in the Bayesian tradition about what amount of case work might be adequate to reach closure. Dion (1998) nicely showed that in a Bayesian framework five or six consistent case studies can easily lead to 90 percent posterior confidence starting with a uniform prior. Clear theorizing in this area might reduce the demands of this approach and thus strengthen its appeal.

More work might also be done on a version of generalization which can be called extrapolation. Using the classic Shadish, Cook & Campbell (2002) units, treatment, outcomes, and settings (UTOS) framework we can ask to what extent when we move along some dimension of UTOS the same results apply. For example, the quest for external validity and generalization in experimental work has been about doing the experiment or treatment on new units. Extrapolation could be about changing the treatments gradually in some direction within populations of case studies as well.

Often a strong generalization that is empirically founded naturally leads to questions about extrapolation. The democratic peace is a well-founded empirical generalization. This should, but does not seem to have, lead to a question about how generalizable it is to less democratic countries. So as we extrapolate from democratic towards authoritarian how far does the democratic peace generalize and extrapolate? We think that this is a clear next step in the analysis of generalization and external validity.

This methodology also has potential applications for experiments as well as large-N statistical analyses, whether observational or based on designs that are causally "well-specified," such as matching, difference-in-difference or regression discontinuity designs. Currently the solution for the generalization problem for experiments is just to do more experiments. We know of almost nothing that systematically tries to analyze what constitutes successful generalization criteria; we can see room for parallel work looking at how experiments on a given issue aggregate across the $X = 1$ column. But this approach may also contribute to internal validity of experimental findings.

The critical cell is the causal mechanism (1,1) cell. One could easily take a random sample of the (1,1) cases in the statistical analysis to see if the causal mechanism is in fact present. This would represent an independent check on the statistical or experimental results.

In short, we think we have barely begun to systematically analyze the crucial decisions and the crucial options in case study–generalization methodologies and in large-N qualitative analysis (LNQA) in particular. But our analysis suggests that there is a lot – e.g., extrapolation – that requires more sustained thought and analysis.

References

Abadie, A., et al. 2015. Comparative politics and the synthetic control method. *American Journal of Political Science* 59:495–510.

Acemoglu, D., and J. Robinson. 2006. *Economic origins of dictatorship and democracy.* Cambridge: Cambridge University Press.

Anscombe, G. 1971. *Causality and determination.* Cambridge: Cambridge University Press.

Armstrong, D. 1983. *What is a law of nature?* Cambridge: Cambridge University Press.

Bardsley, N., et al. 2010. *Experimental economics: rethinking the rules.* Princeton: Princeton University Press.

Braumoeller, B., and G. Goertz. 2000. The methodology of necessary conditions. *American Journal of Political Science* 44:844–58.

Campbell, D., and J. Stanley. 1963. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Dion, D. 1998. Evidence and inference in the comparative case study. *Comparative Politics* 30:127–45.

Druckman, J., and C. Kam. 2011. Students as experimental participants a defense of the "narrow data base". In J. Druckman et al. (eds.) *Cambridge handbook of experimental political science.* Cambridge: Cambridge University Press.

Dunning, T., et al. (eds.). 2019. *Information, accountability, and cumulative learning.* Cambridge: Cambridge University Press.

Fearon, J., and D. Laitin. 2008. Integrating qualitative and quantitative methods. In J. Box-Steffensmier, H. Brady, and D. Collier (eds.) *The Oxford handbook of political methodology.* Oxford: Oxford University Press.

Glennan, S. 2017. *The new mechanical philosophy.* Oxford: Oxford University Press.

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

Goertz, G., and J. Mahoney. 2012. *A tale of two cultures: qualitative and quantitative research in the social sciences.* Princeton: Princeton University Press.

Haggard, S., and R. Kaufman. 2016. *Dictators and democrats: masses, elites, and regime change.* Princeton: Princeton University Press.

Holland, P. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81:945–60.

Kaplan, O. 2017. *Resisting war: how communities protect themselves*. Cambridge: Cambridge University Press.

Levy, J., and W. Thompson. 2005. Hegemonic threats and great power balancing in Europe, 1495–1999. *Security Studies* 14:1–30.

Levy, J., and W. Thompson. 2010. Balancing at sea: do states ally against the leading global power? *International Security* 35:7–43.

Lewis, D. 1973. *Counterfactuals.* Cambridge: Harvard University Press.

Machamer, P., et al. 2000. Thinking about mechanisms. *Philosophy of Science* 67:1–25.

McDermott, R. 2011. Internal and external validity. In J. Druckman et al. (eds.) *Cambridge handbook of experimental political science.* Cambridge: Cambridge University Press.

Narang, V., and R. Nelson. 2009. Who are these belligerent democratizers? Reassessing the impact of democratization on war. *International Organization* 63:357–79.

Przeworski, A. 2009. Conquered or granted? A history of suffrage extensions. *British Journal of Political Science*, 39:291–321.

Ragin, C. 2000. *Fuzzy-set social science.* Chicago: University of Chicago Press.

Ray, J. 1993. Wars between democracies: rare or nonexistent? *International Interactions* 18:251–76.

Ripsman, N. 2016. *Peacemaking from above, peace from below: ending conflict between regional rivals.* Ithaca: Cornell University Press.

Schenoni, L. et al. 2019. Settling resistant disputes: the territorial boundary peace in Latin America. Manuscript. University of Notre Dame.

Schneider, C., and C. Wagemann. 2012. *Set-theoretic methods for the social sciences: a guide to qualitative comparative analysis.* Cambridge: Cambridge University Press.

Sechser, T., and M. Fuhrmann. 2017. *Nuclear weapons and coercive diplomacy.* Cambridge: Cambridge University Press.

Shadish, W., T. Cook, and D. Campbell. 2002. *Experimental and quasi-experimental designs for general causal inference.* Boston: Houghton Mifflin.

Trachtenberg, M. 2012. Audience costs: an historical analysis. *Security Studies* 21:3–42.

Wallensteen, P. 2015. *Quality peace: peacebuilding, victory and world order.* Oxford: Oxford University Press.

Ziblatt, D. 2017. *Conservative parties and the birth of democracy.* Cambridge: Cambridge University Press.

Appendix: Examples of large-N quantitative analysis

Our research suggests that the Ziblatt, and Sechser and Fuhrmann procedures, aka LNQA, has become a standard model for case study–only books or books with a large case study component. In this appendix we provide a brief discussion of some notable books that rely on case studies for generalization. These might be books that exclusively rely on case studies as a generalization strategy, some might also include some statistical analyses.

As we noted above, this paper is also an exercise in methodological ethnography. What we describe and synthesize here also reflects influential practice for the last few years. One thing that makes our chapter methodological ethnography is that these case selection decisions are rarely justified. Often the universe of potential case studies is ambiguous and not clear. Ziblatt does it in one sentence referring to "online" cases.

This is not to stay that all of these works do exactly what we described above: they usually do not. However, at a general level what we describe is what they are doing. We encourage the reader to sample some of these books.

We start by giving the abstract of the book (if there is one) and then a brief discussion often with extended quotes of the basic logic of case selection and generalization the authors follow.

Kaplan (2017): Resisting War: How Communities Protect Themselves

> In civil conflicts around the world, unarmed civilians take enormous risks to protect themselves and stand up to heavily armed combatants. This is not just counterintuitive – it is extraordinary. In this book, Oliver Kaplan explores cases from Colombia, with extensions to Afghanistan, Pakistan, Syria, and the Philippines, to show how and why civilians are able to influence armed actors and limit violence. Based on original fieldwork as well as statistical analysis, the book explains how local social organization and cohesion enables both covert and overt nonviolent strategies, including avoidance, cultures of peace, dispute resolution, deception, protest, and negotiation. These "autonomy" strategies help communities to both retain civilian status and avoid retaliation by limiting the inroads of armed groups. Contrary to conventional views that civilians are helpless victims, this book highlights their creative initiative to maintain decision-making power over outcomes for their communities. (Kaplan 2017, book abstract)

A key question for this book is how generalizable these local solutions are as the civil war gets more more severe? The author locates Columbia as a middle range civil war in terms of deaths. Many of his generalization case studies involve more severe civil wars (e.g., Afghanistan and Syria), not less severe ones. He is thus more or less implicitly thinking about generalization and extrapolation within the scope of civil wars.

Ripsman (2016): Peacemaking from Above, Peace from Below Ending Conflict between Regional Rivals

Ripsman (2016) provides a nice example of how a potentially relatively large scope gets narrowed, allowing him to perform case studies of all cases within his scope. He starts with a much longer list of rivalries (from a standard dataset) and ends with nine within his scope:

> The object of inquiry in this book is peace agreements between regional rivals. To investigate this, I engage in detailed primary and secondary source case studies of highly salient cases of peace settlements (treaties or their functional equivalents) between regional antagonists to determine their causes and the sources of variation in their postagreement stability. (Ripsman 2016, 7)
>
> This left a list of nine rivalries (see below). Of these, I selected three particularly salient cases resulting in multiple wars between the antagonists (indicated in italics in the list below): the Franco-German, Egyptian-Israeli, and the Israeli-Jordanian rivalries. These are the cases that I examine in chapter-length depth. In each of these cases, because language skills allowed me access to archival depositories and the ability to conduct interviews with governmental officials involved in peacemaking, I have utilized both primary and secondary sources. To make certain that case selection has not biased my conclusions, I examine each of the six remaining rivalries as mini case studies in chapter 5. As a result, in this book I explore the entire universe of twentieth-century regional rivalries that terminated with peace settlements. (Ripsman 2016, 11)