

RESEARCH ARTICLE

Qualitative research methods for institutional analysis

David Skarbek 

Brown University, Providence, RI, USA

Corresponding author. Email: David_Skarbek@Brown.edu

(Received 28 October 2019; revised 16 December 2019; accepted 19 December 2019)

Abstract

How can economists use qualitative evidence – such as archival materials, interviews, and ethnography – to study institutions? While applied economists typically rely on quantitative evidence and statistical estimation, many important aspects of institutions and institutional change appear in the form of qualitative evidence. This raises the question if, and how, social scientists trained in quantitative methods can exploit and analyze this evidence. This paper discusses two qualitative research methods that are both commonly used outside of economics: comparative case studies and process tracing. Drawing on existing research about crime and political revolutions, it discusses these two methods and how to implement them to improve institutional analysis.

Keywords: methods; most-similar systems; process tracing; research design

1. Why care about qualitative evidence?

Economics is often defined by its method rather than by a substantive topic area. For instance, Dani Rodrik writes, ‘economics is a way of doing social science, using particular tools. In this interpretation the discipline is associated with an apparatus of formal modeling and statistical analysis’ (Rodrik, 2015: 7, see also 199).¹ While many people studying institutions would disagree with this emphasis, this is a common interpretation of contemporary economic research. In particular, the applied economics literature is now dominated by a focus on causal identification (Ruhm, 2019). The ‘credibility revolution’ is clearly a major advance in social science. However, the widespread emphasis on estimating causal effects raises several important challenges for social scientists interested in understanding institutions and institutional change. First, in the absence of a plausible identification strategy, it removes a substantial range of important questions from investigation, especially those associated with institutions. Big questions – such as what caused the industrial revolution or long-term political and cultural change (Kuran, 2012; McCloskey, 2016) – become difficult or impossible to study.² Related to this, if we rely only on quantitative data as evidence, then we must necessarily exclude many questions from analysis. We simply lack the relevant data to many questions that we might care about. However, there is sometimes a substantial amount of qualitative evidence available that can be useful for adjudicating debates. We miss an opportunity to learn if we exclude this type of evidence from analysis.

A second challenge raised from focusing exclusively on statistical identification is that there might be multiple possible causal pathways that connect x to y . This is known in the methodology literature as ‘equifinality’ (Gerring, 2012: 216). If so, then even a precisely estimated causal effect tells us only about *what* happened but not necessarily about *why* it happened. For example, economists have

¹Kirzner (2009) discusses the development of economics as it has grown from being topic-focused to more method-focused.

²Tilly (1984) discusses the methodological challenges of studying big structures, large processes, and huge comparisons.

recently studied whether increasing the number of police officers on patrol causes a reduction in crime. Using natural experiments that exogenously increase the number of police, studies find that increasing the number of police officers does reduce crime (Draca *et al.*, 2011; Klick and Tabarrok, 2005). Yet, these studies do not tell us much about why this happens. It might be that crime falls because police are more visible to potential offenders, thereby deterring crime. Or, police officers might be more proactive when there are more officers working nearby, making it more difficult to shirk. There might be organizational economies of scale for how dispatchers distribute calls. These quantitative studies do not distinguish between a variety of possible causal pathways. Of course, sometimes data are available that point to particular mechanisms and economists often attempt to use quantitative evidence to identify causal mechanisms. However, qualitative evidence can also provide information about possible causal pathways. In the case of police effectiveness, for example, we can learn about mechanisms with surveys, interviews, and participant observation (Boettke *et al.*, 2016; Ostrom and Smith, 1976; Ostrom *et al.*, 1978). Economists already occasionally use qualitative evidence to pinpoint causal pathways (typically, informally during presentations), but they often do so in *ad hoc* ways. Qualitative research methods can discipline the use of qualitative evidence to be more systematic and rigorous.

A third problem that arises from the exclusive use of quantitative evidence is that it tends to exclude ‘thick’ concepts and theories from analysis. These are theories that are inherently complex, multifaceted, and multidimensional (Coppedge, 1999: 465). Unlike ‘thin’ theories, their empirical predictions are less straightforward to measure and test. The ‘more police, less crime’ hypothesis is a good example of a thin theory. We can measure and count the number of police and the amount of crime in a fairly straightforward way that captures what we mean when we think of ‘police’ and ‘crime’. But, not all concepts and theories are so simple. In particular, institutions tend to be discussed in a far thicker way. Even if we think of them simply as ‘rules of the game’, that opens a wide range of possible important attributes.³ Institutions can vary in many ways, including ranging from legal to extralegal, formal to informal, centralized to decentralized, flexible to rigid, permanent to temporary, and fragile to robust to antifragile. A good description of an institution must often be multifaceted and multidimensional (see discussions and examples in Ostrom, 1990, 2005). It is also often not possible to accurately or precisely measure variation in these characteristics. Taken together, these issues make it difficult or impossible to reduce thick concepts – like institutions – to simple quantitative measures without losing crucial parts of their meaning.

This is not an argument that all research should use qualitative evidence. Nor do I argue that we should abandon any of the important methods used in statistical analysis.⁴ Instead, I argue that the type of evidence that will be useful depends on the questions that we ask (Abbott, 2004). Institutional research, in particular, regularly asks questions that would benefit a great deal from qualitative evidence. This is partly the case because institutional analysis often focuses on causal *mechanisms* more than on causal *effects*. The goal of many studies is to understand the mechanisms that give rise to and sustain a particular organizational or institutional outcome (Greif, 2006: Ch. 11; Munger, 2010; Poteete *et al.*, 2010: 35). Qualitative evidence is often well-suited for describing an institution’s multifaceted and multidimensional characteristics. Moreover, by studying mechanisms (*m*), we can discover which among several possible *m* connects *x* to *y*. This suggests that qualitative evidence can be a valuable addition to empirical research, especially when studying institutions (Boettke *et al.*, 2013).

³There is some debate on how to best define and conceptualize institutions (Greif and Kingston, 2011). Scholars have likewise debated the most appropriate way to measure institutions (Shirley, 2013; Voigt, 2013).

⁴However, it is important to note that, like with any empirical method, causal identification cannot ‘prove’ causality with certainty, but instead can only increase or decrease our confidence in estimated effects. Related to this, Ruhm (2019) discusses limitations of causal inference, including the trade-off between the desire for precise estimates and the importance of the research question. Deaton and Cartwright (2018) likewise discuss a wide range of limitations on the usefulness of randomized control trials.

In short, if we wish to say something about big historical questions, accurately identify causal mechanisms, or engage thoughtfully with thick concepts and theories, then we should be more open to engaging with qualitative evidence. This can include a wide range of materials, including written records (such as constitutions, laws, statutes, company manuals, archival materials, private correspondences, and diaries), interviews, surveys, participant observation, and ethnography. However, most economists are not trained in how to use these types of evidence in a rigorous way. Rodrik goes so far to say that the economics profession ‘is unforgiving of those who violate the way work in the discipline is done’, and this sentiment might well apply to engaging with qualitative evidence (Rodrik, 2015: 199). Nevertheless, social scientists in other disciplines care about many of the same research design challenges as economists do. This paper discusses two common ways that social scientists analyze qualitative evidence: comparative case studies and process tracing. While these methods are commonly used in other social sciences, they are rarely used explicitly in economics. Nevertheless, they provide a productive way to analyze qualitative evidence and to better understand institutions and institutional change.

2. Comparative case studies

What is a case study?

To get a sense of how qualitative research methods work, it is helpful to think about a common context in which they are used: case studies. A case study is ‘an intensive study of a single case or a small number of cases, which draws on observational data and promises to shed light on a larger population of cases’ (Gerring, 2017: 28). People sometimes conflate the distinction between the number of observations (n) and the number of cases (c) in a study. An observation is simply the unit of analysis in a particular study. Many well-known economics papers are case studies. David Card’s study of the Mariel boat lift is a nice example (Card, 1990). It is a single case ($c = 1$) that uses a large number of observations (large- n). Both of the police–crime papers mentioned above are also large- n case studies (of London and Washington, DC). By contrast, a cross-sectional analysis of national GDP would be a large- c study (with c = number of countries included) but might have only a few observations per case. The best case studies tend to have only one or a few cases and many observations within each case. Social scientists working within the case study tradition typically seek to maximize within-case explanatory power. This sometimes comes at the expense of explanatory power across cases, which is why external validity is one of the biggest weaknesses of case studies, regardless of whether it is based on qualitative or quantitative evidence (Gerring, 2017: 244–245).

As noted, both quantitative and qualitative evidence can be used in case study research. However, to be an ‘intensive’ study, case studies often benefit from the fine-grained evidence that qualitative evidence regularly engages with. There are several reasons for this. First, adding qualitative evidence to quantitative evidence expands the amount of potentially informative evidence. More types of evidence can be used to increase our within-case understanding. Second, qualitative evidence can identify information that is not easily quantified, so there are also different types of evidence available to inform our explanation. Third, especially when studying thick concepts, it can allow for a more accurate and complete understanding of institutions.

To return to the police–crime example, we could learn more about this relationship by collecting several types of qualitative evidence (ideally both before and after the exogenous shock that increases the number of officers working). We might survey or interview police officers about how they spend their time on a shift or how they perceive the difficulty or danger of performing their jobs.⁵ Participant observation in a police ride-along would provide the opportunity to learn more about what police actually do and to discover unexpected explanatory factors (i.e. identify important omitted variables).

⁵Economists have sometimes avoided interviews because they fear that respondents will not answer truthfully. Yet, political scientists (among others) have developed numerous experiments that allow us to elicit sensitive information that respondents might be hesitant to state even in an anonymous survey or interview (Lyall *et al.*, 2013).

Observation could also help to confirm the accuracy of written reports and officers' statements (reducing measurement error). Many police departments have adopted body-cameras because this type of qualitative evidence provides more and better evidence about what police actually do while on patrol. Simply riding around in a cruiser would allow a researcher to see, feel, and hear what police are doing 'on the ground'. This would potentially illuminate which mechanisms are most important for deterring crime. For instance, it might be that certain types of police officers or policing tactics are more important than others for reducing offending. In one ethnography, sociologist Peter Moskos worked as a Baltimore city police officer for a year (Moskos, 2008). He found that police are often insufficiently trained, suggesting that the experience level of the officers who are actually working is likely to be important for deterring crime. He likewise found that there was a significant principal-agent problem between police captains and shirking officers. This led captains to use the officer's number of citations and arrests as an indicator of effort, leading some officers to focus on relatively non-serious, high-frequency violations, such as riding a bike at night without a light (Moskos, 2008: 140). This suggests that police oversight and organization is important for understanding when more police actually does reduce crime. In brief, some or all of these types of qualitative evidence would improve our understanding of the relationship between police and crime. Moreover, to the extent that the police-crime hypothesis is a thin theory, the benefits from qualitative evidence are likely to be even greater when studying the thick concepts and theories commonly found in the study of institutions.

Descriptive inference

Social science entails both descriptive inference and causal inference. Regardless of the type of evidence used in pursuing social scientific explanations, the same general principles apply in thinking about what makes for good propositions or hypotheses. In general, they should be logically consistent and falsifiable. They are better when they apply to a broader set of phenomenon, have many empirical implications, make novel predictions, receive substantial empirical support, and are consistent with other established theories (Levy, 2008: 633). Philosophers Quine and Ullian (1978: 39–49) argue that these characteristics determine the 'virtuousness' of a hypothesis. The same virtues hold true regardless of whether the evidence used to test a proposition is quantitative or qualitative.

How can we make descriptive inferences with qualitative evidence? As with the virtues of hypotheses, we do so by focusing on the same concerns and issues that concern us when making descriptive inferences with quantitative evidence. Social scientists who work with qualitative evidence are keenly aware of concerns about cherry-picking, anecdotes, and *ad hoc* claims based on a few, or biased, sources of evidence. In *Designing Social Inquiry: Scientific Inference in Qualitative Research* (1994), Gary King, Robert O. Keohane, and Sidney Verba provide a helpful introduction to how many social scientists outside of economics use qualitative methods. The book is widely known outside of economics and has more than 11,000 Google Scholar citations. The authors argue that there is a unified logic of inference that holds across both qualitative and quantitative evidence (see also Gerring 2017: 165). The general principles of good research design are the same. Description involves making an inference from the facts that we observe to other, unobserved facts. This means that the sampling process is crucial, and that it should be representative of the broader population of interest. It also requires distinguishing between observations that are systematic and observations that are not systematic. We should therefore care about the sample size.

There are also some basic guidelines for thinking about the informational quality of different types of qualitative evidence. Gerring (2017: 172) suggests five criteria: relevance, proximity, authenticity, validity, and diversity. Relevant evidence speaks to the theoretical question of interest. It is pertinent. Proximity means that the evidence is close to, or in a special place to inform, the research question. For a study on urban policing, for instance, interviewing police officers and city dwellers would offer more proximate evidence than interviewing a farmer working in an agricultural area. Authentic evidence is not fabricated or untrue for some reason. Falsified police reports do not make for accurate description. It is preferable to have valid, unbiased sources of evidence. For instance, homicide data tends to be the

least biased crime data because there is less under-reporting and medical examiners have few incentives to misreport. Likewise, choosing which police officers to interview (say, a representative sample *versus* only rookies) will influence whether our descriptive inferences are biased. If one believes that there is bias in the sampling process, it can often be accounted for. For instance, we know that sex offenses tend to be under-reported, so inferences based on interviews should account for that bias as much as possible. Finally, evidence from a diversity of perspectives, people, and methods can allow triangulation if they tell a consistent story. If a victim, police officer, and a witness independently report the same facts, then it should give us greater confidence than if we have only one report. These five criteria provide a guide for assessing the usefulness of pieces of qualitative evidence.

Once we collect qualitative evidence and assess its usefulness, we must decide whether the sampling process allows for an accurate descriptive inference. King *et al.* (1994: 63–69) give two criteria for judging descriptive inference (both of which will be familiar to economists): bias and efficiency. Unbiasedness means that if we apply a method of inference repeatedly, then we will get the right answer on average. If one interviews a biased sample of respondents, then the sample will do a poor job of describing the population. For example, if we only interviewed police officers about the quality of policing in poor and minority communities, then our evidence and descriptions would likely be biased. Likewise, if we did ride-alongs to observe policing only in affluent neighborhoods, we could not draw accurate descriptive inferences about policing behavior in poor communities. A better research design, in this case a more representative sample, would improve our inferences by reducing bias.

Second, we can judge descriptive inferences based on efficiency, the variance of an estimator across hypothetical replications (King *et al.*, 1994: 66). *Ceteris paribus*, the less variance, the better. But, we might be willing to accept some bias if efficiency is greatly improved. This becomes important in considering whether to do an in-depth case study (small-*c*, large-*n*) or to collect only one or a few observations for many cases (large-*c*, small-*n*). King *et al.* (1994: 67) write, ‘although we should always prefer studies with more observations (given the resources necessary to collect them), there are situations where a single case study (as always, containing many observations) is better than a study based on more observations, each one of which is not as detailed or certain’. For example, we might be interested in knowing what crime is like in US cities. There are different ways to make inferences about this. We could collect data on the rates of property and violent crime from the Uniform Crime Reporting Statistics for the hundred largest US cities. Or, we could select a city that is typical or representative in some way (perhaps based on past research or based on a look at the Uniform Crime Data) and collect many additional observations within that single case. This could include interviewing a large numbers of people (police officers, offenders, victims, residents, and shop keepers), spending several weeks on police ride alongs, living in different parts of the city, going to crime hot spots at night, following the local news, and attending city council meetings when crime is discussed. This would generate far more observations about a single case and could give a more precise description of urban crime in American than aggregate, national crime rates.

Causal inference

How can institutional economists working with qualitative evidence make causal inferences? One of the most intuitive approaches is the use of comparative case studies based on a ‘most similar’ design. This is, in a sense, attempting to mimic identification of causal effects with quantitative data. The goal is to select cases where *x* varies, and as much as possible, holds constant other factors *z*, a vector of background factors that may affect *x* and *y* and thus serve as confounders (Gerring, 2017: xxviii). If *z* is identical across cases (and we can rule out reverse causality), then we have good reason to think that a change in *x* is what caused an observed change in *y*. This makes it crucial to find cases where *z* is similar. A deep and rich knowledge of a case helps a researcher to learn how similar *z* is across cases. If *z* does vary, the small number of cases and intensive knowledge of the case can often allow a researcher to determine to what extent these differences in *z* might actually explain variation in *y*.

Comparative case studies based on a most-similar design take two common forms. First, ‘longitudinal’ case studies focus on one (or possibly a few) cases through time. This research design requires that there is within-case variation in x that would allow us to observe changes in y . Focusing on a single case provides two advantages. It allows a researcher to accumulate a much richer and fine-grained knowledge of the case than would be possible if studying a large number of cases. From this perspective, it is desirable for a researcher to have expert knowledge on the topic being studied. In a longitudinal case study, z is also often more similar over time than it is when comparing across cases. For example, in a study on urban policing, z might be more similar in St. Louis over time than z would be in St. Louis compared to Chicago. Second, a ‘most-similar’ research design can entail comparison across two or more cases that are as close as possible on z but have variation in x . For instance, one could study crime in St. Louis, MO and Kansas City, MO. They both are medium-sized, Midwestern cities in the same state with similar local economies. By contrast, comparing cities that are very different – such as a large urban city like Chicago, IL and a small town like Hope, MI – would undermine this type of analysis because z varies a great deal across cases in ways that likely affect the relationship between x and y .

In an article in the *Journal of Political Economy*, Peter Leeson provides an excellent example of how to make causal claims based on comparative case studies (Leeson, 2007; see also, Leeson, 2009a, 2009b, 2010). He compares the internal governance in 18th-century pirate ships to the organizational form found in merchant ships operating at the same time. These are useful comparison cases because they are both seafaring operations, worked by the same types of sailors, traveling in the same waters, at the same historical era, and have access to similar technologies. Leeson uses a range of qualitative evidence to inform his comparisons, including written laws, information about *de facto* legal enforcement, private written correspondence between merchants, accounts of pirate hostages, qualitative histories of piracy, accounts from former pirates, and pirates’ own written constitutions (Leeson, 2007: 1052–1053). What Leeson discovers is that merchant ships were incredibly hierarchical and ship captains ruled in an autocratic way, often physically abusing their sailors. By contrast, pirate ships were highly-egalitarian constitutional democracies and pirate captains rarely preyed on their sailors.

What explains variation in ship organization? Leeson provides a compelling argument that these differences arise from differences in how they are financed. Merchant ships are financed by land-based businessmen who were absent from daily operations. This absence created a serious threat of shirking by merchant sailors and captains. To overcome the principal-agent problem, merchant captains were given a stake in the profitability of the enterprise and great legal authority to use corporal punishment to control the wage-based sailors on board. Pirate ships, on the other hand, did not face the problem of an absentee owner. They stole their ships, so they were all co-owners and residual claimants. Since they were also working onboard, each could monitor each other’s shirking. To prevent free riding, they wrote detailed constitutions to empower pirate captains to lead effectively, but to constrain their power so that it could not be abused. This included special rewards for pirates who gave the most effort and the right to depose an abusive captain and to elect his replacement. Leeson’s paper is a useful illustration of how to use economic theory to make causal claims with qualitative evidence because it provides persuasive evidence that the type of ownership (x) varies across similar cases (z) to explain variation in internal organization (y).

Criteria for judging comparative case studies

Once we think of using qualitative evidence in this comparative way, it suggests several criteria for assessing the plausibility of the causal claims that are made. Here, I present four ways of judging comparative case studies. First, one can question how close the cases match on z . The closer the match, the more reliable the inference. The more dissimilar z is across cases, the more concern we should have about confounders. When z varies across cases, the author should provide evidence on whether, and to what extent, these differences undermine causal claims. For instance, Leeson’s hypotheses could not be tested effectively by comparing 18th century piracy to recent Somalian piracy – which is organized and operated in vastly different ways (Leeson, 2007: 1088–1089).

In both descriptive and causal inference, careful attention must be paid to how the sample relates to a well-specified population of interest. All case studies should explicitly justify why particular cases are selected. Some cases are chosen because they are representative of some broader population. Other cases are chosen because they fit variation in the explanatory variables of a hypothesis we wish to test. Sometimes a case is chosen because it is a 'least-likely' case, which is helpful for testing the robustness of some finding or relationship. For instance, Leeson's study of pirates is a least-likely case to find the emergence of democracy and cooperation. Nevertheless, he finds that self-governance is possible even among history's most notorious rogues, suggesting that cooperation can be robust to agent type.

Second, we can judge the plausibility of causal claims by noting where the variation in x comes from.

Variation arises in several ways. In the most useful cases, x varies for random or exogenous reasons. For example, one challenge in studying the police-crime hypothesis is that more crime often causes more policing. Cities might hire more police in the face of rising crime rates. To get around the problem of reverse causality, the two papers discussed above use natural experiments: an exogenous increase caused by changes in the 'terror alert level' (Klick and Tabarrok, 2005) and by a terror attack itself (Draca *et al.*, 2011), both of which increased the number of police on patrol but are likely to be uncorrelated with the rate of crime. Another helpful type of variation is when x varies both within-cases and across cases. This increases the number of opportunities to observe correlations that can inform causal claims. A slightly less helpful type of variation is when there is only within-case variation in x . For example, in a longitudinal case study of the California prison system, I argue that growth in the size of the prison population caused the growth in prison gang activity (Skarbek, 2014). The change in x was not the result of an exogenous shock, and this raises concerns about endogeneity. In response to this concern, I argued that there is no evidence to support the claim that an anticipated rise of prison gangs in the future somehow caused the earlier growth in the prison population (Skarbek, 2014: 47–72). Finally, when there is no within-case variation in x , one can simply compare across cases. However, with cross-case comparisons, a major concern is that some aspect of z differs across cases and it is the differences in z that actually explain variation in y across cases. For instance, prisons in Norway do not have prison gangs like those found in California. It might be that this is because they have a small prison population, but it could also be that officials provide better governance, more resources, or there are systematic differences in the corrections culture (Skarbek, 2016). In sum, the nature of the variation itself is an important criterion for judging the quality of causal inferences based on comparative cases.

Third, we can be more confident in causal claims the more that one can increase the number of observable implications of a theory (King *et al.*, 1994; Levy, 2008: 208–228). This increases our ability to subject a claim to both qualitative and quantitative evidence. This means that developing better theory is an important way to improve the plausibility of claims based on qualitative evidence. Novel or surprising empirical implications are especially informative. Expanding the number of empirical implications for alternative explanations likewise provides more opportunities to assess the value of one's own proposition *versus* competing explanations. Likewise, identifying empirical implications that can distinguish between theories is extremely useful. Compared to someone working with a large number of cases, deep knowledge of a particular case can often allow a researcher to observe more, and possibly more subtle, facts predicted by a theory. The more that an author identifies the most relevant alternative explanations and provides evidence on their empirical implications, the more confident we can be in the argument.

Finally, we can assess causal claims by focusing on what role, if any, potential confounders play. This can actually be easier with case studies because they are, by definition, based on an intensive knowledge of only one or a few cases. Moreover, the thick knowledge found in qualitative evidence can be highly informative. For instance, spending several 12-hour police shifts in a car with an officer can reveal much about their personal disposition, tact in interacting with the public, and how they perceive their job. This information cannot be easily entered into a data file, but it provides insights about the presence of potential confounders.

One possible limitation facing comparative case studies is the lack of a sufficiently similar case to compare against. However, recent advances at the intersection of qualitative and quantitative methods provide a potential solution to this problem. In particular, the synthetic control method provides a way to create a counterfactual case for comparison. It does so based on ‘the weighted average of all potential comparison units that best resembles the characteristics of the case of interest’ (Abadie *et al.*, 2015: 496). Synthetic control is especially helpful when comparing across cases that focus on aggregate or large-scale cases, such as regional economies, state-level crime rates, or national political institutions (e.g. Freire, 2018). Because the treatment takes place on such a large scale in these types of cases, there might simply be no untreated units to compare them to (Abadie *et al.*, 2010: 493).⁶ Creating a synthetic counterfactual helps to solve this problem. In fact, even when there are untreated units available to study, they might still differ systematically in ways that undermine our ability to draw causal inferences. Synthetic control can thus sometimes still provide a better comparison than any available untreated units (Abadie *et al.*, 2010: 494). This method also provides the additional advantage of providing transparency to other researchers about how the synthetic case is weighted and created in the first place (Abadie *et al.*, 2015: 494). Moreover, it can reduce the uncertainty involved in selecting which case is the best one to compare with. This is, more broadly, part of a growing emphasis on using quantitative methods to organize qualitative research, such as the use of algorithms for selecting which cases to compare (Gerring, 2012: 118–128). Related to this, the synthetic control method provides a nice example of how to wed quantitative and qualitative research methods. The synthetic case provides an estimate of a causal effect and qualitative evidence can help identify mechanisms.

3. Process tracing

Basic approach

Process tracing is another method that social scientists use to understand causal relationships. It is substantially different from the most-similar method discussed in the previous section. Unlike comparative case studies, this case study method does not require a good counterfactual case. It is not trying to mimic causal inference in the way that it is used in a potential outcomes framework (Mahoney, 2010: 124). It emphasizes collecting evidence to trace the unfolding of some event or process one step at a time. The focus can be either on some recurring process within a case or on the unfolding of a process in a single event. This makes process tracing an ideal method for understanding questions having to do with institutional change and for understanding what factors sustain institutional outcomes.

There is no standard definition of process tracing, and one recent article actually counted 18 different definitions (Trampusch and Palier, 2016: 440–441). So with that caveat, here are a few of the more common descriptions of process tracing.⁷ Collier (2011: 824) describes process tracing as ‘an analytic tool for drawing descriptive and causal inferences from diagnostic pieces of evidence – often understood as part of a temporal sequence of events or phenomena’. Bennett and Checkel (2015: 6) describe process tracing as examining ‘intermediate steps in a process to make inferences about hypotheses on how that process took place and whether and how it generated the outcome of interest’. This method is frequently used by people interested in understanding causal and temporal mechanisms so it is likely to be useful to people studying institutions and institutional change (Beach, 2016: 438; Falletti, 2016; Trampusch and Palier, 2016). For example, Lorentzen *et al.* (2017: 472) explains that, ‘rather than trying to assess whether or not changes in X produce changes in Y, the goal is to evaluate a particular mechanism linking the two’. This argument suggests that we can view process tracing as a way to use quantitative analysis and qualitative evidence as complements instead of substitutes.

⁶This same challenge is also a reason for examining variation in units at the subnational level (Giraudy *et al.*, 2019).

⁷For a longer, thoughtful debate on how to conceptualize process tracing, see the ‘Causal Mechanisms and Process Tracing’ symposium in the *Qualitative & Multi-Method* (Spring/Fall 2016) newsletter and the symposium on ‘Process Tracing: the understandings of causal mechanisms’ in *New Political Economy*, introduced by Trampusch and Palier (2016).

There are three main components to process tracing. First, a researcher collects evidence that is relevant to the case. This evidence is often referred to as ‘causal process observations’ (in contrast to data matrices used in quantitative analysis that are often referred to in this literature as ‘data-set observations’). A causal process observation (CPO) is ‘an insight or piece of data that provides information about context, process, or mechanism, and that contributes distinctive leverage in causal inference’ (Mahoney, 2010: 124). Unlike in most-similar designs, these are not simply called ‘observations’ because each CPO might be a totally different type of evidence. As a result, there might not be a single unit of analysis, and it will instead look more like a series of one-shot observations (Gerring, 2017: 159). The evidence for CPOs can come from both qualitative and quantitative sources. Second, the CPOs are used to provide a careful description of the process or mechanism, with an emphasis on identifying and closely tracing change and causation within the case (Collier, 2011: 823). Finally, process tracing places a major emphasis on understanding the sequences of independent, dependent, and intervening variables (Bateman and Teele, 2019; Collier, 2011). The goal is to ‘observe causal process through close-up qualitative analysis within single cases, rather than to statistically estimate their effects across multiple cases’ (Trampusch and Palier, 2016: 439). This requires accurately describing each of these at points in time, which can eventually accurately describe the process over time.

Gerring (2017: 158) illustrates how process tracing works by dissecting one of the most famous studies in political science: Theda Skocpol’s *States and social revolutions: A comparative analysis of France, Russia and China* (Skocpol, 1979). In the book, Skocpol uses detailed historical evidence to argue that social revolution occurs because of agrarian backwardness, international pressure, and state autonomy. This argument informs major debates on institutions, including questions about state capacity, political upheaval, social movements, and state-based governance. To show how Skocpol uses her evidence as an example of process tracing, Gerring (2017) identifies 37 claims based on Skocpol’s qualitative evidence that substantiate how these processes unfolded and led to revolution (Figure 1). The arrows in the figure indicate causal linkages between particular CPOs. For example, as interpreted by Gerring (2017: 158), Skocpol argues that agricultural backwardness (CPO 4) was caused by the combination of three factors: property rights relations that prevented agricultural innovations (CPO 1), a tax system that discouraged agricultural innovation (CPO 2), and sustained growth that discouraged agricultural innovation (CPO 3). As a result, there was a weak domestic market for industrial goods (CPO 5), which prevented France from having an industrial breakthrough (CPO 8). The lack of an industrial breakthrough (CPO 8) meant that it could not compete with England (CPO 10). In addition to lack of innovation, France failed to have sustained economic growth (CPO 9) because of internal transport problems (CPO 6) and population growth (CPO 7). The failure to have sustained growth (CPO 9) and the inability to compete successfully with England (CPO 10) caused the state to have major financial issues (CPO 25). Combined with other factors, this helped cause the outbreak of social revolution in France. While I will not discuss the remainder of the CPOs in the figure, the full graphic illustrates how a broader argument can be marshaled with this method. Historical evidence on a substantial number of empirical claims can support larger, historical causal arguments.

This example of process tracing is clearly not an attempt to identify average treatment effects, or even to discern causality based on a most-similar cases research design. Instead, process tracing uses deep knowledge of a case to trace causality from changes in independent variables, through intervening variables, leading to changes in the dependent variable. Social scientists working in this tradition argue that the rich qualitative evidence can make it easier to observe causal relationships in the evidence, compared to thin sources of quantitative data that provide far less detailed information.

To return to a crime example, I used process tracing to explain variation in jail-based extortion based on the case of the Mexican Mafia prison gang in the Los Angeles County jail system (Skarbek, 2011). The gang has a credible threat to harm Hispanic prisoners in the jail, but not prisoners from other ethnic backgrounds. With this threat, the gang is able to extort a portion of profits from drug sales (called a ‘gang tax’) from non-incarcerated street gang members. The study draws on

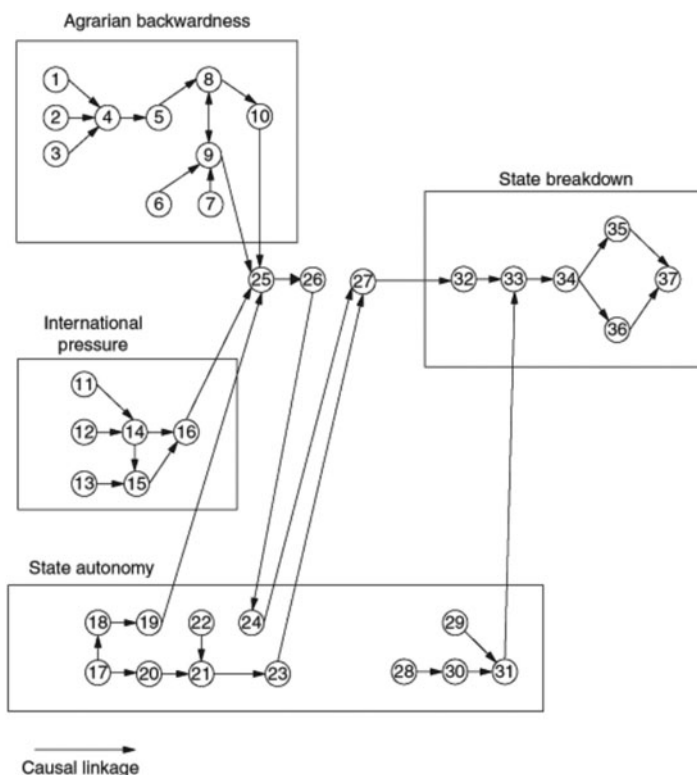


Figure 1. Example of process tracing from Gerrig (2017).

qualitative evidence, including law enforcement documents, interviews with prisoners and prison staff, and criminal indictments.

Figure 2 provides a simple set of mechanisms that predict when someone will pay the gang's extortion demand. The theory predicts that a street-based drug dealer will pay the gang tax in two scenarios. First, he will do so if he anticipates being incarcerated in the local jail and that members of the Mexican Mafia can and will harm him if he has not paid (strong gang). Second, he will pay even if he does not anticipate incarceration, if he has friends who are incarcerated who the gang can harm. The article provides qualitative evidence for each step in this process, showing that when these conditions are met, drug dealers do, in fact, pay the gang tax. This theory also suggests empirical implications where jail-based extortion should not occur. In particular, drug dealers who do not anticipate incarceration and who do not have friends incarcerated will not pay. If a drug dealer can rely on a rival prison gang to stay safe from the Mexican Mafia, then he will not pay. If someone is engaged in a legal business and does not anticipate incarceration (and furthermore can rely on police for protection), he or she will not be targeted. The qualitative evidence supports these predictions as well. Taking these propositions together, the evidence describes how the sequence of extortion (or not) varies across a wide range of potential victims. The qualitative evidence shows that these intervening variables are important and are consistent with the causal claim about what explains variation in the dependent variable (paying gang taxes).

Assessing process tracing

There are two common ways for assessing how much confidence we should have in claims based on process tracing. The first is similar to the above criteria for assessing comparative cases (see the section 'Criteria for judging comparative case studies'). The evidence should provide substantial, useful

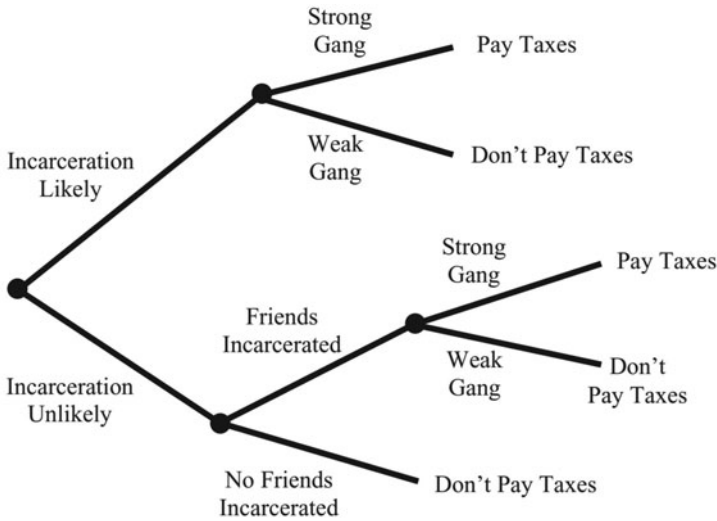


Figure 2. Example of process tracing from Skarbek (2011).

information, the sampling procedure should prevent bias, and the evidence should adjudicate competing explanations. In addition, we should find an explanation more plausible if it can effectively explain out-of-sample cases.

Second, recent methodology research proposes several tests for assessing how informative individual pieces of evidence are in testing a causal claim. The tests are based on whether passing a test is a necessary and/or sufficient condition for accepting the causal claim (Bennett, 2010; Collier, 2011; Gerring, 2012; Mahoney, 2010). A *hoop* test is necessary, but not sufficient, for affirming a causal inference. At a minimum, the evidence has to pass this test, but even if it does, it does not rule out other plausible explanations. However, if the evidence fails the hoop test, then the causal claim is unlikely to be true. In the example of jail-based extortion by a prison gang, there has to be a strong gang in the jail for this hypothesis to be true. If the gang did not have a credible threat of violence, then it cannot be the case that they use it to extort drug taxes. Such a finding would essentially refute the claim. However, it might be that a strong prison gang receives payments for some entirely different reason, such as organizing the trafficking of drugs into the country. So, finding that a strong gang exists in the jail does not increase our confidence that much.

A *smoking-gun* test examines evidence that is not necessary for a causal claim to be true, but is sufficient for affirming the causal claim. Contrary to the hoop test, failing a smoking-gun test does not eliminate it, but passing it provides strong evidence that the causal claim is true. Passing a smoking-gun test also substantially weakens rival explanations. For example, if we had secretly-recorded phone conversations where prison gang members demanded that drug dealers pay taxes or else be subject to violence (and perhaps evidence of them killing people who refused to pay), then that evidence would pass the smoking-gun test (Skarbek, 2011: 708). It is sufficient to show that the proposed mechanism is in operation. However, we could conceivably find other evidence to support this claim, so this specific piece of evidence is not necessary.

Finally, a *doubly decisive* test is one that is both necessary and sufficient for affirming a causal claim. Evidence that passes this test provides strong evidence for the causal claim and eliminates rival explanations. Consider a plausible alternative explanation: the gang is simply providing a form of insurance that drug dealers voluntarily buy to stay safe if incarcerated in the future. A key difference in these mechanisms is whether people are forced to buy the insurance or not. Evidence that drug dealers are forced to pay and harmed if they try to avoid doing so, would pass the doubly decisive test. Evidence showing prison gangs use violence is necessary for the extortion mechanism to be true. This evidence is perhaps not as compelling as phone-recorded conversations, but it might well be

sufficient to support the proposed mechanism. In addition, evidence of violence provides strong evidence against the alternative explanation that it is merely a voluntary insurance scheme.⁸

To be sure, these ‘tests’ are far more subjective than standard statistical tests. First, the tests lack the precision and decisiveness of formal tests. This makes it less clear whether some particular piece of evidence passes the test or not. Second, a theory can be ambiguous in terms of which test is appropriate to apply to a particular piece of evidence. This raises the concern that people will disagree over which test is appropriate for a particular piece of evidence. Nevertheless, these tests provide a way to connect a theory to its empirical implications and to weight (ideally prior to observation) how important particular CPOs are for substantiating causal claims.⁹

4. Conclusion

Even if one concedes that the best way to identify causal effects is with quantitative data and a good identification strategy, qualitative evidence is still capable of complementing formal and quantitative research in critical ways (Seawright, 2016; Trampusch and Palier, 2016). First, as discussed above, case studies and process tracing can help us discover causal mechanisms that causal effects often overlook or are unable to observe (Seawright, 2016: 73). Second, the use of qualitative evidence can sometimes allow us to test assumptions that are typically untested when using only a single method (Lorentzen *et al.*, 2017: 473; Seawright, 2016: 1). For example, regressions are based on assumptions about measurement, the absence of confounding variables, and post-treatment bias, which cannot usually be tested with regression-type methods (Seawright, 2016: 14, 73). Qualitative evidence can inform the soundness of these assumptions. For example, the deep knowledge of a case can allow a researcher to develop, evaluate, or amend measurements and observe potential confounders (Seawright, 2016: 48–49, 73).

Process tracing and qualitative evidence can also complement the use of formal models. They both often specify key actors, choices, and how their interactions affect outcomes.¹⁰ Both formal models and process tracing often focus on understanding causal mechanisms (Lorentzen *et al.*, 2017). Lorentzen *et al.* (2017: 476) write ‘Even if the covariational pattern between a parameter and outcome fits the model’s predictions, strong empirical support for the model requires evidence of the intervening causal process implied by the model’. Detailed process tracing can provide evidence to answer this important question (Lorentzen *et al.*, 2017: 474). This is crucial for determining whether a specific model applies in a particular situation. Of course, all models are simplifications, so some assumptions will not accurately reflect reality. Rodrik (2015) argues that the focus should be on whether *critical* assumptions are true. If not, then the model needs revision. If assumptions that are not critical to the analysis are false, then this is not problematic. Lorentzen *et al.* (2017: 475) likewise argue that assumptions are only problematic if they do not fit the empirical evidence and if changing them substantially changes the implications of the model. Qualitative evidence and process tracing provide a way to test whether a model’s assumptions are sufficiently accurate to help us understand the world.

In conclusion, I have argued that there is value in incorporating more qualitative evidence into the study of institutions and institutional change. Qualitative evidence can help us understanding big historical questions, identify causal mechanisms, and better test thick theories in the social sciences. Social scientists working with qualitative evidence care about many of the same fundamental research design challenges that economists doing quantitative work care about. In some cases, qualitative evidence might not provide much helpful evidence. However, in many instances, qualitative evidence can

⁸A fourth, weaker test is the *straw-in-the-wind* test, which is evidence that is neither necessary nor sufficient for affirming a causal inference. If the evidence is consistent with a hypothesis, then it affirms or slightly raises its plausibility. If it is not consistent with the theory, then it slightly reduces the likelihood of the causal claim. Overall, it provides relatively little information.

⁹Trampusch and Palier (2016: 449–450) recommend 10 steps to help improve process tracing in practice.

¹⁰Economists and political scientists have often studied institutions with ‘analytical narratives’ that combine game theory and rich qualitative case studies (Bates *et al.*, 1998; Gailmard, 2019; Greif, 2006).

provide substantial explanatory value. Not all economists or all questions need to incorporate qualitative evidence, but there is a significant range of questions for which it would generate significant benefits, especially in the study of institutions and institutional change.

Acknowledgements. I am grateful to three anonymous referees, Ennio Piano, Emily Skarbek, and participants at a conference hosted by the Ronald Coase Institute for helpful comments and discussion. I thank the Center for the Study of Law & Society at University of California, Berkeley for hosting me as a Visiting Scholar while I wrote portions of this article.

References

- Abadie, A., A. Diamond and J. Hainmueller (2010), 'Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program', *Journal of the American statistical Association*, **105**(490): 493–505.
- Abadie, A., A. Diamond and J. Hainmueller (2015), 'Comparative Politics and the Synthetic Control Method', *American Journal of Political Science*, **59**(2): 495–510.
- Abbott, A. D. (2004), *Methods of Discovery: Heuristics for the Social Sciences*, New York, NY: W.W. Norton & Company.
- Bateman, D. A. and D. L. Teele (2019), 'A Developmental Approach to Historical Causal Inference', *Public Choice*. doi: 10.1007/s11127-019-00713-4.
- Bates, R. H., A. Greif, M. Levi, J.-L. Rosenthal and B. R. Weingast (1998), *Analytic Narratives*, Princeton, NJ: Princeton University Press.
- Beach, D. (2016), 'It's All About Mechanisms – What Process-Tracing Case Studies Should be Tracing', *New Political Economy*, **21**(5): 463–472.
- Bennett, A. (2010), 'Process Tracing and Causal Inference', in H.E. Brady and D. Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. 2nd edn. Plymouth: Rowman & Littlefield Publishers, pp. 207–219.
- Bennett, A. and J. T. Checkel (2015), 'Process Tracing: From Philosophical Roots to Best Practices', in A. Bennett and J. T. Checkel (eds), *Process Tracing: From Metaphor to Analytic Tool*, New York, NY: Cambridge University Press, pp. 3–37.
- Boettke, P. J., C. J. Coyne and P. T. Leeson (2013), 'Comparative Historical Political economy', *Journal of Institutional Economics*, **9**(3): 285–301.
- Boettke, P. J., J. S. Lemke and L. Palagashvili (2016), 'Re-evaluating Community Policing in a Polycentric System', *Journal of Institutional Economics*, **12**(2): 305–325.
- Card, D. (1990), 'The Impact of the Mariel Boatlift on the Miami Labor Market', *ILR Review*, **43**(2): 245–257.
- Collier, D. (2011), 'Understanding Process Tracing', *PS: Political Science & Politics*, **44**(4): 823–830.
- Coppedge, M. (1999), 'Thickening Thin Concepts and Theories: Combining Large *n* and Small in Comparative Politics', *Comparative Politics*, **31**(4): 465–476.
- Deaton, A. and N. Cartwright (2018), 'Understanding and Misunderstanding Randomized Controlled Trials', *Social Science & Medicine*, **210**: 2–21.
- Draca, M., S. Machin and R. Witt (2011), 'Panic on the Streets of London: Police, Crime, and the July 2005 Terror Attacks', *American Economic Review*, **101**(5): 2157–2181.
- Falletti, T. G. (2016), 'Process Tracing of Extensive and Intensive Processes', *New Political Economy*, **21**(5): 455–462.
- Freire, D. (2018), 'Evaluating the Effect of Homicide Prevention Strategies in Sao Paulo, Brazil: A Synthetic Control Approach', *Latin American Research Review*, **53**(2): 231–249.
- Gailmard, S. (2019), 'Game Theory and the Study of American Political Development', *Public Choice*. doi:10.1007/s11127-019-00705-4.
- Gerring, J. (2012), *Social Science Methodology: A Unified Framework*, Cambridge, UK: Cambridge University Press.
- Gerring, J. (2017), *Case Study Research: Principles and Practices*, Cambridge, UK: Cambridge University Press.
- Giraudy, A., E. Moncada and R. Snyder (2019), *Inside Countries: Subnational Research in Comparative Politics*, Cambridge, UK: Cambridge University Press.
- Greif, A. (2006), *Institutions and the Path to the Modern Economy: Lessons From Medieval Trade*, New York, NY: Cambridge University Press.
- Greif, A. and C. Kingston (2011), 'Institutions: Rules or Equilibria?', in N. Schofield and G. Caballero (eds), *Political Economy of Institutions, Democracy and Voting*. New York, NY: Springer, pp. 13–43.
- King, G., R. O. Keohane and S. Verba (1994), *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, NJ: Princeton University Press.
- Kirzner, I. M. (2009), *The Economic Point of View*, Indianapolis, IN: Liberty Fund, Inc.
- Klick, J. and A. Tabarrok (2005), 'Using Terror Alert Levels to Estimate the Effect of Police on Crime', *The Journal of Law and Economics*, **48**(1): 267–279.
- Kuran, T. (2012), *The Long Divergence: How Islamic law Held Back the Middle East*, Princeton, NJ: Princeton University Press.
- Leeson, P. T. (2007), 'An-arrrgh-chy: The Law and Economics of Pirate Organization', *Journal of political economy*, **115**(6): 1049–1094.

- Leeson, P. T. (2009a), 'The Calculus of Piratical Consent: The Myth of the Myth of Social Contract', *Public Choice*, **139**(3–4): 443–459.
- Leeson, P. T. (2009b), *The Invisible Hook: The Hidden Economics of Pirates*, Princeton, NJ: Princeton University Press.
- Leeson, P. T. (2010), 'Pirational Choice: The Economics of Infamous Pirate Practices', *Journal of Economic Behavior & Organization*, **76**(3): 497–510.
- Levy, J. S. (2008), 'Counterfactuals and Case Studies', in Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier (eds), *The Oxford Handbook of Political Methodology*, Oxford: Oxford University Press, pp. 627–644.
- Lorentzen, P., M. T. Fravel and J. Paine (2017), 'Qualitative Investigation of Theoretical Models: The Value of Process Tracing', *Journal of Theoretical Politics*, **29**(3): 467–491.
- Lyall, J., G. Blair and K. Imai (2013), 'Explaining Support for Combatants During Wartime: A Survey Experiment in Afghanistan', *American Political Science Review*, **107**(4): 679–705.
- Mahoney, J. (2010), 'After KKV: The New Methodology of Qualitative Research', *World Politics*, **62**(1): 120–147.
- McCloskey, D. N. (2016), *Bourgeois Equality: How Ideas, not Capital or Institutions, Enriched the World*, Chicago, IL: University of Chicago Press.
- Moskos, P. (2008), *Cop in the Hood: My Year Policing Baltimore's Eastern District*, Princeton, NJ: Princeton University Press.
- Munger, M. C. (2010), 'Endless Forms Most Beautiful and Most Wonderful: Elinor Ostrom and the Diversity of Institutions', *Public Choice*, **143**(3): 263–268.
- Ostrom, E. (1990), *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge, MA: Cambridge University Press.
- Ostrom, E. (2005), *Understanding Institutional Diversity*, Princeton, NJ: Princeton University Press.
- Ostrom, E. and D. C. Smith (1976), 'On the Fate of 'Lilliputs' in Metropolitan Policing', *Public Administration Review*, **36**(2): 192–200.
- Ostrom, E., R. B. Parks and G. P. Whitaker (1978), *Patterns of Metropolitan Policing*, Cambridge, MA: Ballinger Publishing Company.
- Poteete, A. R., M. A. Janssen and E. Ostrom (2010), *Working Together: Collective Action, the Commons, and Multiple Methods in Practice*, Princeton, NJ: Princeton University Press.
- Quine, W. and J. S. U. Ullian (1978), *The Web of Belief*, New York, NY: McGraw-Hill.
- Rodrik, D. (2015), *Economics Rules: The Rights and Wrongs of the Dismal Science*, Oxford, UK: WW Norton & Company.
- Ruhm, C. J. (2019), 'Shackling the Identification Police?', *Southern Economic Journal*, **85**(4): 1016–1026.
- Seawright, J. (2016), *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*, Cambridge, UK: Cambridge University Press.
- Shirley, M. M. (2013), 'Measuring Institutions: How to be Precise Though Vague', *Journal of Institutional Economics*, **9**(1): 31–33.
- Skarbek, D. (2011), 'Governance and Prison Gangs', *American Political Science Review*, **105**(4): 702–716.
- Skarbek, D. (2014), *The Social Order of the Underworld: How Prison Gangs Govern the American Penal System*, New York, NY: Oxford University Press.
- Skarbek, D. (2016), 'Covenants Without the Sword? Comparing Prison Self-Governance Globally', *American Political Science Review*, **110**(4): 845–862.
- Skocpol, T. (1979), *States and Social Revolutions: A Comparative Analysis of France, Russia and China*, Cambridge, MA: Cambridge University Press.
- Tilly, C. (1984), *Big Structures, Large Processes, Huge Comparisons*, New York, NY: Russell Sage Foundation.
- Trampusch, C. and B. Palier (2016), 'Between x and y : How Process Tracing Contributes to Opening the Black Box of Causality', *New Political Economy*, **21**(5): 437–454.
- Voigt, S. (2013), 'How (not) to Measure Institutions', *Journal of Institutional Economics*, **9**(1): 1–26.