Qualitative investigation of theoretical models: the value of process tracing

Peter Lorentzen
University of California, Berkeley, USA

M Taylor Fravel
Massachusetts Institute of Technology, USA

Jack Paine
University of Rochester, USA

Abstract
Political scientists frequently use qualitative evidence to support or evaluate the empirical applicability of formal models. Despite this widespread practice, neither the qualitative methods literature nor research on empirically evaluating formal models systematically address the topic. This article makes three contributions to bridge this gap. First, it demonstrates that formal models and qualitative evidence are indeed frequently combined in current research. Second, it shows how process tracing can be as important a tool for empirically assessing models as statistical testing, because models and process tracing share a common focus on understanding causal mechanisms. Last, it provides new guidelines for using process tracing that focus on issues specific to the modeling enterprise, illustrated with examples from recent research.

Keywords
Comparative politics; empirical implications of theoretical models; game theory; international relations; process tracing

Corresponding author:
Peter Lorentzen, University of California, Berkeley, 210 Barrows Hall, Berkeley, CA 94720-1950, USA.
Email: lorentzen@berkeley.edu

1. Introduction

Since the introduction of formal models into political science, their relationship with the rest of the discipline has provoked vigorous discussion. Early debates about whether formalization or rational choice approaches contribute to our understanding of politics have largely given way to more applied discussions about the appropriate relationship between formal theories and empirical research.[1] Yet, these discussions have centered on how (or even whether) formal models should be tested with statistical methods, providing little guidance for scholars interested in assessing formal theories with qualitative evidence. Likewise, although the literature on qualitative methods has seen a strong resurgence of activity in recent years, it does not address the particular issues that arise when seeking to use qualitative methods to evaluate formal theories.

This methodological gap is perplexing, because in practice articles introducing a new formal model frequently include qualitative evidence. We surveyed all such articles published from 2006 to 2013 on topics in international relations and comparative politics from a set of prominent journals and found that qualitative evidence appeared in over half the articles that empirically assessed the model. Furthermore, recent critiques of prominent models of audience costs and of regime change have sparked vigorous debates about core arguments in these subfields.[2]

Yet despite its common use, formal theorists often dismiss the value of qualitative evidence. Coming from the other direction, scholars using qualitative evidence to evaluate existing formal models do not express these qualms, but modelers often question the conclusions drawn by these studies. The lack of agreed-upon standards can lead scholars from different methodological backgrounds to talk past each other, hindering scientific progress. The applied formal modeling enterprise is now at an unsatisfying middle ground whereby qualitative evidence is routinely used to convince readers of the value of a model, yet the evidence is not presented or evaluated rigorously because it is qualitative.

The frequency with which scholars from diverse methodological backgrounds use qualitative evidence to evaluate formal models, along with a lack of clear methodological guidance, suggests two options: develop better standards or abandon the practice. We argue for developing better standards. In particular, formal models and the increasingly prominent qualitative method known as process tracing share an under-recognized affinity—a focus on causal mechanisms—that makes process tracing a valuable tool for the empirical investigation of formal models. Qualitative evidence that strongly supports or contradicts a model's causal processes can help to either strengthen or to weaken the model's claims to empirical applicability. Such process tracing can be as convincing as statistical tests.

A sharp need exists for systematic standards on using process tracing to evaluate formal models. In this article, we take a first step toward such standards, focusing on issues that are unique to or are particularly important for the modeling endeavor rather than surveying process tracing in general. We address (1) how to use process tracing to evaluate the causal mechanisms embodied in a formal model, (2) what types of evidence to collect to conduct process tracing, and (3) how to

select cases. Although we draw from existing scholarship on process tracing, formal models raise specific concerns including how to (or even whether) to test a model's assumptions and how to assess actors' beliefs. Furthermore, the guidelines provided improve upon the somewhat ad hoc approaches used at present for evaluating models. By including examples in which qualitative evidence is intended to validate a model as well as examples in which the evidence challenges a formal model, we hope to engage scholars from diverse backgrounds to help bridge the gap between two closely related yet surprisingly disconnected research traditions.

In order to discuss how to evaluate models, we must first specify what this means. Some scholars have implied that formal models have no value unless they can be and have been empirically tested (Hill, 2005). Others have argued that the notion of empirically testing theoretical models misunderstands both formal theory and empirical research (Clarke and Primo, 2012). We take a middle ground that we believe reflects the consensus view of most applied modelers, but that may require some explanation for other scholars. In particular, we do not believe it is possible to test whether a model is true or correct in the same way that one can test a statistical hypothesis. All models are simplifications and therefore are false in the strictest sense. However, political scientists are frequently interested in evaluating whether a particular model serves a useful guide to understanding a specific empirical setting.[3] For instance, are redistributive conflicts an important driver of regime change? Do audience costs play a key role in inter-state conflict? Or are these factors relatively unimportant?

Even many who accept this stance favor quantitative over qualitative evidence. However, qualitative and quantitative methods can both play an important role in evaluating formal models. Moreover, these two approaches complement each other at all stages of research. Qualitative evaluation is not simply prep work for statistical testing. Even if one believes that in principle any qualitative evaluation is inferior to a quantitative test with sufficient high-quality data (King and Powell, 2008), this is rarely the relevant issue in the real world. Instead, one generally chooses between imperfect methods, such as an experiment with uncertain external validity, an observational study with uncertain causal implications, or a qualitative study with only a few cases. Which approach proves more helpful for evaluating a model depends on what kind of data can be gathered or generated. For many topics in comparative politics and international relations, only a small number of cases may fit the scope conditions of the model. It is frequently difficult to measure key parameters and choice variables specified by a model or to assess actors' intentions. Moreover, the focus of process tracing on evaluating the causal process differs from the goal of typical statistical tests, which estimate the covariational relationship between a parameter and an outcome. Consequently, process tracing can evaluate key components of a model that conventional statistical tests cannot.

The next section of this article demonstrates the major role qualitative evidence already plays in validating and evaluating formal models, and discusses problems that arise from the lack of rigorous standards. Section 3 shows that formal modeling and process tracing share the common goal of understanding causal pathways, suggesting strong complementarities between the two endeavors. Three sections

provide a practical how-to guide: Section 4 discusses unique challenges that arise when considering what aspects of a model to evaluate and how to use process tracing effectively, Section 5 presents standards for gathering qualitative evidence, and Section 6 discusses case selection. Section 7 briefly explores what to do if the evidence goes against a model before we conclude.

2. Formal models and qualitative evidence: problems with current practice

The current use of qualitative evidence and formal models highlights two interrelated problems. First, although qualitative evidence often accompanies the introduction of new models, this is done with little methodological self-consciousness. Second, a gap exists in the methods literature. With few exceptions, the literatures on the empirical testing of formal models and on qualitative methods do not address how to evaluate formal models using qualitative evidence.

2.1. Contradictions in current practice

To understand current practice, we surveyed six prominent political science journals and identified all articles with a formal model published between 2006 and 2013.[4] Of the 182 such articles that addressed topics in international relations or comparative politics, 70 percent (128 articles) included either a quantitative or qualitative empirical component. Among these 128 articles, more than half (72) provided qualitative evidence. Articles that provided qualitative evidence dedicated an average of two-and-a-half pages to discussing this evidence. This is notable when one considers the tight space limitations of journal articles. Moreover, in book-length expositions of formal theory, qualitative evidence has become almost mandatory.[5]

This widespread use of qualitative evidence has two implications. First, since models are already frequently presented in conjunction with qualitative evidence, a need exists for more explicit standards against which to evaluate using such evidence. Second, the importance of qualitative evidence in the formal modeling enterprise suggests scholars with skills in qualitative research should be encouraged to systematically evaluate evidence offered for formal theories and to gather new evidence to evaluate these theories. Recent qualitative challenges to influential formal models, discussed below, represent an important step in this direction, despite the counter-criticisms they have attracted.

Unfortunately, scholars introducing a formal model often disparage their own qualitative evidence even after devoting considerable thought to this evidence. For example, in an article discussing political barriers to industrialization, Acemoglu and Robinson (2006b) devote four journal pages to country case studies that contain extensive evidence that these cases are consistent with their model. Yet they state their 'interpretations are necessarily speculative and more conclusive evidence requires proper statistical testing of the ideas we develop here' (125). Similarly, Slantchev (2011) opens a chapter-long analysis of the expansion of the Korean

War by saying, 'I remain skeptical about the extent to which we should trust either supporting or disconfirming evidence' (192).

Certainly, these models constitute important theoretical advances regardless of how well they explain particular cases. In addition, one might argue that these case studies are only intended as illustrations to make the models more accessible to non-technical readers. However, political scientists often link their models with historical events because they hope readers will find the empirical connection plausible and meaningful. Indeed, after Acemoglu and Robinson (2006b) describe their qualitative evidence as inferior to statistical testing, they go on to claim their case studies show that 'there is direct historical evidence that the mechanism which is the central focus of this paper was important to nineteenth-century industrialization' (125). Slantchev contends that his Korean War case study 'challenge[s] some existing interpretations and show[s] how the model can illuminate some of the complex dynamics [of the case]' (192). These claims certainly transcend mere illustration— even though Slantchev carefully argues that 'the main purposes of [his] theoretical model [are] explicative and generative.' If the case studies deliver what the authors suggest, it is unnecessary to shroud motivation for the qualitative evidence in heavy qualifications.

2.2. Shortcomings of existing standards

Little methodological guidance exists for combining qualitative evidence and formal models. Discussions of empirically testing formal models may mention case studies in passing, but typically characterize them as merely a source of ideas after which formalization and quantitative testing can begin (e.g., Granato and Scioli, 2004; Morton, 1999: 133–134). A recent book-length challenge to the very premise of testing models has 'little to say directly about qualitative research…because the conversation in political science in recent years has revolved around formal models and statistical models' (Clarke and Primo, 2012: 18). The neglect has been mutual. The terms 'game theory' and 'formal theory' appear nowhere in the widely used qualitative methods handbook *Rethinking Social Inquiry* (Brady and Collier, 2010), nor in two recent book-length treatments of case study methods (Beach and Pedersen, 2013; Rohlfing, 2012). The topic receives only passing mention in George and Bennett's (2005) influential guide to case study methodology (e.g., 34–35), Goertz and Mahoney's (2012) study of the 'two cultures' of quantitative and qualitative research (e.g., 106), and Bennett and Checkel's (2014) guide to process tracing.

A decade-and-a-half after its launch, the *Analytic Narratives* project remains the most prominent attempt to combine formal models and qualitative research (Bates et al., 1998).[6] Although ideas from this project remain important, more attention to integrating models and qualitative evidence is needed. First, the past decade has witnessed a resurgence of thinking on qualitative methods, producing important insights relevant for evaluating formal theories. Second, and perhaps more fundamentally, Bates et al. (1998) sought to show how formal and non-formal rational choice theories could contribute to the idiographic tradition by 'account[ing] for particular events or outcomes' (10). Their concern with using models to understand

particular cases differs from our broader goal of using qualitative evidence to evaluate existing models that aspire to greater generality. Our literature review suggests this objective is far more common than an analytic narrative's goal of using a model to elucidate a single case. Among the articles we reviewed, only Nalepa (2010) claimed to follow the analytic narratives template—and this article includes three country case studies.[7]

3. The distinct advantages of process tracing

The broad goal of empirically evaluating a model is often conflated with using statistics to evaluate a model's covariational predictions (e.g., Granato et al., 2012; Morton, 1999).[8] However, despite the importance of comparative statics predictions, a model is not simply a claim about the relationship between an X and a Y. Instead, a model embodies claims about a *causal process* by which X influences Y. This corresponds directly with the goals of process tracing, which involve 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' (Bennett and Checkel, 2014: 6). That is, 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. The analyst uses many facts from a case, sometimes called 'causal process observations,' to evaluate the plausibility of a particular causal mechanism (Collier et al., 2004).

Modeling and process tracing are also closely related in the kind of thinking they require. Using process tracing requires a scholar to 'clarify as much as possible…[w]hich actors should have known, said, and did what, and when' (Bennett and Checkel, 2014: 30). By definition, a model does this. Indeed, writing down and solving a model often forces the modeler to ask new questions about what choices are available to decision-makers and at what points, what their motivations and goals might be, and what information they had available to them. These considerations generate numerous implications about events in a particular case that can be evaluated empirically, whether by the scholar that produced the model or by others with knowledge of relevant cases. The shared concern with understanding the processes leading to an outcome makes process tracing a natural tool to use when evaluating a formal model.

Current practice reflects this affinity between causal mechanisms and process tracing. In our survey of research articles, 12 percent presented qualitative cross-case comparisons while almost three times as many (34 percent) provided qualitative evidence focused on mechanisms. This took two forms. About one-quarter (23 percent) of the articles offered qualitative evidence to justify key assumptions. Just under one-fifth (18 percent) of articles included detailed case studies, averaging almost four pages in length. Despite disclaimers of the type mentioned above, these case studies are clearly intended to demonstrate that a model offers insight into the events discussed.

One view might be that such case studies are merely preliminary work, throwaway 'suggestive' evidence that serves as a placeholder until the day when the real

work of statistical testing can be conducted. This is not our view. No single test, whether quantitative or qualitative, can be conclusive, but each moves the needle (in a loosely Bayesian sense) by increasing or decreasing our confidence that a model captures key aspects of a particular empirical context. Moreover, the shared concern of models and process tracing with causal mechanisms—as distinct from narrowly defined causal relationships between two variables—means that a good process tracing exercise can be as or even more valuable than a statistical test for assessing the empirical value or limitations of a model. By contrast, standard statistical evaluations of models often fail to provide insight either into the empirical validity of key assumptions that generate the model's comparative statics predictions, or into a model's intervening implications.

Process tracing also offers several practical advantages for evaluating models, which the following sections discuss in detail. Quite often, off-the-shelf datasets do not adequately measure key parameters and choice variables. Even more rarely do they provide insight into actors' perceptions of their strategic setting. A close examination of a smaller number of cases can be preferable to a large-*N* test with more questionable measures (see Section 5). Furthermore, especially in models with country-level implications, the number of cases that correspond to the scope conditions of the formal model might be small enough to make conventional statistical tests infeasible (see Section 6). Finally, if a model falls short in a particular case, then the kind of information provided by detailed process tracing plays a crucial role in generating new formal models (Section 7).

Our intent here is not to advocate abandoning statistical tests in favor of process tracing. The value-added of quantitative evaluation is well-known. Our point is simply that process tracing has distinct strengths for evaluating formal models that make it just as important for assessing formal models as statistical tests. However, using process tracing to evaluate formal models raises particular challenges that the qualitative methods literature does not consider. The remainder of this article discusses characteristics that distinguish convincing process tracing evaluations of formal models from unconvincing ones. We illustrate these features with examples from research that introduce a new model and provide qualitative evidence to support it, as well as from research that assesses existing formal models. We focus on three key areas in qualitative methods research: how to evaluate mechanisms, how to gather evidence, and which cases to select. Table 1 summarizes the standards.

4. Using process tracing to evaluate mechanisms from a formal theory

In order to empirically evaluate mechanisms from a formal model, it is useful to distinguish between two different objects of investigation usually conflated in the process tracing literature. First, a model makes assumptions about the causal process. Formal models specify what choices players must make, in what order, with what information, and with what objectives. Second, a model's assumptions create observable implications for the causal process, including which choices and outcomes should be observed.

Table 1. A checklist for evaluating a formal model using process tracing.

Topic	Aspects to consider	Advice
Evaluating mechanisms	Assumptions	• Distinguish between a model's assumptions and implications • Recognize all assumptions are simplifications • Assess how changing the assumption could change the implications
	Implications about the causal process	• Identify observable implications of the causal process implied by the model • Evaluate one implication across more cases or many implications in one case-
	Alternative explanations	• Examine evidence that distinguishes among alternative explanations (formalized or not) • To challenge a model most effectively, present an alternative or a sketch of one
Gathering evidence	Accurate measurement	• Take advantage of small sample size to carefully measure variables • Consider multiple ways to measure key variables • Provide detailed coding rules to facilitate replication
	Transcript evidence	• Use transcript evidence to provide insight into actors' information, beliefs, and perceptions • Recognize that speech is a strategic choice
Selecting cases	Fit with model	• Select cases that match key assumptions of the model as closely as possible
	'Biased' selection	• Look for cases that isolate the model's mechanisms relative to alternatives • Look for cases likely to produce relevant transcript evidence

Let us begin with a prominent example, Acemoglu and Robinson's model of regime change (2006a). Their core model *assumes* that all redistribution occurs through a flat tax, whose proceeds are handed out equally to all members of society. It also *assumes* that in a democracy, this tax rate will be determined through a one-person-one-vote system. By contrast, it *implies* that the wealthy will prefer autocracy to democracy and the poor will prefer the opposite, and that given the opportunity, each group may mobilize politically to achieve its desired regime. These two assumptions and two implications can each be evaluated using process tracing. Our confidence in the empirical value of any model depends significantly on our evaluation of assumptions like these, even though they cannot be statistically tested.[9]

Below, we first discuss how to use process tracing to evaluate a model's assumptions and implications, and then discuss the importance of comparing the model's

mechanisms to alternative explanations. These topics constitute the core elements of using process tracing to evaluate formal models, assuming the researcher has already chosen cases and collected data (discussed later in the article).

4.1. Evaluating assumptions about the causal process

Evaluating a model's assumptions is a particularly contentious issue. At one extreme, Milton Friedman (1966 [1953]: 14) argues that assumptions should never be evaluated because they are not meant to be accurate descriptions of reality. Instead, Friedman asserts that we should restrict attention solely to a model's predictions. However, this position does not help to adjudicate between different models that explain the same set of facts, nor does it provide guidance for revising models with weak predictive power. At the other extreme, Bates et al. (1998: 14) present as a key criterion of a good analytic narrative that 'the assumptions fit the facts.' Yet, whether an assumption fits the facts is also a poor standard. All models and indeed all theories inherently simplify and therefore contain assumptions that do not exactly 'fit the facts,' so this standard provides an easy target for unproductive critiques.[10]

Developing a satisfactory middle ground between Friedman and Bates et al. is crucial. As noted above, about one-quarter of the articles in our sample use qualitative evidence to justify the model's assumptions. In other words, the authors of these models view the empirical appropriateness of their assumptions as an important criterion of evaluation.

Although it is impossible to draw a bright line between satisfactory and unsatisfactory assumptions, we propose two criteria that establish a pragmatic middle ground.[11] The most compelling critique replaces a disputed modeling assumption with an alternative assumption that (a) more closely matches the empirical context in question and (b) significantly changes the model's conclusions. The first criterion captures the importance of empirical relevance, but the second criterion rules out easy criticisms that a formal model simplifies.[12] Any theory, formal or not, simplifies the social world to facilitate a tractable analysis. Individuals frequently behave in what appear to be irrational ways. States, parties, corporations, interest groups, and other organizations do not always behave as unitary actors. People's beliefs about the world do not exactly follow Bayes' rule. Demonstrating these facts tells us very little about whether a model is a useful simplification for understanding an important strategic tradeoff.

Other assumptions, however, are more important to justify when presenting a model or to evaluate when critiquing a model. For example, Svolik (2009) shows how a moral hazard model provides insights into a strategic interaction between a dictator that desires to concentrate power and a ruling coalition that attempts to maintain a power-sharing arrangement. A key assumption in this model is that the principal (the ruling coalition) receives informative but imperfect signals about the agent's (dictator's) actions and must reward or punish based on this information (by attempting a coup or not). Svolik demonstrates the empirical relevance of this assumption by providing examples in which leaders' attempts to consolidate power generated these signals. In the Soviet Union, Lavrentiy Beria merged formal

ministries after Stalin's death to concentrate more power in his hands. In Iraq, Saddam Hussein gradually replaced earlier supporters with loyalists from his hometown.

With regard to critiquing a model's assumptions, it is easier to evaluate the impact of changing an assumption if two competing formal models differ only on that one assumption. Although this is rarely the case, we can still make a more subjective judgment of whether a particular change in assumptions would likely meaningfully change the conclusions. Consider, for example, recent evaluations of audience cost models with qualitative evidence. This family of models builds on Fearon (1994, 1997), who presented a model of international disputes in which 'the side with a stronger domestic audience (e.g., a democracy) is always less likely to back down' (1994: 577). Audience costs, or the cost a leader would pay for backing down in a dispute, more rapidly lock democratic leaders into a stated position. Snyder and Borghard (2011) challenge this theory in part because they disagree with the key assumption that 'publics care a great deal about consistency between threats and deeds, independent of their preferences on policy substance' (439), among others.

Snyder and Borghard therefore examine crises in which democratic leaders made threats and failed to carry them out. Although they only found one case in which domestic audiences imposed *no* costs on the leader, they provide considerable evidence against the theory even in the apparently confirmatory cases: domestic audiences punished leaders only when those audiences were already pro-war. This contrasts with the model's assumption that audiences will punish leaders primarily for failing to back their threats up. Of course, proponents of audience costs can counter that this evidence does not fully rule out the counterfactual implied by the model: those leaders had weaker public support after making these threats and then backing down than if they had never made the threats in the first place (Fearon, 2013). Nonetheless, the lack of evidence for this effect suggests the trade-off assumed in the formal model is not highly relevant in the cases examined.

Regardless of one's assessment of Snyder and Borghard's evidence,[13] their critique satisfies our standard for how to assess an assumption. The assumption that audiences want their leaders to back up threats independently of whether they actually believe the prize is worth fighting for is not merely a simplification, but instead is fundamental to the theoretical logic of this family of models. If this assumption is rarely empirically relevant or its effect is negligible in practice, then audience cost models offer little additional explanatory leverage relative to a simpler theory in which democratic leaders largely follow the electorate's preferences.

4.2. Evaluating observable implications about the causal process

Process tracing can also evaluate whether events from empirical cases match a model's implications. 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. The absence of such evidence provides grounds to question whether the mechanisms implied by the model help to explain empirical cases. Observable implications can be assessed

either by examining small numbers of implications across a larger number of cases or by examining many implications within a single case.

For example, in Acemoglu and Robinson's (2006a) regime transition model, high inequality (independent variable) affects reversions from democratic rule (dependent variable) because the rich want to end high taxation the poor impose on them (intervening variable). Haggard and Kaufman (2012) measure these three key variables for a large sample of democratic reversions. They find that of the seven democratic reversions in high-inequality countries, only two cases exhibited evidence that elites ended democracy to prevent redistribution. In other cases, redistribution either played a negligible role or populist autocrats came to power offering *more* redistribution than had occurred under democracy. These findings cast doubt on the empirical relevance of Acemoglu and Robinson's theory for explaining regime transitions, even if the predicted correlation between independent and dependent variables was robust.

Alternatively, one can evaluate more mechanisms implied by a model with a single case. For example, Dunning (2008) argues that under certain conditions resource wealth (independent variable) can promote democratic stability (dependent variable) by mitigating the wealthy elite's concerns that a democratic government pandering to the masses will expropriate their income (intervening variable). Dunning's qualitative analysis of Venezuela starts by closely tracing the rise and fall of oil rents and of democratic stability. However, he goes beyond this qualitative analog of time series analysis to provide various pieces of evidence that the core mechanisms implied by the model were operating. For instance, when oil rents were high, elites did not object to the high levels of public benefits provided to the masses (163–166). This is consistent with the model's implication that elites should only protest public spending funded by high tax rates. Similarly, as oil rents fell, Dunning shows politics became polarized around classes and redistributive conflicts and ultimately facilitated the rise of the populist Hugo Chavez (166–183).[14]

4.3. Examining alternative explanations

Whether validating or challenging a model, process tracing should also address alternative hypotheses. Bennett (2010) places this consideration at the heart of the method: 'Process tracing involves the examination of "diagnostic" pieces of evidence within a case that contribute to supporting or overturning alternative explanatory hypotheses… What matters is not the amount of evidence, but its contribution to adjudicating among alternative hypotheses' (208–209).

For example, in a study on trust in international relations, Kydd (2005) formalizes Robert Jervis's well-known spiral model. One important implication of the formalization is that spirals of 'unjustified mistrust'—circumstances in which both states are security-seekers that nevertheless view each other as expansionist and threatening—should be quite rare. Kydd then evaluates this model in the context of the origins of the Cold War between 1945 and the outbreak of the Korean War in 1950. Kydd directly addresses many existing explanations that historians and political scientists have posited for these events, which greatly increases confidence that

the model provides valid insights into this important series of events. Similarly, Slantchev (2011) argues the expansion of the Korean War resulted from deliberately ambiguous military signals made by the Chinese, and explicitly evaluates this explanation against previous analyses that had attributed the war to an American failure to accurately read Chinese intentions.

Frequently, however, political scientists introducing a new model often provide evidence consistent with their own explanation but do not consider existing alternative hypotheses. In addition, even when scholars do consider alternative explanations, they often only address rival formal models. For instance, Miller and Schofield (2003) set their explanation for the gradual shift in American party alignments over the past century against formal models they interpret as implying 'chaos' (constantly changing or unpredictable policy positions) or 'convergence' (policies fixed at the median voter), but do not address any competing non-formalized explanations for this shift.[15]

Evaluating alternative explanations is also vital when challenging formal models. Judging a model against the ideal of perfect consistency with all available facts rather than against a specific alternative explanation can result in unconstructive critiques of oversimplified assumptions, such as appeared in Elster's (2000) critique of *Analytic Narratives*. This concern may even raise questions about more pertinent critiques. Schultz (2012) and Slantchev (2012) both argue that the arguments against audience costs theories made by Trachtenberg (2012a) and Snyder and Borghard (2011) are weakened because they do not explicitly delineate an internally consistent alternative explanation, although Trachtenberg (2012b) rejects this criticism.

5. Gathering evidence to evaluate a model's mechanisms

Process tracing both permits and requires more attention to gathering evidence than is generally possible in statistical analyses. This has two implications. First, key variables should be operationalized and measured carefully and precisely. Second, scholars should use a broader range of evidence, such as transcript evidence (the public or private communications of actors directly involved in the case).

5.1. Carefully measuring key variables

Properly measured variables are a necessary foundation for any convincing empirical analysis. Models, especially in international relations, frequently encompass difficult-to-measure concepts such as actors' perceptions of the distribution of power. It may be preferable to study covariational patterns among a small number of personally measured cases rather than to use an off-the-shelf measure that does not adequately measure the underlying concept—even if the latter facilitates a larger sample. Ideally, these variables should be measured in multiple ways to assess whether different measures produce divergent findings. If they do, then it is important to defend why one measure should be preferred over others. Although this advice applies for any case study, researchers presenting qualitative support for their own models rarely consider multiple measures.

Slantchev's (2011) Korean War case study in his book on military threats illustrates the importance of considering different measures. This study operationalizes the 'distribution of power' in terms of the number of troops in the theater of operations, which in this case means Chinese forces in Manchuria and U.S.-led United Nations (UN) forces on the Korean Peninsula. Because China enjoyed an almost 70 percent share of troops in-theater in September 1950, this operationalization implies China would have deterred the United States from crossing the parallel by issuing a clear military threat at this time. By October, however, since China had as little as 57 percent of the troops in-theater, it would appear that even a clear Chinese threat to intervene would not have deterred the United States.

However, this conclusion is sensitive to the choice of measure. Consider two reasonable alternative operationalizations of the distribution of power. On the one hand, if military capabilities and odds of victory are measured in terms of China's entire standing army of more than 5 million soldiers in 1950 that it could bring to bear in Korea, set against all U.S. forces in Asia, China should have been able to continue to deter the United States even in October. On the other hand, the qualitative inferiority of the Chinese army suggests that China would not have been able to deter the United States even in September with only its troops in Manchuria. The Chinese army was primarily a guerilla force consisting of light infantry and some artillery, and the American army had a clear advantage in armored vehicles, air support, and other technologies. Because these different measures of power yield different predictions than Slantchev's operationalization, his otherwise careful presentation of evidence is less compelling than it would have been had he addressed whether his chosen measure improves upon these alternatives.

Considering multiple measures is equally important when arguing against the empirical applicability of a model. Haggard and Kaufman (2012) strengthen their critique of redistributional regime change theories by considering a wide variety of inequality measures. This reduces the possibility of counting a case against the theory because it scored low on one inequality measure but not another. Although the authors do not provide multiple measures when creating a new variable—a coding of whether redistributive conflict plausibly played a causal role in a particular transition—they provide transparent coding rules and detailed documentation for how each case was coded (Haggard et al., 2012).

5.2. Using transcript evidence to assess actors' beliefs

Effective process tracing also requires and facilitates using a broader range of evidence than is typical of statistical analyses. In particular, when treated with appropriate caution, transcript evidence—what decision-makers in the case and those around them actually said or wrote—can help assess whether actors appeared to perceive key tradeoffs from the formal model and how they perceived outcomes that they took actions to avoid.[16]

For example, Goemans and Fey (2009) leverage transcript evidence when using process tracing to evaluate a model that explains risky behavior by national leaders in international conflicts. Their model suggests leaders with weak domestic support

will take military actions that decrease the overall probability of winning a war if those actions simultaneously increase the odds their country will decisively win the war—the only outcome in which the tenuously supported leaders will stay in power. Here, the key intervening variable is how leaders perceive their actions will affect variance in the war's outcome. Empirically, they examine Germany's decision in January 1917 to return to unrestricted submarine warfare. Consistent with the model's implications, Goemans and Fey provide evidence from contemporary speeches, reports, and armistice negotiations that German leaders risked this gamble because, without additional action, Germany would not have been able to achieve a deal favorable enough for its leaders to keep their jobs—even though this gamble lowered Germany's expected settlement terms by raising the likelihood the United States would enter the war.[17]

Of course, individuals often have incentives to lie to justify their actions. This highlights the importance of addressing possible biases when presenting transcript evidence—which the formal model may provide insight into. It would go too far to treat direct statements from political actors as necessary evidence for a particular interpretation of their actions. Gartzke and Lupu (2012) argue that critics of audience cost theory make this mistake by treating the lack of instances in which political leaders admit they were trying to create audience costs as evidence against the theory. It may not be wise for political leaders to admit, even after the fact, that they were trying to make war over an issue unavoidable. Or they may simply not be introspective enough to write down their thoughts afterwards.

Nonetheless, statements made after a political actor steps down from a formal post can provide important supportive evidence. For example, Johns' (2007) model of international bureaucrats with multiple principals suggests that when one principal has a strong outside option that gives it substantial bargaining leverage over the final outcome, a bureaucrat whose preferences are more moderate will disclose less information than it would have if either (a) both principals were equally powerful or (b) its preferences were more closely aligned with those of the stronger principal. Johns supports this finding by quoting UN representative Hans Blix, who admitted to obscuring important information when reporting the results of his weapons inspections in Iraq to the UN Security Council because he knew U.S. preferences would likely determine the outcome.

Overall, transcript evidence offers a unique opportunity to assess actors' beliefs. The potential pitfalls of transcript evidence indicate that researchers should carefully interpret this evidence and acknowledge that these interpretations are subject to debate. However, these pitfalls do not mean that such evidence should be disregarded and cannot be used (contra Bates et al., 1998: 698; Beck, 2010). Furthermore, the difficulty of incorporating such evidence into a statistical framework highlights another distinct advantage of qualitative tests.

6. Selecting cases to isolate the model's mechanisms

The above discussion presumes the analyst has chosen cases for which we should expect a model's assumptions and implications to find support. However, no

model will provide insight into all cases. Qualitative methodologists have made notable progress in the past decade by articulating different ways cases can be selected and by explaining how particular case selection procedures serve different research goals.[18] However, distinct issues arise when evaluating formal models, leading us to make two recommendations for selecting cases. First, the case should match the model's assumptions as closely as possible. Second, and contrary to standard statistical practices, sampling on particular values of the main independent and/or dependent variables may be more likely to distinguish the model from alternatives and to yield useful transcript evidence.

First, recommending that the cases selected should match the model's assumptions as closely as possible resembles the goal of choosing cases that fit the model's scope conditions. Although no case will perfectly match the precise technical setup of the model—nor will a bright line distinguish cases that are conclusively in or out of scope—a model should be more likely to correspond with empirical cases for which the model's assumptions match more closely. This provides an important practical advantage of process tracing. For many important theories and models, the total number of cases meeting this criterion will often yield low-powered statistical tests. This is particularly true for the country-level theories that are often the focus of comparative politics and international relations research.

For instance, Slantchev (2011) solves his model of military threats in the context of two-party conflicts for reasons of tractability. The model likely offers insights into threat-making in conflicts with three or more actors. However, because the model neglects additional dynamics that might arise with more parties, Slantchev examines the two-party setting of U.S.–Chinese interactions in the expansion of the Korean War. Similarly, Dunning (2010) presents a model of endogenous oil rents for which the tradeoffs from the model should be relevant in 'democratic polities in which the prospect of continued and regular elections is relatively high' (382). His case of Venezuela meets these conditions, whereas the actions of authoritarian rulers who maintain their position via family succession rules or through brute force may not be well-explained by a model that assumes electoral competition. For both studies, examining cases that fit less well with the model's assumptions would have been less useful.

Second, sampling cases with particular values of the independent and/or dependent variables is often the best strategy for assessing whether a model's mechanisms help to explain known outcomes (Collier and Mahoney, 1996),[19] in contrast to case selection principles derived from statistical research (Geddes, 1990). A model's predictions may only be distinguished from prominent alternative explanations for certain values of the independent variable(s). Similarly, actors may only be likely to produce transcript evidence relevant for distinguishing the model from alternatives when certain conditions are present.

Goemans and Fey's (2009) analysis of Germany's strategy in the final years of the First World War illustrates this approach. As discussed, they use transcript evidence to show that the logic of their model explains the outcome of risky escalation to unrestricted submarine warfare better than alternative explanations assuming incomplete information. In particular, they summarize transcript evidence in which

German leaders expressed fears about whether they would be able to stay in office if they were unable to bring about a 'sufficient peace' that would yield material spoils of war for the German public. By contrast, we would not expect to find transcript evidence that would distinguish their theory from alternatives in cases with different values of either the independent or dependent variable. To see this, suppose the leaders would retain their positions regardless of the war outcome. Even if the officials explicitly stated they never considered escalating war tactics, this transcript evidence would not distinguish Goemans and Fey's model from an alternative explanation that leaders never take risky gambles. Furthermore, we would certainly not expect secure leaders to produce transcript evidence about whether they would consider taking a risky gamble were their jobs less secure. Finally, regardless of leaders' job status, if Goemans and Fey studied a minor power that could not have escalated a war on its own, once again we would not expect to encounter transcript evidence useful for evaluating their institutionally induced escalation model.

Similarly, Dunning's (2008) model discussed earlier implies that greater resource wealth may tend to preserve rather than weaken democracy for a particular value of a conditioning factor: non-resource income is distributed very unequally. Examining the case of Venezuela, an oil-rich country with high non-resource inequality, he finds considerable evidence that the association predicted by the model is generated by the mechanism in the model. However, suppose he had instead analyzed a country with no resource wealth, or a country with resource wealth but a very even income distribution. Table 2 compares the predictions of Dunning's model with Acemoglu and Robinson's (2006a) model under different combinations of the inequality and resource variables.

In case types 1–3, Dunning's model does not offer different predictions than Acemoglu and Robinson. Therefore, although Dunning may have found no evidence that elites were using resource wealth to mitigate redistributive pressures in these countries—consistent with his model—such a finding would do little to enhance our confidence that his model offers novel insights. Only case type 4 distinguishes Dunning's model from Acemoglu and Robinson.

These standards for case selection are useful both when a researcher is trying to provide initial empirical support for a model and when the goal is to challenge an established model. When developing a model, finding strong evidence for the model's explanatory power in one or a few cases is sufficient to make the argument that

Table 2. Comparing predictions from Dunning and Acemoglu and Robinson.

Case type	Non-resource income inequality	Natural resources	A&R predict elites fear redistribution?	Dunning predicts elites fear redistribution?
1	Low	Low	No	No
2	Low	High	No	No
3	High	Low	Yes	Yes
4	High	High	**Yes**	**No**

it warrants further empirical investigation. Similarly, when the research goal is to evaluate an existing model, attention should be focused on cases where its processes should be most apparent because the model's proponents may rightfully dismiss contrary evidence from less pertinent cases.

Both recent challenges to audience cost models follow this principle. Snyder and Borghard (2011) survey all post-1945 international crises involving the United States, and then focus on cases in which a democratic leader made a public threat and domestic public opinion was mobilized around the issue. Democratic leaders' anticipation of audience costs should also affect their escalation decision even if they do not make such threats, but it would presumably be more difficult to find evidence of these costs. Trachtenberg (2012a) instead restricts attention to great power crises that involve a democracy and do not end in war. Both studies aim to evaluate the model in a best-case scenario by looking for cases that fit the scope conditions of the model as closely as possible, minimize confounding factors, and in which the mechanisms of the model should be isolated from alternative explanations.

Relatedly, as discussed, Haggard and Kaufman (2012) examine every regime transition worldwide from 1980 to 2000 to assess redistributive regime transition models. These theories have implications for politically stable regimes as well, but there are two good reasons to focus on transitions. First, a society at the crux of transition is most likely to produce transcript evidence useful for assessing how important are inequality-driven social tensions. Second, the originators of these models advance general claims: Boix (2003) aims to offer 'a theory of political transitions and regime choice' (10) and Acemoglu and Robinson (2006a) offer a 'framework for analyzing the creation and consolidation of democracy' (front matter). Such claims of broad applicability have led these models to be so influential in our thinking. However, by the same token, failure to find evidence of the models' mechanisms in nearly every recent transition case suggests either that their scope is more sharply delimited than initially claimed, or that the strategic factors and tradeoffs they highlight are considerably smaller in magnitude than had been thought.

Notably, despite Snyder and Borghard's, Trachtenberg's, and Haggard and Kaufman's concerted attempts to sample cases in which the model's mechanisms are highly visible, much of the pushback against their conclusions also focuses on case selection. Boix (2013: 12), for example, argues that the inequality/redistribution link is most important and therefore more easily observed in much earlier democratizations, whereas this mechanism is less central and more easily obscured by confounding factors in regime changes during the later period Haggard and Kaufman examine. Similarly, Slantchev (2012) defends audience cost theory by pointing out that: 'in several of the cases Trachtenberg studies, one would not even expect [audience cost theory's] mechanism to be at work for the simple reason that both sides had eschewed the use of force' (377). Case selection plays a central role in these debates, highlighting the need for careful attention to scope conditions and the likely prominence of a model's central mechanism in a case.

7. What to do when process tracing does not support a model

Having addressed how to assess a model, the question remains of what to do if the evidence goes against it. We argue that qualitative evidence plays an equally important role as quantitative evidence in updating our confidence about the empirical relevance of a model. Moreover, the kind of qualitative evidence that results from process tracing is usually more useful than quantitative findings for generating new theories or for revising existing ones.

The debate about Acemoglu and Robinson's (2006a) and Boix's (2003) redistributive models of regime transitions illustrates these arguments. Above we summarized Haggard and Kaufman's (2012) research, which found little evidence from cases in recent decades for the central idea that redistributive pressures have affected regime transitions. Slater et al. (2014) have additionally shown that certain core assumptions do not find support in Southeast Asian cases that otherwise fit the models' scope conditions: the military is not a consistent ally of economic elites, nor do many post-colonial states have the infrastructural capacity to implement widespread redistribution. This qualitative evidence is crucial for casting doubt on some key assumptions and implications of these models and therefore complements statistical tests that demonstrate weak support for the core comparative statics predictions linking inequality and regime type (Acemoglu et al., 2013; Soifer, 2013). The sum of this evidence strongly supports the conclusion that redistributional pressure is not a central factor to analyze for understanding modern regime changes.

What should be done if process tracing casts doubt on the empirical relevance of a model? The next step is to work toward a new theory that will serve as a better guide to understanding the particular context. Regardless of whether one plans to formalize the theory, further progress will require careful thought about which assumptions were empirically irrelevant in the rejected model and what alternative assumptions might serve better. This will usually be founded in qualitative case knowledge, which highlights another distinct advantage of process tracing research. For instance, one could take Slater et al.'s (2014) findings and explore the implications of models with an independent military, or in which the total amount of possible redistribution is greatly circumscribed. Or one could build off Haggard and Kaufman's (2012) observation that autocrats may build support through populist redistributive strategies.

By contrast, statistical findings provide little guidance for future theorizing. Suppose one does not accept that qualitative evidence can meaningfully cast empirical doubt on a model, but accepts Acemoglu et al.'s (2013) conclusion that inequality has 'no robust causal relationship' with regime transitions in any methodologically sound statistical analysis. Where should one go next? Qualitative research, most likely including some form of process tracing, will still be required in order to lay foundations for a better theory.

As a final point about the role of models in the larger social science enterprise, models have many uses and even 'wrong' models can be valuable in moving our thinking forward.[20] The central issues raised by redistributive transition models and some of the

techniques used to analyze them will play an important part in subsequent theories developed to explain political transitions (Boix, 2013) and even other topics.[21] Moreover, the difficulty of conclusively evaluating the empirical relevance of a particular model and the value a model can have even if it falls short empirically demonstrate why we should reject the position that formal theories should not be published for wider consideration until they have been empirically tested.[22] The highly informative debates about redistributive causes of regime transitions and about audience costs simply would not have been possible without the parsimonious and clear formal frameworks provided by these foundational formal models.

8. Conclusion

Although there has been extensive discussion of when and how to link formal theories and statistical evidence, the role of qualitative evidence in evaluating formal theories has been largely neglected. Nevertheless, as we have demonstrated, using qualitative evidence to validate formal models is actually common practice among formal theorists—particularly in international relations and comparative politics—and qualitative scholars are also becoming more actively engaged in evaluating formal models. This is not accidental, but results from the focus that case studies, especially those using process tracing, share with game-theoretic models on understanding causal mechanisms. Yet, rather than capitalize on this affinity, scholars versed in formal theory are often vague about or explicitly denigrate inferences drawn from qualitative evidence, leaving readers with no clear standard by which to judge the evidence provided. Relatedly, many formal theorists appear to be largely unaware of the substantial body of recent research refining and advancing techniques for qualitative analysis. On the other hand, the qualitative methods literature pays almost no attention to the specific concerns that arise with empirically assessing formal models using case studies. Modelers and qualitative scholars often seem to talk past each other for the simple reason that criteria for empirical validation remain unclear.

This disconnect between formal models and current qualitative research is unfortunate. Both approaches share a common concern with mechanisms underlying causal relationships, and would benefit from greater integration in the collective effort to understand political phenomena. Toward this end, we have offered a new set of practical standards for using process tracing to evaluate formal models. These standards apply both to formal modelers aiming to convince a broader audience that a new model has empirical purchase in substantively important settings, and to scholars who seek to evaluate the empirical value of a model. This article represents one step in bridging the gap between two vibrant but often disconnected intellectual communities.

Appendix: survey of evidence accompanying formal models

To quantify how frequently and in what manner qualitative evidence is used in conjunction with formal theory in studies of comparative politics and international relations, we surveyed all articles with formal models published from 2006 to 2013

across seven major journals, including both general interest journals and ones focused on these subfields. Articles were identified in the *American Journal of Political Science, American Political Science Review, Comparative Political Studies, International Organization, Journal of Politics*, and *World Politics*. *Comparative Politics* was also surveyed but no articles were identified in this journal during this period that used formal models. Articles were classified as having formal models and therefore included in our survey if they included analytical propositions and specified an equilibrium concept. Rational actor assumptions were not required, but models analyzed using primarily computational techniques were excluded for simplicity, although their empirical evaluation faces many similar issues. To slightly narrow the scope of our study, we also excluded articles that explicitly addressed institutions unique to American politics, or which referred only to examples or empirical studies from American politics. One hundred eighty-two articles fit these criteria. Our rules for coding the empirical methods used in these articles follow. The coding database lists all 182 modeling articles in our survey and is posted on www.peterlorentzen.com.

Coding rules

- *Quantitative empirical evaluation*: the article offers statistical evidence that an implication of the model is borne out, usually with an explicit hypothesis test and confidence interval. Such claims generally involve original data collection or analysis, and follow the presentation of the model's key results. References to previously established statistical regularities were not counted.
- *Qualitative empirical evidence*: the article offers significant qualitative evidence and argumentation to support the applicability of the theory. Brief examples motivating the modeling exercise, such as references to well-known events or previous empirical findings that the model aims to explain, were not coded in this category. We further divided articles in this category into three subcategories (not mutually exclusive): cross-case comparison, case studies, and evidence for assumptions.
  - *Cross-case comparison*: the article compares two or more cases (either across space or across time), arguing that the patterns of covariation support the model's predictions. This is the classic 'small-*n*' comparison.
  - *Case studies*: the article aims to convince the reader that the model accurately characterizes or explains a sequence of events in one or a small number of closely examined cases. These case studies engage in process tracing, if only implicitly.
  - *Qualitative evidence for assumptions*: the article provides significant qualitative justification for some aspect of the setup of the model or the key tradeoffs the model examines. Evidence is offered that these assumptions are frequently true, or at least true in certain important cases. This kind of evidence typically appears prior to presenting the model, or is concurrent with the model presentation.

Acknowledgements

Declaration of conflicting interests

Funding

Notes

1. For examples of the former, Green and Shapiro (1994) and Shepsle (1995) provide influential early statements. Little and Pepinsky (2016) and Svolik (2015) revisit the topic. For examples of the latter, see Morton (1999), Clarke and Primo (2012), and Granato et al. (2012).
2. We focus on international relations and comparative politics because qualitative evidence has been far less prominent in formal theoretical research on American politics.
3. Rodrik (2015) addresses the closely related but more concrete challenge of assessing whether a particular economic model serves as a useful guide for solving a specific policy problem.
4. The Appendix provides additional details on this survey, and an accompanying document lists all 182 modeling articles in the survey.
5. Prominent examples from recent years include Kydd (2005), Acemoglu and Robinson (2006a), Slantchev (2011), Gailmard and Patty (2012), and Gingerich (2013).
6. Levi (2004) and Greif (2006) discuss and develop the analytic narratives methodology further. Buthe (2002) discusses how to effectively use historical narratives to evaluate models, broadly construed, but does not address the specific challenges of evaluating formal theories. Laitin (2002, 2003) argues for the complementarity of qualitative, quantitative, and formal approaches. Recent work by Kuehn (2013) and Goemans and Spaniel (2016) also advocate the combination of process tracing and formal models. We build upon these contemporaneous contributions by comprehensively marrying formal models with a broad spectrum of best-practice process tracing methods and by thoroughly summarizing existing practice.
7. See also Lorentzen and Scoggins (2015), although they do not use the term.
8. The qualitative analog of this is cross-case comparisons, which can provide some basis to evaluate such predictions where a large sample is unavailable. The literature on this approach is extensive, dating back in political science to Lijphart's (1971) article on the comparative method. Tarrow (2010) and Slater and Ziblatt (2013) provide more recent

statements that could easily translate to testing formal models, so we will not discuss this approach here.

9. By contrast, another implication of this model is that countries with high income inequality should be less likely to democratize. This kind of implication about outcomes is natural to test statistically.
10. Healy (2016) provides a recent discussion of similar issues in sociology.
11. Inexact standards are not limited to qualitative methods. For example, there are only gradations of plausibility assessing whether a natural experiment is well-designed or whether a lab experiment has external validity. Clarke and Primo (2012: 97–100) provide a broader discussion of 'the illusion of precise standards' for evaluating a formal model.
12. For example, Elster's (2000) review of Bates et al. (1998) attacks the authors for assuming rational behavior and for treating collectives as unitary actors, among other issues.
13. For a vigorous discussion of this article and the closely related Trachtenberg (2012a), see Schultz (2012), Slantchev (2012), Trachtenberg (2012b), and other symposium articles in the same issue of *Security Studies*.
14. Lorentzen (2014) provides another example.
15. Mayhew (2000) reviews this extensive literature.
16. In other words, what were their perceptions about outcomes off the equilibrium path? Weingast (1996) provides a lengthier discussion of how scrutinizing off-the-equilibrium path behavior in game-theoretic models can inform historical analysis.
17. In a similar vein, Lorentzen (2013) uses evidence from interviews with Chinese protesters to validate a model of protest as an information gathering institution.
18. Lieberman (2005), Seawright and Gerring (2008), and Fearon and Laitin (2008) discuss how to select cases from a large dataset after an initial statistical analysis is complete. Bennett and Elman (2006) survey earlier developments in this literature.
19. Gerring (2007) suggests that cases of this type be referred to as 'pathway' cases and offers additional useful suggestions on their selection and usage. This approach is related to but distinct from Eckstein's (1975) conception of 'most likely' cases.
20. Powell (1999) and Clarke and Primo (2012) discuss these points more thoroughly.
21. For instance, the role of natural resources in preventing civil wars (Paine, 2016).
22. Hill (2005), for instance, asserts that this view was near-universal among reviewers for the *American Journal of Political Science* for at least a period of time.

References

Acemoglu D, Naidu S, Restrepo P, et al. (2013) Democracy, public policy, and inequality. *Newsletter for the Comparative Democratization Section of the American Political Science Association* 11(3): 2, 16–20.

Acemoglu D and Robinson JA (2006a) *Economic Origins of Dictatorship and Democracy*. New York: Cambridge University Press.

Acemoglu D and Robinson JA (2006b) Economic backwardness in political perspective. *American Political Science Review* 100(1): 115-131.

Bates RH, Greif A, Levi M, et al. (1998) *Analytic Narratives*. Princeton, NJ: Princeton University Press.

Beach D and Pedersen RB (2013) *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor, MI: University of Michigan Press.

Beck N (2010) Causal process 'observation': oxymoron or (fine) old wine. *Political Analysis* 18(4): 499–505.

Bennett A (2010) Process tracing and causal inference. In: Brady HE and Collier D (eds) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield, pp. 207–220.

Bennett A and Checkel JT (2014) Process tracing: from philosophical roots to best practices. In: Bennett A and Checkel J (eds) *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool*. New York: Cambridge University Press.

Bennett A and Elman C (2006) Qualitative research: recent developments in case study methods. *Annual Review of Political Science* 9: 455–476.

Boix C (2003) *Democracy and Redistribution*. New York: Cambridge University Press.

Boix C (2013) RMDs. *Newsletter for the Comparative Democratization Section of the American Political Science Association* 11(3): 2, 12-15.

Brady HE and Collier D (eds) (2010) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Buthe T (2002) Taking temporality seriously: modeling history and the use of narratives as evidence. *American Political Science Review* 96(3): 481–493.

Clarke KA and Primo DM (2012) *A Model Discipline: Political Science and the Logic of Representations*. New York: Oxford University Press.

Collier D, Brady H and Seawright J (2004) Sources of leverage in causal inference: toward an alternative view of methodology. In: Brady He and Collier D (eds) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield, pp. 229–266.

Collier D and Mahoney J (1996) Insights and pitfalls: selection bias in qualitative research. *World Politics* 49(1): 56–91.

Dunning T (2008) *Crude Democracy: Natural Resource Wealth and Political Regimes*. New York: Cambridge University Press.

Dunning T (2010) Endogenous oil rents. *Comparative Political Studies* 43(3): 379–410.

Eckstein H (1975) Case study and theory in political science. In: Greenstein FJ and Polsby NW (eds) *Handbook of Political Science*. Vol. 7. Reading, MA: AddisonWesley, pp. 79–137.

Elster J (2000) Rational choice history: a case of excessive ambition. *American Political Science Review* 94(3): 685–695.

Fearon JD (1994) Domestic political audiences and the escalation of international disputes. *American Political Science Review* 88(3): 577–592.

Fearon JD (1997) Signaling foreign policy interests tying hands versus sinking costs. *Journal of Conflict Resolution* 41(1): 68–90.

Fearon JD (2013) 'Credibility' is not everything but it's not nothing either. Available at: http://themonkeycage.org/2013/09/07/credibility-is-not-everything-either-but-its-not-nothing-either/ (accessed 13 December 2013).

Fearon JD and Laitin DD (2008) integrating qualitative and quantitative methods. In: Collier D, Brady HE and Box-Steffensmeier JM (eds) *Oxford Handbook of Political Methodology*. New York: Oxford University Press, pp. 756–777.

Friedman M (1966 [1953]) The methodology of positive economics. In: Friedman M (ed.) *Essays in Positive Economics*. Chicago, IL: University of Chicago Press, pp. 3–43.

Gailmard S and Patty JW (2012) *Learning While Governing: Expertise and Accountability in the Executive Branch*. Chicago, IL: Chicago University Press.

Gartzke E and Lupu Y (2012) Still looking for audience costs. *Security Studies* 21(3): 391–397.

Geddes B (1990) How the cases you choose affect the answers you get: selection bias in comparative politics. *Political Analysis* 2(1): 131–150.

George AL and Bennett A (2005) *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gerring J (2007) Is there a (viable) crucial-case method? *Comparative Political Studies* 40(3): 231–253.

Gingerich D (2013) *Political Institutions and Party-Directed Corruption in South America: Stealing for the Team*. New York: Cambridge University Press.

Goemans H and Fey M (2009) Risky but rational: war as an institutionally induced gamble. *Journal of Politics* 71(1): 35–54.

Goemans H and Spaniel W (2016) Multi-method research: the case for formal theory. *Security Studies* 25(1): 25–33.

Goertz G and Mahoney J (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton, NJ: Princeton University Press.

Granato J, Lo M and Wong MC (2012) A framework for unifying formal and empirical analysis. *American Journal of Political Science* 54(3): 783–797.

Granato J and Scioli F (2004) Puzzles, proverbs, and omega matrices: the scientific and social significance of empirical implications of theoretical models (EITM). *Perspectives on Politics* 2: 313–323.

Green DP and Shapiro I (1994) *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven, CT: Yale University Press.

Greif A (2006) *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. New York: Cambridge University Press.

Haggard S and Kaufman RR (2012) Inequality and regime change: democratic transitions and the stability of democratic rule. *American Political Science Review* 106(3): 1–22.

Haggard S, Kaufman RR and Teo TK (2012) Distributive Conflict and Regime Change: A Qualitative Dataset. Coding document to accompany Haggard and Kaufman (2012).

Healy K (2016) Fuck nuance. *Sociological Theory*, forthcoming.

Hill K (2005) AJPS rejection without review for theory papers. Posted on *H-POLMETH* 24 March 2005. Accessed online 13 December 2013.

Johns L (2007) A servant of two masters: communication and the selection of international bureaucrats. *International Organization* 61(2): 245–275.

Kuehn D (2013) Combining game theory models and process tracing: potential and limits. *European Political Science* 12(1): 52–63.

King G and Powell EN (2008) How not to lie without statistics. Available at: http://gking.harvard.edu/files/gking/files/nolie.pdf (accessed 6 August 2016).

Kydd AH (2005) *Trust and Mistrust in International Relations*. Princeton, NJ: Princeton University Press.

Laitin DD (2002) Comparative politics: the state of the subdiscipline. In: Katznelson I and Milner HV (eds) *Political Science: State of the Discipline*. New York: W. W. Norton & Company, pp. 630–659.

Laitin DD (2003) The Perestroikan challenge to social science. *Politics & Society* 31(1): 163–184.

Levi M (2004) An analytic narrative approach to puzzles and problems. In: Shapiro I, Smith RM and Masoud TE (eds) *Problems and Methods in the Study of Politics*. New York: Cambridge University Press, pp. 201–226.

Lieberman ES (2005) Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review* 99(3): 435–452.

Lijphart A (1971) Comparative politics and the comparative method. *American Political Science Review* 65(3): 682–693.

Little AT and Pepinsky TB (2016) Simple and formal models in comparative politics. Available at http://www.andrewlittle.com/papers/cpformal_web.pdf (accessed 8 July 2016).

Lorentzen P (2013) Regularizing rioting: permitting public protest in an authoritarian regime. *Quarterly Journal of Political Science* 8(2): 127–158.

Lorentzen P (2014) China's strategic censorship. *American Journal of Political Science* 58(2): 402-14.

Lorentzen P and Scoggins S (2015) Understanding China's rising rights consciousness. *China Quarterly* 223: 638–657.

Mayhew DR (2000) Electoral realignments. *Annual Review of Political Science* 3: 449–474.

Miller G and Schofield N (2003) Activists and partisan realignment in the United States. *American Political Science Review* 97(2): 245–260.

Morton RB (1999) *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science*. New York: Cambridge University Press.

Nalepa M (2010) Captured commitments: an analytic narrative of transitions with transitional justice. *World Politics* 62(2): 341–380.

Paine J (2016) Rethinking the conflict resource curse: how oil wealth prevents center-seeking civil wars. *International Organization* 70(4): 727–761.

Powell R (1999) *In the Shadow of Power: States and Strategies in International Politics*. Princeton, NJ: Princeton University Press.

Rodrik D (2015) *Economics Rules: The Rights and Wrongs of the Dismal Science*. New York: W.W. Norton & Company.

Rohlfing I (2012) *Case Studies and Causal Inference: An Integrative Framework*. Basingstoke, UK: Palgrave Macmillan.

Schultz KA (2012) Why we needed audience costs and what we need now. *Security Studies* 21(3): 369–375.

Seawright J and Gerring J (2008) Case selection techniques in case study research. *Political Research Quarterly* 61(2): 294–308.

Shepsle KA (1995) Statistical political philosophy and positive political theory. *Critical Review* 9(1): 213–222.

Slantchev B (2012) Audience cost theory and its audiences. *Security Studies* 21(3): 376–382.

Slantchev BL (2011) *Military Threats: The Costs of Coercion and the Price of Peace*. New York: Cambridge University Press.

Slater D and Ziblatt D (2013) The enduring indispensability of the controlled comparison. *Comparative Political Studies* 46(10): 1301–1327.

Slater D, Smith B and Nair G (2014) Economic origins of democratic breakdown? The redistributive model and the postcolonial state. *Perspectives on Politics*, 12(2): 353–374.

Snyder J and Borghard ED (2011) The cost of empty threats: a penny, not a pound. *American Political Science Review* 105(3): 437–456.

Soifer HD (2013) State power and the economic origins of democracy. *Studies in Comparative International Development* 48(1): 1–22.

Svolik M (2009) Power sharing and leadership dynamics in authoritarian regimes. *American Journal of Political Science* 53(2): 477–494.

Svolik M (2015) Equilibrium analysis of political institutions. In: Gandhi J and Ruiz-Rufino R (eds) *Routledge Handbook of Comparative Political Institutions*. Abingdon-on-Thames, UK: Routledge.

Tarrow S (2010) The strategy of paired comparison: toward a theory of practice. *Comparative Political Studies* 43(2): 230–259.

Trachtenberg M (2012a) Audience costs: an historical analysis. *Security Studies* 21(1): 3–42.

Trachtenberg M (2012b) A comment on the comments. *Security Studies* 21(3): 405-415.

Weingast BR (1996) Off-the-path behavior: a game-theoretic approach to counterfactuals and its implications for political and historical analysis. In: Tetlock PE and Belkin A (eds) *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton, NJ: Princeton University Press.