A Dialogue with the Data:
The Methodological Foundations of Iterative Research

Tasha Fairfield (LSE)[1]
Dept. of International Development,
London School of Economics

Andrew Charman
Dept. of Physics,
University of California, Berkeley

Version 4.0

Abstract: This paper advances efforts to explicate and improve inference in qualitative research that iterates between theory development, data collection, and data analysis, rather than proceeding linearly from hypothesizing to testing. We draw on the school of Bayesian "probability as extended logic" from the physical sciences, where probabilities represent rational degrees of belief in propositions given limited information, to provide a solid foundation for iterative research that has been lacking in the qualitative methods literature. We argue that mechanisms for distinguishing exploratory from confirmatory stages of analysis that have been suggested in the context of APSA's transparency initiative are unnecessary for qualitative research that is guided by logical Bayesianism, because new evidence has no special status relative to old evidence for testing hypotheses within this inferential framework. Bayesian probability not only fits naturally with how we intuitively move back and forth between theory and data, but also provides a framework for rational reasoning that mitigates confirmation bias and ad-hoc hypothesizing—two common problems associated with iterative research. Moreover, logical Bayesianism facilitates scrutiny of findings by the academic community for signs of sloppy or motivated reasoning. We illustrate these points with an application to recent qualitative research on state building.

[1] Corresponding author, T.A.Fairfield@lse.ac.uk

1. Introduction

In the context of the replicability crisis, APSA's transparency initiative, and surrounding debates (www.dartstatement.org, www.qualtd.net), scholars have sought to revalue, explicate, and improve inference in qualitative research that proceeds in an inherently iterative manner, where prior knowledge informs hypotheses and data gathering strategies, evidence inspires new or refined hypotheses along the way, and there is continual feedback between theory and data.[2] This iterative style of research, which is common in process tracing and comparative historical analysis, diverges from prevailing norms that mandate clearly differentiating and sequencing theory-building (exploratory, inductive) and theory-testing (confirmatory, deductive) stages of research. Theory-testing requires new data that did not contribute to inspiring hypotheses, and any deviations from a specified research design should be reported (e.g., Humphreys et al. 2013:1, Monogan 2015). Furthermore, theory testing is generally granted higher status (Bowers et al. 2015:7, Lieberman 2016:1057, Jacobs 2017:14).[3]

Advocates of iterative qualitative research have suggested the key to enhancing its status and improving inference lies in finding ways to conform to the norms of differentiating exploration from confirmation and testing theory with new evidence. Scholars have called for greater transparency about analytical sequencing (Yom 2015:11, Büthe and Jacobs 2015:55) and advocate various mechanisms for keeping track of when a hypothesis was devised relative to specific stages of data collection, including pre-registration (Bowers et al. 2015, Jacobs 2017) or maintaining logs that time-stamp data as "used" or "unused" over the course of fieldwork and analysis (Kapiszewski et al. 2015b). Meanwhile, the recent joint-committee proposal for a political science registry from members of APSA's Political Methodology, Qualitative and Multi-Method Research, and Experimental Research sections asserts: "The basic analytical difference between induction and testing is as relevant to qualitative analysis as to quantitative. ...The clearest evaluation of explanatory or theoretical propositions derives from a new set of observations, independent of those that inspired the propositions in the first place," (Bowers et al. 2015:15). Similar suggestions have been raised during the Qualitative Transparency Deliberations.

This paper presents a different view of iterative research that is grounded in Bayesian probability. We draw on expositions of "probability as extended logic" from the physical sciences (Cox 1961, Jaynes 2003),[4] where probabilities represent rational degrees of belief in propositions given the inevitably limited information we possess. From a logical Bayesian perspective, prescriptions for separating theory-building from theory-testing draw on false dichotomies between old vs. new evidence and inductive vs. deductive reasoning. Theory testing—understood in Bayesian terms as inference to best explanation using probabilistic reasoning—takes *all* evidence into account, regardless of whether or not it was known to the investigator at the time hypotheses were devised; new evidence has no special status relative to old evidence. Scientific inference invariably entails a "dialogue with the data," where we go back and forth between theory development, data collection, and data analysis, rather than a linear sequence from hypothesizing to testing.

[2] Iterative research occurs not just in qualitative research, but throughout comparative politics and political science (Laitin 2013:44, Yom 2015, Kapiszewski et al. 2015:336, Büthe and Jacobs 2015:53, Lieberman 2016:1057).
[3] Lieberman notes an "unspoken presumption that the best work ought to be confirmatory or a test of an ex-ante specified hypothesis."
[4] This approach can also be found in machine learning (MacKay 2003), econometrics (Zeller), and other fields.

Our perspective highlights and aims to resolve an underlying tension in contemporary efforts to understand and improve qualitative research. On the one hand, much of the best such research implicitly and intuitively, albeit not consciously, approximates the logic of Bayesian reasoning.[5] On the other hand, proposals advocating crisp delineations between exploratory and confirmatory research are grounded in the frequentist inferential framework that underpins most large-N analysis—a framework that is inapplicable to small-N case-study analysis. Whereas separating theory-building from theory-testing is imperative within frequentism, it is unnecessary for Bayesian inference.

Accordingly, this paper aims to make two central contributions. First, we advance efforts to revalue iterative research by elucidating its Bayesian foundations and thereby providing a solid methodological basis that is currently lacking in the qualitative methods literature. Second, we explicate the safeguards Bayesianism provides against confirmation bias and *ad-hoc* hypothesizing, which make firewalls between theory building and theory testing unnecessary. We therefore argue that time-stamping and pre-registration (binding or non-binding) are not useful tools in qualitative research, regardless of the practical (in)feasibility of these approaches in particular research programs (e.g. analysis of existing historical data vs. generation of original data through expert interviews). We hope this paper will help inform ongoing discussion among multi-method and qualitative scholars on the nature of inference in case-study research, as well as the relative costs and *analytical* benefits of measures that have been suggested for improving research transparency, beyond advocating transparency for transparency's sake.

We begin by overviewing the trajectory of methodological thinking on iterative research and situating our contribution within recent work on Bayesian process tracing (§2). We then introduce the "logical" approach to Bayesian probability (§3). We clarify how this framework differs from the frequentist paradigm, and we elucidate fundamental tenets of logical Bayesianism that mitigate the need for distinctions between exploratory and confirmatory research. The key lies in recognizing that the terms "prior" and "posterior," as applied to our degree of belief in whether a proposition is true or false, are not temporal notions. Instead, they are purely logical concepts that refer to whether we have incorporated a given body of evidence into our analysis via Bayes' rule. Section 4 illustrates these points with an application to recent qualitative research on state building.

Section 5 considers potential concerns regarding our arguments that within a logical Bayesian framework, there is no need to keep track of what the investigator knew when and that "old" evidence is just as good as "new" evidence for assessing rival hypotheses. Our response emphasizes that Bayesian probability in and of itself provides a framework for rational reasoning in the face of uncertainty that simultaneously helps inoculate against cognitive biases and opens analysis to scrutiny by other scholars for signs of such pitfalls. While there are no magic bullets for ensuring and signaling honest and unbiased assessments of evidence in practice, drawing on Bayesian reasoning more consciously in qualitative research, discussing rival explanations more explicitly, and openly addressing observations that run counter to overall conclusions could help further those goals.

[5] While we recognize that a wide range of epistemological views are debated within qualitative methods, we follow Humphreys and Jacobs (2015:672), Bennett (2015:297), and Fairfield and Charman (2017) in espousing Bayesianism as the most appropriate logic of inference and contending that to the extent narrative-based qualitative research makes valid causal inferences, it implicitly follows Bayesian reasoning.

2. Perspectives on Iterative Research in Qualitative Social Science

Iterative research has a long tradition in social science. Classic methodological discussions include Glaser and Strauss (1967) and Ragin (1997), which emphasize jointly collecting and analyzing data while developing and refining theory and concepts. Yet these authors largely describe their goal as theory building, not theory testing or some combination thereof. Glaser and Straus (1967:103) for example remark that theory testing entails "more rigorous approaches" that "come later in the scientific enterprise."

Differentiating between theory building and testing remains prevalent even in qualitative methods literature that questions KKV's (1994) application of standards from large-N statistical inference to case studies. Ragin (1997:3) expressly criticizes KKV's (1994:22) assertion that "we should not make it [our theory] more restrictive without collecting new data to test the new version of the theory," but his response stops short of providing a methodological rationale; he simply notes the infeasibility of KKV's prescription: "When the number of relevant cases is limited by the historical record to a mere handful...it is simply not possible to collect a 'new sample' to 'test' each new theoretical clarification." Brady and Collier's (2010) groundbreaking volume stresses the contribution of inductive research to theory innovation and notes: "for qualitative researchers, the refinements of theory and hypotheses through the iterated analysis of a given set of data is an essential research tool," (Collier, Seawright & Munck 2010:62). But in emphasizing tradeoffs between different objectives, the volume leaves the dichotomy between theory development and theory testing largely intact.

Similarly, contemporary process-tracing literature retains language that discriminates between induction and deduction. Authors refer to inductive vs. deductive process tracing (Bennett and Checkel 2015:7-8, Schimmelfennig 2015:101), theory-building vs. theory-testing process tracing (Beach and Pedersen 2013), and similar variants (Mahoney 2015, Bowers et al. 2015:15). Even when acknowledging that process tracing in practice involves a complex combination of both theory construction and evaluation, these modes are still treated as analytically distinct (Mahoney 2015:201-02) and ideally sequential, where "inductive discovery is followed by deductive process tracing" using "evidence independent of that which gave rise to the theory," (Bennett and Checkel 2014:268).[6]

The relationship between theory building and theory testing is receiving renewed attention in the context of debates over transparency and the crisis of replicability. Yom's (2015:4) valuable contribution seeks to elevate the status of disciplined "inductive iteration" while highlighting "truly destructive" practices like "data mining, selective reporting, and ignoring conflicting results." Yet like previous authors, he does not articulate a clear methodological foundation for iterative research. Yom's (2015:11) emphasis on "transparency in practice," which calls for scholars to report when they "had to reconceptualize a causal mechanism as new information comes to light, ...tighten a theoretical argument in light of how rival explanations perform with the data, or rewrite a process-tracing narrative due to an initial misunderstanding," essentially falls back on the linear research template he critiques, in that the only rationale for requiring such information about the temporal trajectory of the intellectual process lies in standard prescriptions to test inductively-inspired theory with new evidence, otherwise we are promoting transparency purely for the sake of transparency. While we agree that scholars should be forthright when conducting iterative research, we will argue that there are few analytical benefits to reporting temporal details about how the research process unfolded.

[6] Van Evera (1997:45-6) offers a dissenting view.

In reevaluating the relationship between theory building and theory testing, we take inspiration not only from the physical sciences but also from early work on Bayesian underpinnings of case-study research. McKeown (1999) instigated a pioneering methodological agenda by observing that KKV's statistical world-view is at odds with a logic of "folk Bayesianism" that governs case study research:

> Researchers in the social sciences... are 'interactive processors.' They move back and forth between theory and data, rather than taking a single pass through the data. ...one can hardly make sense of such activity within the confines of a classical theory of [frequentist] statistics. A [Bayesian] theory of probability that treats it as a process involving the revision of prior beliefs is much more consistent with actual practice...

Subsequent scholarship makes important strides towards explicitly applying Bayesian reasoning in process tracing (Bennett 2015, Humphreys and Jacobs 2015, Fairfield and Charman 2017). Yet this research has not yet explored the implications of McKeown's central observation about moving back and forth between theory and data. Formal treatments of Bayesian process tracing have been cast in a deductive, theory-testing framing that emphasizes prospective anticipations about the evidence we might encounter, without elucidating the importance of inferential feedback and the role played by induction in conjunction with retrospective analysis of data actually obtained.

We build on McKeown's insights by arguing that logical Bayesianism provides a firm methodological foundation for iterative research. In the apt phrase of astrophysicist Stephen Gull, Bayesian analysis involves a "dialogue with the data" (quoted in Sivia 2006). We draw new insights through a continuous, iterative process of analyzing data differently and/or more deeply, revising and refining theory, asking new questions, and deciding what kinds of additional data to collect. Inference is always provisional, in that theories are rarely definitively refuted and never definitively confirmed—they are constantly amended in light of new ideas and new data. But in these inferential cycles we never "use up" or "throw away" previous information—Bayesianism mandates learning from accumulated knowledge by virtue of the fact that all probabilities must be *conditional* probabilities that take into account all known information relevant to the question of interest. Confidence in one proposition depends on what else we know and generally changes when we make new observations. There is no need within logical Bayesianism to temporally sequence inductive and deductive stages of reasoning. Bayes' rule allows us to move back and forth fluidly between reasoning about the empirical implications of hypotheses and drawing inferences about possible causes from observed effects, and Bayesian probability allows us to assess the weight of evidence whether it was collected before or after formulating hypotheses.

3. The Bayesian Logic of Iterative Research

We begin by reviewing conceptual distinctions between Bayesianism and frequentism, the dominant approach to quantitative inference which often informs how qualitative research is evaluated, and introducing the logical school of Bayesianism, which provides a prescription for rational reasoning given incomplete information (§3.1). We then briefly review the mathematical framework of Bayesian inference (§3.2). Section 3.3 resolves the false dichotomies of new vs. old evidence and deductive vs. inductive research by focusing on the logical—not temporal—nature of "prior" and "posterior" probabilities. Section 3.4 discusses safeguards built into logical Bayesianism that help curtail confirmation bias and *ad-hoc*

hypothesizing—two potential pitfalls often associated with iterative research that underpin conventional guidelines that theory building must be segregated from theory testing.

3.1 Bayesian Foundations

Frequentism conceptualizes probability as a limiting proportion in an infinite series of random trials or repeated experiments. For example, the probability that a die shows "2" on a given throw would be equated with the fraction of times it turns up "2" in an infinite sequence of throws. In this view, probability reflects a state of nature—e.g, a property of the die (fair or weighted) and the throwing process (random or rigged). In contrast, Bayesianism understands probability as a degree of belief. Two individuals observing the same die might rationally assign different probabilities to the proposition "the next throw will produce 2," based on whatever information they know about the die and throwing procedure.

The Bayesian notion of probability offers multiple advantages—most centrally: it is much closer to how people intuitively reason in the face of uncertainty; it can be applied to any proposition, including causal hypotheses, which would be nonsensical from a frequentist perspective; it is well-suited for explaining unique events, working with a small number of cases, and/or analyzing limited amounts of data; and inferences can be made using any relevant information, above and beyond data generated from stochastic processes. These features make Bayesianism especially appropriate for qualitative research, which evaluates competing explanations for complex sociopolitical phenomena using evidence that cannot naturally be conceived as random samples (e.g., information from expert informants, legislative records, archival sources). Strictly speaking, frequentist techniques are unsuitable for such data. In Jackman and Western's (1994:413) words: "*frequentist inference is inapplicable to the nonstochastic setting*."

The school of Bayesianism we advocate as the methodological foundation for scientific inference—logical Bayesianism—seeks to represent the rational degree of belief we should hold in propositions given the information we possess, independently of hopes, subjective opinion, or personal predilections. In ordinary logic, the truth-values of all propositions can be known with certainty. But in most real-world contexts, we have limited information, and we are always at least somewhat unsure about whether a proposition is true or false. Bayesian probability is an "extension of logic" (Jaynes 2003) in that it provides a prescription for how to reason when we have incomplete knowledge and are thus uncertain about the truth of a proposition. When our degrees of belief assume limiting values of zero (impossibility) or one (certainty), Bayesian probability automatically reduces to ordinary logic.

A central tenet of logical Bayesianism is that probabilities should encode knowledge in a unique, consistent manner. Incorporating information in different but logically equivalent ways (e.g. learning the same pieces of information in different orders) must produce identical probabilities, and individuals who possess the same information must assign the same probabilities. Cox (1961), Jaynes (2003), and subsequent scholars (e.g. Gregory 2005) show that if we represent our level of confidence in the truth of propositions with real numbers and impose these consistency requirements, we are led directly to the sum and product rules of probability, which in turn give rise to all other operations within Bayesian analysis for manipulating and updating probabilities.

The consistency requirements of logical Bayesianism are more demanding than requirements imposed in social-science approaches that draw on the "psychological" or "subjective" school of Bayesianism common in philosophy of science and conventional

Bayesian statistics textbooks. In this latter approach, rationality requires degrees of belief to follow the sum rule and product rule of probability, such that utility-maximizing gamblers decline "Dutch Book" bets (where loss is certain). But as long as probabilities satisfy these rules, they can be based on pure psychology—whatever happens to motivate an individual to hold some particular subjective degree of belief. Accordingly, within psychological Bayesianism, individuals possessing the same information need not assign identical probabilities.

We will show that the consistency requirements are the key to understanding the powerful methodological foundation that logical Bayesianism provides for iterative research. First, however, we review the mechanics of Bayesian inference.

3.2. Bayesian Inference in Brief

Bayesian inference generally proceeds by assigning "prior" probabilities to a set of plausible rival hypotheses using all relevant background information we possess. These prior probabilities represent our degree of confidence in the truth of each hypothesis taking into account salient knowledge accumulated from previous studies and/or experience. We then consider evidence obtained during the investigation at hand. The evidence includes all relevant observations (beyond our background information) that bear on the plausibility of our hypotheses. We ask how likely the evidence would be if a particular hypothesis were true, and we update our beliefs in light of that evidence using Bayes' rule to derive "posterior" probabilities on our hypotheses.

Formally, Bayes' rule is expressed in terms of conditional probabilities $P(A|B)$, representing the rational degree of belief in proposition A if we consider B to be true. Bayes' rule is a rearrangement of the product rule of probability:

$$P(AB)=P(BA)=P(A|B)\times P(B)=P(B|A)\times P(A). \quad (1)$$

For a hypothesis H, evidence E, and background information I, Bayes' rule states:

$$P(H|EI)=P(H|I)\times P(E|HI)/P(E|I), \quad (2)$$

where $P(H|EI)$ is the posterior probability on the hypothesis given the evidence and the background information, $P(H|I)$ is the prior probability on the hypothesis given our background information alone, $P(E|HI)$ is the *likelihood* of the evidence—the conditional probability of the evidence given the hypothesis and the background information—and $P(E|I)$ is the unconditional likelihood of the evidence (regardless of whether H is true).

Because causal inference always involves comparing hypotheses,[7] it is easier to work with the odds-ratio form of Bayes' rule:

$$\frac{P(H_i|EI)}{P(H_j|EI)}=\frac{P(H_i|I)}{P(H_j|I)}\times\frac{P(E|H_iI)}{P(E|H_jI)} \quad (3)$$

The factor on the left-hand side of equation (3) is the *posterior odds* on hypothesis H_i relative to H_j in light of the evidence. The posterior odds equals the *prior odds* (the first factor on the right-hand side) multiplied by the *likelihood ratio* (the second factor on the right-hand side).

Assessing the likelihood ratio, $P(E|H_iI)/P(E|H_jI)$, is the key inferential step that tells us whether the evidence should make us more or less confident in one hypothesis relative to another. The likelihood ratio can be thought of as the probability of observing evidence E in a

[7] In practice we cannot evaluate $P(E|I)$ in equation (2) unless I restricts our attention to a finite number of plausible rival hypotheses.

hypothetical world where H_i is true, relative to the probability of observing E in an alternative world where H_j is true (recall that in the notation of conditional probabilities, all *conditioning information* that appears to the right of the vertical bar is taken to be true when assessing degrees of belief). In qualitative research, we need to "mentally inhabit the world" of each hypothesis (Hunter 1984) and ask how surprising (low probability) or expected (high probability) the evidence would be in each respective world. If the evidence is less surprising in the "H_i world" relative to the "H_j world," then that evidence will increase the odds we place on H_i vs. H_j, and vice versa. We gain confidence in one hypothesis vs. another to the extent that it makes the evidence we find more plausible.

Elsewhere, we elaborate guidelines for formal Bayesian analysis in qualitative research, which entails quantifying all probabilities. To illustrate how Bayesian logic can be applied heuristically (without quantification), consider an example drawing on Kurtz's state-building research (2009). We wish to ascertain whether the resource-curse hypothesis, or the warfare hypothesis (assumed mutually exclusive), better explains institutional development in Peru:

> H_R =*Mineral resource abundance is the central factor hindering institutional development. Easy money from mineral exports precludes the need to collect taxes and creates incentives to spend public resources on inefficient subsidies and patronage networks, instead of investing in administrative capacity.*
>
> H_W =*Absence of warfare is the central factor hindering institutional development. Threat of conquest requires states to extract resources from society and develop strong administrative capacity in order to build and sustain armies. In the absence of external threats, state leaders lack these institution-building incentives.*

For simplicity, suppose we have no relevant background knowledge about state-building in Peru. Since both hypotheses find substantial support in literature on other countries, we might reasonably assign even prior odds. We now learn the following:

> E_1=*Peru was consistently threatened by international military conflict following independence, its economy has been dominated by mineral exports since colonial days, and it never developed an effective state.*

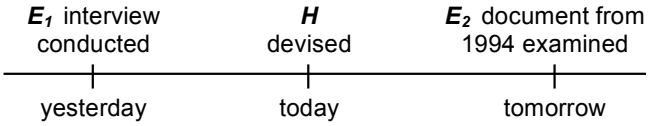
Intuitively, this evidence strongly favors the resource-curse hypothesis. Applying Bayesian reasoning, we must evaluate the likelihood ratio $P(E_1|H_RI)/P(E_1|H_WI)$. Imagining a world where H_R is the correct hypothesis, mineral dependence in conjunction with weak state capacity is exactly what we would expect. Furthermore, although H_R makes no direct predictions about presence or absence of warfare, external threats are not surprising given that a weak state with mineral resources could be an easy and attractive target. In the alternative world of H_W, however, the evidence would be quite surprising; something very unusual, and hence improbable, must have happened for Peru to end up with a weak state if the warfare hypothesis is nevertheless correct, because weak state capacity despite military threats contradicts the expectations of the theory. Because the evidence is much more probable under H_R relative to H_W, the odds in favor of H_R increase substantially, even though the evidence does not exhibit the resource-curse logic in action.

3.3. Prior vs. Posterior Probabilities and Old vs. New Evidence

While testing hypotheses with new evidence is pervasively espoused, distinctions between old vs. new evidence (relative to the formulation of hypotheses), and hence exploratory vs. confirmatory research, are far less consequential within logical Bayesianism. To be clear, "new

evidence" refers to information that was unknown to the scholar before the hypothesis was devised regardless of the historical timing of when that information was generated. For example, in Figure 1, E_1 is old evidence relative to H, whereas E_2 is new evidence, even though E_2 existed in the world before E_1.

Figure 1



The key to unraveling the false dichotomies lies in understanding that the terms *prior* and *posterior* are not temporal notions—they are *logical* notions. In the words of astrostatistician Tom Loredo (1990:87):

> There is nothing about the passage of time built into probability theory. Thus, our use of the terms... 'prior probability,' and 'posterior probability' do not refer to times before or after data is available. They refer to logical connections, not temporal ones. Thus, to be precise, a prior probability is the probability assigned before consideration of the data.

To reiterate these crucial points, the descriptions *prior* and *posterior* refer to degrees of belief before and after a piece of evidence is incorporated into our analysis—not to the timing of when we happened to learn or obtain that piece of evidence. *Prior* and *posterior* refer simply to idealized states of knowledge without and with specific pieces of evidence included. Of course, hypotheses can contain temporal structuring, and evidence can contain information about timing. However, probabilities themselves carry no intrinsic time-stamps.

These points merit expounding. Recall that within logical Bayesianism, only the data at hand and the background knowledge are relevant for assessing the reasonable degree of belief that is warranted in a hypothesis. Nothing else about our state of mind, hopes, or predilections should influence the probabilities we assign. The relative timing of when we stated the hypothesis, worked out its potential implications, and gathered data falls into this later category of logical irrelevance when assigning and updating probabilities.

To further stress the logical irrelevance of keeping track of what we knew when, notice that the rules of conditional probability mandate that we are free to incorporate evidence into our analysis in any order without affecting the final posterior probabilities we derive via probability theory. Using the product rule (1) and commutativity, the joint likelihood of two pieces of evidence can be written in any of the following equivalent ways:

$$P(E_1E_2|HI)=P(E_2E_1|HI)=P(E_1|E_2HI)\times P(E_2|HI)=P(E_2|E_1HI)\times P(E_1|HI). \quad (4)$$

Evidence learned at time one (E_1) may thus be treated as logically posterior to evidence learned at time two (E_2), if we choose to incorporate E_2 into our analysis before E_1. If in practice conclusions are found to differ depending on the order in which evidence was incorporated, there is an error in our reasoning somewhere that should be corrected. Otherwise we have violated the fundamental notion of rationality that lies at the heart of logical Bayesianism (§3.1)—information incorporated in equivalent ways should lead to the same conclusions.

Once we recognize that timing is not relevant in probability theory, it follows that each of the analytical steps below is a logically distinct endeavor:

- drawing on evidence E to inspire hypotheses;
- assigning prior probabilities to those hypotheses given background information I that does not include E;
- assessing the likelihood of evidence E under alternative hypotheses in order to derive posterior probabilities.

Information is neither "exhausted" nor "double-counted" in this inferential process. All relevant knowledge can be sorted as convenient into background information I on which all probabilities in Bayes' theorem are conditioned, and evidence E that we use to update probabilities.

Psychological/subjective approaches to Bayesianism often diverge from logical Bayesianism on these points, because the former focus on individuals' personal degrees of belief and how their psychological states evolve over time. Jeffrey's (1983) "probability kinematics" is a prominent example; his approach introduces non-standard rules for updating that violate the laws of probability (4) and hence imply that the order in which evidence is analyzed does matter.

Another salient example from psychological/subjective Bayesianism is the "problem of old evidence" in philosophy of science (e.g., Glymour 1980, Earman 1992). Glymour argued that if probabilities are evaluated at a time when the evidence E is known, then $P(E|I)=1$, which in turn directly implies that $P(E|HI)=1$. Substituting into Bayes' rule (2), he then find that $P(H|EI)=P(H|I)$, such that "old" evidence purportedly cannot alter our degree of belief in hypothesis H. From a logical Bayesian perspective, the flaw in this reasoning lies in confusing temporal relationships with logical ones. If we wish to evaluate probabilities in the light of knowing evidence E, then E must appear as conditioning information. In essence, Glymour can only assert that $P(E|EI)=1$, and his argument collapses, because Bayes' rule accordingly yields $P(H|EEI)=P(H|EI)\times P(E|HEI)/P(E|EI)=P(H|EI)$, which we already knew from $EE=E$. As astrophysicist Bill Jefferys (2007:7) notes, "what Glymour has actually proved is the (well-known) fact that...quite sensibly...[we] cannot validly manipulate the Bayesian machinery to get additional information out of information that has already been used." The crucial point is that when evaluating probabilities, the conditioning information does not include whatever is in our heads at a particular moment in time. Instead, we condition on propositions located to the right of the vertical bar, which are explicitly specified and assumed to be true.

In probability theory, we must keep track of what information has been incorporated into our analysis, not the time at which that information was acquired. The "problem of old evidence" is therefore a red herring. Time-stamps indicating when hypotheses were composed or when evidence was observed or incorporated are not relevant to the logic of scientific inference.[8]

3.4. Curtailing Confirmation Bias and Ad-hoc Theorizing

Legitimate concerns about objectivity, rigor, and transparency underlie prescriptions that theory building and theory testing should proceed sequentially, observable implications of hypotheses should be identified before gathering data, and hypotheses should be tested on new evidence. However, careful application of Bayesian logic can itself help guard against both confirmation bias and *ad-hoc* hypothesizing—two of the most salient pitfalls commonly associated with iterative research.

Among multiple variants of confirmation bias (Nickerson 1998, Klayman 1995), two common tendencies entail focusing too much on data that fits a particular hypothesis and/or

[8] In natural sciences, hypotheses often derive support from evidence known long before they were developed; for example, quantum mechanics was devised to explain known facts about blackbody radiation, atomic stability, and the photoelectric effect.

overlooking data that casts doubt on it, and focusing on only a single favored hypothesis while forgetting to consider whether data consistent with that hypothesis might be equally or more supportive of a rival hypothesis. A common recommendation for precluding such biases entails identifying observable implications of rivals as well as the main working hypothesis before gathering data.[9] However, this advice can be problematic for two reasons.

First, deducing observable implications beforehand may be infeasible, because any hypothesis may be compatible with a huge number of possible evidentiary findings—just with varying probabilities of occurrence. In the context of qualitative research on complex socio-political phenomena, there is essentially no limit to the different kinds of evidence we might encounter, and there is no way to exhaustively catalogue the infinite possibilities in advance.

Second, anticipating observable implications may foster even *greater* bias. If we have conducted the exercise of spelling out hypotheses to be considered and evidence expected under each, we are now *better situated* to seek out the sorts of evidence that will support our pet theory, compared to a situation where we collect evidence without necessarily anticipating what will support which hypothesis. This caveat is classic advice from Sherlock Holmes: "It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts," (*A Scandal in Bohemia*).

Risks of confirmation bias can be better controlled by conscientiously endeavoring to follow logical Bayesian reasoning. First and foremost, tendencies to seek evidence that supports a favored hypothesis, interpret evidence as overly favorable to that hypothesis, and underweight evidence that runs against that hypothesis are counteracted by following the prescription to condition probabilities on all relevant information available, without presuming anything beyond what is in fact known, or bringing mere opinions or desires into the evaluation. Furthermore, remembering that the key inferential step in Bayesian inference entails assessing likelihood *ratios* of the form $P(E|H_jI)/P(E|H_kI)$ precludes the pitfall of restricting attention to a single hypothesis—we must always ask whether a given explanation makes the evidence more or less likely *compared to* a rival explanation.

In contrast to confirmation bias, the complementary problem of *ad-hoc* hypothesizing involves over-tailoring an explanation to fit a particular, contingent set of observations. This danger underpins calls for distinguishing exploration from confirmation and testing hypotheses with new data. Within logical Bayesianism, however, an *ad-hoc* hypothesis that is too closely tailored to fit arbitrary details of the data incurs a low prior probability, which protects us from favoring it over a simpler hypothesis that adequately explains the data. If an explanation is *ad-hoc*, careful consideration should reveal that it is just one member of a large family of more or less equally *ad-hoc* hypotheses, characterized by multiple parameters or arbitrary choices that must be fine-tuned to the data. Each of these related hypotheses might explain a different set of contingent facts, yet none of them would seem any more credible than the others in the absence of the particular body of observations obtained. Even if the overall prior probability of the family of hypotheses $\{H_1 \text{ or } H_2 \text{ or} \ldots H_N\}$ is appreciable, this prior probability must be spread over all of the constituent possibilities, such that the prior for any particular H_i will be small.

Consider an example adapted from Jefferys (2003). A stranger at a party shuffles a deck of cards, and you draw the six of spades. We might reasonably hypothesize that this card was

[9] Re. process-tracing, see Bennett and Checkel (2015:18), Bennett and Elman (2006:460). Specifying observable implications deductively is widely advocated, often without any explicit link to avoiding confirmation bias, and sometimes with regard only to the working hypothesis (e.g. Schimmelfennig 2015:108, Beach and Pedersen 2014:20, Rohlfing 2012:187).

arbitrarily selected from a randomly-shuffled deck (H_R). A rival hypothesis proposes that the stranger is a professional magician relying on a trick deck that forces you to draw the six of spades ($H_{6\spadesuit}$). While the likelihood of selecting this particular card is 1/52 under H_R, it is far larger under $H_{6\spadesuit}$. However, $H_{6\spadesuit}$ must be penalized by a factor of 1/52 relative to H_R, because without observing your draw, there would be no reason to predict the six of spades as the magician's forced card. $H_{6\spadesuit}$ should be treated as one of 52 equally plausible related hypotheses whereby the magician forces some other card.[10] Accordingly, our single draw provides insufficient evidence to boost the credibility of $H_{6\spadesuit}$ above H_R.

Logical Bayesianism thus penalizes complex hypotheses if they do not provide enough additional explanatory power relative to simpler rivals, in line with Occam's razor and Einstein's dictum that things should be as simple as possible, but no simpler. In quantitative analysis, this task is accomplished via *Occam factors* that are automatically built into Bayesian probability (Jaynes 2003:601-07, MacKay 2003:343-356, Gregory 2005:45-50). Appendix A discusses Occam factors in more detail and illustrates how the penalty of 1/52 in our card-draw example emerges when we formally apply Bayesian analysis. In qualitative research that heuristically follows Bayesian reasoning, there are no universal prescriptions for assessing how *ad-hoc* a hypothesis is. However, one useful stratagem entails carefully scrutinizing a new hypothesis to evaluate how much additional complexity it introduces compared to rivals. If the hypothesis invokes many more causal factors or very specific conjunctions of causal factors, good practice would entail penalizing its prior probability relative to the rivals.

In sum, within logical Bayesianism, likelihood ratios help guard against confirmation bias, while priors help guard against *ad-hoc* hypothesizing. These safeguards are absent within frequentism, where hypothesis testing focuses on the probability of the data only under the null hypothesis, rather than relative likelihoods under rival hypotheses, and where the concept of probability applies only to data obtained through a stochastic sampling procedure, not to hypotheses. Frequentist inference therefore requires pre-specifying sampling and analysis procedures to avoid confirmation bias, and strictly separating data used in theory-building from data used for theory-testing to prevent *ad-hoc* hypothesizing, whereas such standards are unnecessary for Bayesian inference.

4. Iterative Research in Practice

We have argued that within logical Bayesianism, there is no need for firewalls between theory-building and theory-testing, and no need to rely exclusively on "new evidence" when testing hypotheses. All we must do is carefully assign prior probabilities in light of our background information, and carefully assess likelihood ratios for all relevant evidence under our rival hypotheses. This section illustrates how these points apply to qualitative research by extending the Peruvian state-building example introduced in §3.2. We emphasize that we make no claims about how Kurtz's research process unfolded. Instead, we draw on hypotheses and evidence from his published work to show how an iterative dialog with the data can give rise to inferences that are as valid as in a purely deductive approach, where all hypotheses were devised prior to data collection.

Suppose that after comparing the resource-curse and warfare hypotheses in light of E_1 (military threats, mineral abundance, and weak state), we learn the following additional information:

[10] We might additionally discount the magician hypothesis considering the chances of encountering a magician at the party.

> E_2=*Throughout the 1880s, agricultural production in Peru relied on an enormous semiservile labor force. When Chile invaded, Peruvian elites were far more concerned that peasants remain under control than they were with contributing to national defense. The mayor of Lima openly hoped for a prompt Chilean occupation for fear that subalterns might rebel. The agrarian upper class not only refused to support General Cáceres' efforts to fight back, but actively collaborated with the Chilean occupiers because of Cáceres' reliance on armed peasant guerillas. (Kurtz 2009:496)*

This evidence might inspire a new hypothesis:

> H_{LRA}=*Labor-repressive agriculture is the central factor hindering institutional development. Elites resist taxation and efforts to centralize authority, especially control over coercive institutions, because they anticipate that national leaders may be unable or unwilling to keep their rebellion-prone local labor forces under control. (Kurtz 2009:485)*

To assess which hypothesis better explains the evidence acquired thus far, we must go back and reassign prior probabilities across the new hypothesis set: H_R, H_W, and the inductively-inspired H_{LRA}. We must then assess likelihood ratios for the aggregate evidence E_1E_2.

For priors, strictly speaking we should assess the plausibility of each hypothesis taking into account all information accumulated in previous state-building literature. However, systematically incorporating all of our background information when assessing priors is infeasible in social science. Given practical limitations, one reasonable approach is to keep equal odds on H_R vs. H_W but give H_{LRA} a moderate penalty relative to each rival, thereby acknowledging the novelty of this hypothesis with respect to existing research on state-building and anticipating skepticism among readers. Another reasonable option entails equal odds on all three hypotheses, considering that H_{LRA} is grounded in a longstanding research tradition originated by Barrington Moore. As Kurtz (2009:485) documents, while H_{LRA} is not discussed in state-building literature, labor-repressive agriculture has been identified as a crucial factor affecting other macro-political outcomes including regime type, so *a-priori* we might expect this factor to be salient for state-building as well. Furthermore, although H_{LRA} was introduced *post-hoc* (in light of E_2), it is no more or less *ad-hoc* compared to the rivals—on inspection, none of the three hypotheses seems appreciably more complex than the others (Figure 2). Each identifies a single structural cause that operates by shaping the incentives of key actors.

Figure 2

H_R = *Mineral resource abundance is the central factor hindering institutional development. Easy money from the mineral sector undermines administrative capacity by precluding the need to collect taxes, and public resources are directed toward inefficient subsidies and patronage networks that sustain elites in power.*

H_W = *Absence of warfare is the central factor hindering institutional development. Threat of military annihilation requires states to extract resources from society and develop strong administrative capacity in order to build and sustain armies. In the absence of external threats, state leaders lack these institution-building incentives.*

H_{LRA} = *Labor-repressive agriculture is the central factor hindering institutional development. Elites resist taxation and efforts to centralize authority, especially control over coercive institutions, because they anticipate that national leaders may be unable or unwilling to keep their rebellion-prone local labor forces under control.*

Turning to the evidence, the easiest way to proceed entails assessing likelihood ratios for H_{LRA} vs. H_R and for H_{LRA} vs. H_W.[11] Since the overall likelihood ratio can be decomposed as:

$$\frac{P(E_1E_2|H_iI)}{P(E_1E_2|H_jI)} = \frac{P(E_1|H_iI)}{P(E_1|H_jI)} \times \frac{P(E_2|E_1H_iI)}{P(E_2|E_1H_jI)} \quad (5)$$

we first consider E_1 and then E_2.

Evidence E_1 moderately favors H_R over H_{LRA}. As explained in §3.2, E_1 fits quite well with the resource-curse hypothesis. However, E_1 is not surprising under H_{LRA}; a weak state with mineral resources would still be an easy and attractive target for invasion if labor-repressive agriculture were the true cause of state weakness. Nevertheless, the presence of resource wealth in conjunction with state weakness makes E_1 more expected under H_R. In contrast, E_1 strongly favors H_{LRA} over H_W. We have argued that this evidence is unsurprising under H_{LRA}, but as previously discussed, it is highly unlikely under H_W.

E_2 very strongly favors H_{LRA} over each alternative. Neither H_W nor H_R makes predictions about the nature of agricultural labor, but under either of these hypotheses, the behavior of Peruvian elites described in E_2 would be highly surprising—we would instead expect them to resist the Chilean incursion (however ineffectively, given state weakness). In contrast, E_2 fits quite well with H_{LRA} in showing that concern over maintaining subjugation of the labor force trumped concern with national sovereignty and statehood. Of course, we know E_2 fits well with H_{LRA} since the former inspired the latter; however, the critical inferential point is that E_2 is much *more* plausible under H_{LRA} *relative to* the alternatives. Accordingly, this evidence very strongly increases the odds in favor of H_{LRA} vs. each rival.

Overall, the likelihood ratio (5) strongly favors H_{LRA} over the rivals. E_2 overwhelms the moderate support that E_1 provides for H_R. And all of the evidence weighs strongly against H_W. Accordingly, H_{LRA} emerges as the best explanation given the evidence acquired thus far. If we begin with a moderate penalty on H_{LRA}, the posterior still favors that hypothesis, although the higher the prior penalty, the more decisive the overall evidence needed to boost the plausibility of H_{LRA} above its competitors.

In essence, we have now "tested" an inductively-inspired hypothesis with "old evidence." What matters is not when H_{LRA} came to mind or which evidence was known before vs. after that moment of inspiration, but simply which hypothesis is most plausible given our background information and all of the evidence. Imagine that a colleague is familiar with all three hypotheses from the outset but has not seen E_1E_2. This colleague would follow a logically identical inferential process in evaluating which hypothesis provides the best explanation for the Peruvian case: assessing the likelihood of E_1E_2 under these rival hypotheses. It would be irrational for two researchers with the same knowledge to reach different conclusions merely because of when they learned the evidence.

To further emphasize the irrelevance of relative timing, we do not know from reading Kurtz (2009) whether he invented H_{LRA} before or after finding E_2, but that chronological information would not make E_2 any more or less cogent. Our goal is not to reproduce the order in which the neurons fired inside the author's brain; our goal is to independently assess which hypothesis is most plausible in light of the evidence presented.

Of course "new evidence" is often valuable for improving inferences by providing additional weight of evidence. In this example, readers probably would not be satisfied if Kurtz's analysis ended with E_2. However, the goal of obtaining new evidence is not to supplant

[11] $P(E_1E_2|H_RI)/P(E_1E_2|H_WI)$ is determined by the other two ratios.

existing evidence that inspired the hypothesis, but rather to supplement that evidence and ideally strengthen our inference. Information is never intentionally disregarded in logical Bayesianism; any new, differentiated stage of research following the inductive inspiration of a hypothesis must take all previously-obtained evidence into account through the prior probability on that hypothesis. In our example, E_2, which inspired H_{LRA}, contributes to the strong posterior odds in favor of H_{LRA}, which would in turn become the "prior odds" when analyzing additional evidence.[12]

5. Addressing Anticipated Concerns

We recognize that logical Bayesianism is a mathematical ideal that usually cannot be fully realized in practice without approximations. In qualitative social science, some degree of subjectivity must inevitably enter when assigning probabilities. There is no mechanical procedure for objectively translating complex, narrative-based, qualitative information into precise probability statements. We may still commit analytical errors despite conscientious efforts to follow Bayesian reasoning.

Accordingly, this section considers potential concerns with our argument that qualitative research need not demarcate theory-building vs. theory-testing. Our overarching response draws on the premise that research is not only a dialog with the data, but also a dialog with a community of scholars. Knowing the trajectory of authors' thought processes should not matter to how readers scrutinize inferences. If scholars disagree with an author's conclusions, logical Bayesianism provides a clear framework for pinpointing the locus of contention, which may lie in different priors and background information, and/or different interpretations of particular pieces of evidence. The Bayesian framework itself, whether applied explicitly or heuristically, thereby lays analysis open for all to scrutinize on its own terms. In contrast, it would be misguided to assume that if authors time-stamp hypotheses and evidence, their analysis is sound, whereas if such information is not reported, their inferences lack credibility. Regardless of whether temporal details about how the research process unfolded are provided, scholars must scrutinize hypotheses and evidence with their own independent brainpower.

Our discussion below includes guidelines for facilitating scholarly dialog and improving inferences within a Bayesian framework, while highlighting shortcomings of prescriptions for labeling and/or separating exploratory/inductive vs. confirmatory/deductive research stages. We address anticipated concerns regarding biased priors, biased likelihoods, and scholarly integrity in turn.

5.1. Biased priors

Concerns:
(a) Given psychological difficulties in "getting something out of our mind," we may be unable to assign priors that are not influenced by what we already know about our data.
(b) Given vulnerabilities to cognitive biases, we may over-fit inductively-devised hypotheses to the evidence without adequately penalizing their priors.

[12] Two further points merit emphasis. First, Bayesian "test strength" is simply a function of the extent to which the evidence fits better with one hypothesis relative to rivals. Second, the goal in this example is to explain a single case. Whether H_{LRA} better explains Peru than the rivals says nothing about how well this theory holds beyond Peru—assessing generalizability requires examining evidence from other cases.

Pre-specifying priors is not a sensible solution to these concerns. We cannot assess a prior before devising the hypothesis, and once we formulate the hypothesis, all relevant information—both background knowledge and evidence E_{prev} acquired thus far—must inform $P(H|E_{prev}I)$, which serves as the "prior" moving forward. Moreover, whether we evaluate $P(H\backslash I)$ and then the likelihood for the total evidence $E_T=E_{prev}E_{post}$ ultimately collected, $P(E_T|I)$, or whether we update along the way, evaluating $P(H|E_{prev}I)$ and then $P(E_{post}|E_{prev}HI)\times P(H\backslash E_{prev}I)$, the final inference must be the same—consistency checks can be conducted to ensure equivalence. The timing of when we assess or record priors is therefore irrelevant.

To guard against subconsciously-biased priors (concern (a)), best practices should include the following. First and foremost, describe the most salient background information and explain why it motivates a particular choice of priors. If priors are obviously biased in favor of an inductively-derived hypothesis, beyond what is justified by the background information discussed, readers should notice the discrepancy. For instance, in our state-building example, readers might balk if our prior odds strongly favored H_{LRA} over the well-established resource-curse and warfare hypotheses. Likewise, if a well-known study or salient literature is overlooked, readers will request reconsideration of priors in light of that further background information.

Second, consider conducting the analysis with equal prior odds, which avoids biasing the initial assessment in favor of any hypothesis. This approach shifts the focus to likelihood ratios, with the aspiration that even if scholars disagree about priors—which will be almost inevitable given that everyone has different background information—we may still concur on the direction in which our odds on the hypotheses should shift in light of the evidence. Third, consider using several different priors to assess how sensitive conclusions are to these initial choices along the lines of our analysis in §4.

For qualitative research that follows Bayesian logic heuristically, the first guideline entails carefully discussing the strengths and weaknesses of rival explanations based on existing literature, which is common practice. The second guideline entails recognizing that readers may initially view a hypothesis with much more skepticism than the author, such that all parties in the scholarly dialog should pay close attention to scrutinizing the evidence and the inferential weight it provides in favor of the author's explanation relative to rivals. Accordingly, authors should be conservative with their inferential claims until the weight of evidence becomes strong.

Regarding concern (b), scholarly dialog can again serve as a corrective to sloppy analysis. If an inductive hypothesis manifesting multiple fine-tuned variables or inordinate complexity is granted too much initial credence, readers should notice and demand additional evidence to overcome an unacknowledged or underestimated Occam penalty (§3.4).

Beyond the simple advice to treat inductively-devised hypotheses with a healthy measure of skepticism, the following suggestions can help curtail *ad-hoc* hypothesizing: start with reasonably simple theories and add complexity incrementally as needed; critically assess whether all casual factors in the theory actually improve explanatory leverage; and ask whether the explanation might apply more broadly.

It is important to emphasize that transparency in reporting the temporal sequencing of the research process in and of itself is not useful for ascertaining how severe an Occam penalty a hypothesis should suffer on its prior. The critical point is that a hypothesis that is *post-hoc*—devised after the evidence—is not necessarily *ad-hoc*—arbitrary or overly complex. These are distinct concepts. As argued in §4, H_{LRA} is *post-hoc* (relative to key evidence E_2), but not *ad-*

hoc, because it is no more arbitrary or complex than its rivals. In contrast, the following illustrates an *ad-hoc* hypothesis that is clearly over-tailored to case evidence:
$H_{ad\text{-}hoc}$*=The conjunction of three factors hinders institutional development: (1) cultural-linguistic affinity between Peruvian and Chilean elites, (2) attempted peasant uprisings within a nine-month period preceding invasion, and (3) distrust on the part of domestic elites in generals' commitment to upholding the social order and their ability to maintain discipline over soldiers.*

5.2. Biased likelihoods

Concern: We may succumb to confirmation bias in overstating how strongly evidence favors an inductively-derived hypothesis.

Recent suggestions for pre-registration and time-stamping in qualitative research (Bowers et al. 2015, Kapiszewski et. al. 2015b, Jacbos 2017) aim to address these concerns, on the premise that differentiating exploratory from confirmatory analysis allows us to more credibly evaluate inductively-inspired hypotheses. Importing this prescription into a Bayesian framework would entail assigning likelihoods in advance to clues we might encounter before we continue gathering data.

Even in light of human cognitive limitations we find this approach unhelpful. Although a scholar's prospective assessment of likelihoods for "new evidence" might be less prone to confirmation bias than retrospective analysis of "old evidence," confirmation bias could just as easily intrude when gathering additional evidence—by subconsciously looking harder for clues that favor the working hypothesis and/or overlooking those that do not (§3.4).

Moreover, we reiterate the impossibility of foreseeing all potential evidentiary observations in the complex world of qualitative social science. It is essential to recognize that anticipating course-grained categories of observations is not adequate for specifying likelihoods for any actual, concrete evidence that might fit within that class, because specific details of evidence obtained can matter immensely to likelihoods under different hypotheses. To illustrate the problem, consider the example Bowers et al. (2015:16-17)[13] present in their discussion of pre-analysis plans for qualitative research: a government has cut taxes, and we wish to assess hypothesis H_K*=The tax cuts were motivated by an interest in Keynesian demand management*. Bowers et al. delineate evidence *E=Records of deliberations among cabinet officials about the tax cut show "prominent mention of the logic of Keynesian stimulus,"* and they judge the probability of finding such evidence if H_K is true to be very high. Although the suggestion that we should consult meeting records and look for discussion of Keynesian ideas is sound advice, *E* as stated above is too vague to assign a likelihood in advance. Here are two different pieces of evidence we might encounter in the records:

> *E′ =The Finance Minister invokes Keynesian demand stimulus when explaining the proposed tax cut and its rationale to other cabinet members present in the meeting.*
>
> *E″ =One of the cabinet members in the meeting notes that the tax cut is consistent with Keynesian demand stimulus, whereafter discussion is interrupted by laughter and derisive jokes about Keynesian economics.*

Suppose further that the amount of time and attention devoted to these mentions of Keynesian stimulus are similar for *E′* and *E″,* such that both qualify as instances of *E* as articulated above, even though they carry very different import. Whereas the likelihood of *E′* might well be high if

[13] See also Jacobs (2017:29).

H_K is true, the likelihood of E'' certainly is not high—E'' would be extremely surprising in a world where the Keynesian-stimulus hypothesis is correct.

Bowers et al. (2015:16-17) recognize this "problem of precision," noting: "of course [E as defined above] still leaves some things open. Just how prominent do mentions of Keynesian logic have to be...? How many actors have to mention it? What forms of words will count as the use of Keynesian logic?" In our view, however, they underestimate the problem. As our example demonstrates, the issue is not just how many mentions or how many actors or what terms we associate with Keynesianism, but an endless array of other possibilities and nuances that depend on the context and manner in which Keynesian logic is discussed. However much additional detail we aim to specify before gathering data, we can always invent—and the real world may well produce—another twist or tweak that matters nontrivially. And despite efforts to anticipate what might surprise us ahead of time, science advances most when evidence surprises us in unforeseen ways.

Pre-registration advocates respond that despite the problem of precision, a pre-analysis plan is still useful because it "allows the reader to compare the researcher's interpretation of unexpected observations to the pre-announced tests and to arrive at her own judgment about the extent to [which] the interpretation of the evidence is consistent with the analysis plan's broad logic," (Jacobs 2017:29). Yet this assertion implies that scrutinizing findings depends on knowing what was in the scholar's head at given time. As we have argued, such psychological and chronological information is logically irrelevant for inference.

Jaynes (2003:421) reinforces these key points: "The orthodox line of thought [holds] that before seeing the data one will plan in advance for every possible contingency and list the decision to be made after getting every conceivable data set. The problem...is that the number of such data sets is usually astronomical; no worker has the computing facilities needed… We take exactly the opposite view: it is only by delaying a decision until we know the actual data that it is possible to deal with complex problems at all. The defensible inferences are the post-data inferences." What matters is not what scholars anticipated they might find, but rather what they did find. Likewise, we care about how sound the inferences are in light of the arguments and evidence presented, not in comparison to every twist and turn of analysis before the author arrived at the final conclusions, or what the author would have thought had the data turned out differently.

Returning to the core concern of mitigating bias when assessing likelihoods, first, recall that inference always requires assessing likelihood *ratios,* which keeps us from forgetting to ask how well the evidence fits with rival explanations. Second, we reiterate our central point regarding scholarly scrutiny. If despite efforts to follow logical Bayesian prescriptions, a scholar nevertheless over-estimates how much the likelihood ratio favors an inductively-inspired hypothesis, readers can independently weigh the evidence and critically assess the author's judgments. Subsequent debate may encourage the author to bring more background information to light that was previously used implicitly, or acknowledge that the evidence is not as strong as previously maintained. In our state-building example, a reader might contest our assessment that E_1E_2 very strongly favors H_{LRA} over H_R, perhaps suggesting that this evidence only moderately favors the inductively-inspired hypothesis. Open discussion would then result in greater consensus or at least greater clarity on why scholars interpret the evidence differently.

5.3. Scholarly integrity

Concern: We need mechanisms to discourage scholars from deliberately choosing procedures after the fact to get the results they want, or manipulating evidence to strengthen results.

The first malpractice—*post-hoc* choice of analytical procedures—is a bigger concern for frequentist inferential techniques, which require predefined stochastic data-generation models. Within a Bayesian framework for case-study research, we must make judgments about which hypotheses to consider, where and how to acquire evidence, and how to interpret that evidence. However, the underlying inferential procedure remains the same: apply probabilistic reasoning to update beliefs regarding the plausibility of rival hypotheses in light of relevant evidence. The analysis always involves assessing priors, assessing likelihoods, and updating probabilities in accord with Bayes' rule. Unlike frequentist statistical analysis, there is no need to choose among sampling procedures, stopping rules, estimators, tests statistics, or significance levels.

The second type of malpractice can certainly occur in qualitative research. Consider cherry-picking, where scholars "selectively pluck supportive quotations, statements, and other data out of context to maintain the fiction of complete corroboration," (Yom 2015:22). However, time-stamping does little to deter such abuses. Any scholar intent on exaggerating results or willing to commit fraud can find ways to do so regardless. Ansell and Samuels (2016:1810) make similar observations regarding the related issue of results-blind review; they note that it is always possible to "sweep dirt an author wants no one to see under a different corner of the publishing carpet." As a device for signaling integrity, mechanisms like pre-registration or time-stamping risk imposing a substantial burden of time and effort on honest scholars without preventing dishonest scholars from sending the same credibility signals.

The only viable strategy in our view involves disciplinary norms. First, the profession must instill a commitment to truth-seeking and scientific integrity. As Van Evera (1997:46) observed long before APSA's transparency initiative, "Infusing social science professionals with high standards of honesty is the best solution." Second, adjusting publication norms regarding requisite levels of confidence in findings would mitigate incentives for falsely bolstering results.[14] For qualitative research, embracing Bennett and Checkel's (2015:30) dictum that "conclusive process tracing is good, but not all good process tracing is conclusive" would be a major step in the right direction for reducing temptations to overstate the case in favor of a given hypothesis. An associated best practice could entail explicitly addressing the pieces of evidence that on their own run most counter to the overall inference; transparency of this type could both encourage critical thinking and signal integrity in a more meaningful way.

6. Conclusion

We share the transparency movement's goals of improving the reliability and quality of inference. We recognize that some research programs might benefit from pre-registration and time-stamping, or related practices such as data-blinding, and theories should certainly be subject to ongoing re-evaluation based on additional evidence.

However, we are skeptical of imposing constraints that often clash with the underlying logic and nonlinear practice of scientific reasoning. We have argued that standards such as pre-registration and time-stamping are neither necessary for nor suited to iterative qualitative case research that follows by the principles of Bayesian inference. From a logical

[14] Avoiding publication bias towards unexpected or counterintuitive findings is also advisable.

Bayesian perspective, such constraints are based on false dichotomies between theory building vs. theory testing and old vs. new evidence. Science invariably involves a dialogue with the data. Progress is nonlinear, iterative, serendipitous, and provisional. Scholars must interrogate data from different angles, re-think assumptions, and consider new hypotheses.

We have argued that within logical Bayesianism, devising hypotheses, assigning prior probabilities, and deriving posterior probabilities on hypotheses in light of our evidence and salient background information are all logically distinct steps, where temporal ordering is irrelevant. Testing hypotheses with evidence already used to develop the theory simply requires following the rules of probability and striving to assign degrees of belief that are based on the information we possess, independently of subjective hopes, intensions, and desires—exactly the same critical thinking necessary for assessing new evidence. Once analysis is completed, what matters is whether experts agree that priors are justified and likelihood ratios well reasoned. Details about what was know when and how research evolved over the course of fieldwork and analysis are logically irrelevant.

Applying Bayesian reasoning in qualitative research remains a methodological frontier. As this program advances, we envision training in logical Bayesianism as a good way to leverage intuition and improve inference, without needing to formally apply the full mathematical apparatus in qualitative research. Although some subjectivity and approximation will inevitably intrude in real-world applications, logical Bayesianism in itself is a prescription for systematic, rational reasoning. This inferential framework counteracts cognitive biases—confirmation bias when collecting or assessing new evidence and *ad hoc* hypothesizing when analyzing old evidence—and helps scholars scrutinize analysis for signs of sloppy or motivated reasoning, rather than making presumptions based on accidents of timing.

References

Ansell, Ben, and David Samuels. 2016. "Journal Editors and 'Results-Free' Research: A Cautionary Note." *Comparative Political Studies* 49(13):1809-1815.

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

Bennett, Andrew, and Jeffrey Checkel, eds. 2015. *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool*. New York: Cambridge University Press.

Bennett, Andrew, and Colin Elman. 2006. "Qualitative Research: Recent Developments in Case Study Methods." *Annual Review of Political Science* 9:455-76.

Bowers, Jake, Jonathan Nagler, John Gerring, Alan Jacobs, Don Green, and Macartan Humphreys. 2015. "A Proposal for a Political Science Registry." http://blogs.bu.edu/jgerring/files/2015/09/AproposalforaPoliticalScienceRegistry.pdf

Büthe, Tim, and Alan Jacobs. 2015. "Conclusion: Research Transparency for a Diverse Discipline." *Qualitative and Multimethod Research: Newsletter of the American Political Science Association's QMMR Section* 13(1):50-63.

Brady, Henry, and David Collier. 2010. *Rethinking Social Inquiry.* Lanham: Rowman and Littlefield.

Collier, David, Jason Seawright and Gerardo Munck. 2010. "The Quest for Standards." In Brady and Collier, eds. *Rethinking Social Inquiry.* Rowman and Littlefield: 33-63.

Cox, Richard. 1961. *The Algebra of Probable Inference.* Johns Hopkins University Press.

Earman, John. 1992. *Bayes or Bust?* Cambridge: MIT Press.

Fairfield, Tasha, and Andrew Charman. 2017. "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats." *Political Analysis.*

Glaser, Barney, and Anselm Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine Publishers.

Glymour, C. 1980. *Theory and Evidence.* Princeton University Press.

Gregory, Phil. 2005. *Bayesian Logical Data Analysis for the Physical Science.* Cambridge.

Humphreys, Macartan, and Alan Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109(4):653-73.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21:1-20.

Hunter, Douglas. 1984. *Political/Military Applications of Bayesian Analysis*. Boulder: Westview.

Jackman, Simon, and Bruce Western. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88(2):412-423.

Jacobs, Alan. 2017. "Of Bias and Blind Selection: Pre-registration and Results-Free Review in Observational and Qualitative Research." Draft chapter.

Jaynes, E.T. 2003. *Probability Theory: The Logic of Science.* Cambridge University Press.

Jeffrey, Richard. 1983. *The Logic of Decision.* University of Chicago Press.

Jefferys, William. Book review: "Bayes' Theorem," *Journal of Scientific Exploration* 17(3:537-42).

_____. 2007. "Bayesians *Can* Learn from Old Data." https://repositories.lib.utexas.edu/bitstream/handle/2152/29425/BayesiansOldData.pdf?sequence=1

Kapiszewski, Diana, Lauren M. MacLean, and Benjamin L. Read. 2015a. *Field Research in Political Science.* Cambridge University Press.

_____. 2015b. "Reconceptualizing Field Research." Unpublished Manuscript.

King, Gary, Robert Keohane, and Sidney Verba (KKV). 1994. *Designing Social Inquiry.* Princeton University Press.

Klayman, Joshua. 1995. "Varieties of Confirmation Bias." *Psychology of Learning and Motivation* 32:385-418.

Kurtz, Marcus. 2009. "The Social Foundations of Institutional Order: Reconsidering War and the 'Resource Curse' in Third World State Building." *Politics & Society* 37(4):479–520.

Laitin, David. 2013. "Fisheries Management." *Political Analysis* 21:42-27.

Lieberman, Evan. 2016. "Can the Biomedical Research Cycle be a Model for Political Science." *Perspectives on Politics* 14(4):1055-66.

Loredo, T.J. 1990. "From Laplace to Supernova SN1987A: Bayesian Inference in Astrophysics." In P.F. Fougere, ed., *Maximum Entropy and Bayesian Methods,* The Netherlands: Kluwer Academic Publishers.

MacKay, David. 2003. *Information Theory, Inference, and Linear Algorithms.* Cambridge.

Mahoney, James. 2015. "Process Tracing and Historical Explanation." *Security Studies* 24:200-218.

McKeown, Timothy. 1999. "Case Studies and the Statistical Worldview." *International Organization* 53(1):161-190.

Monogan, James. 2015. "Research Preregistration in Political Science." *PS: Political Science and Politics* 48(3)425-29.

Nickerson, Raymond. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2(2):175-220.

Ragin, Charles. 1997. "Turning the Tables: How Case-Oriented Research Challenges Variable-Oriented Research." *Comparative Social Research* 16:27–42.

Rohlfing, Ingo. 2012. *Case Studies and Causal Inference.* Palgrave Macmillan.

Schimmelfennig, Frank. 2015. "Efficient Process Tracing," in Andrew Bennett and Jeffrey Checkel, eds,. *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool*. New York: Cambridge University Press.

Sivia, D.S., 2006, "Data Analysis—A Dialogue With The Data," in Advanced Mathematical and Computational Tools in Metrology VII, P. Ciarlini, E. Filipe, A.B. Forbes, F. Pavese, C. Perruchet and B. Siebert (eds.), World Scientific Publishing Co.:108-118.

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

Yom, Sean. 2015. "From Methodology to Practice: Inductive Iteration in Comparative Research." *Comparative Political Studies* 48(5):616-644.

Appendix A. *Ad-Hoc* Hypotheses and Occam Factors

Section 3.4 introduced the logical Bayesian concept of Occam factors, which penalize hypotheses that over-fit the data. This appendix discusses Occam factors in more detail and provides two examples to show how they can arise in qualitative research.

To appreciate the importance of Occam factors, it is worth stressing that over-fitting can be a major problem within a frequentist framework that does not allow prior probabilities on hypotheses or fixed parameters. When working with quantitative datasets, analytical models can be made arbitrarily complex with a multitude of adjustable parameters that end up fitting not just the signal of interest, but the noise as well. Detecting over-fitting can be particularly challenging in orthodox statistics, because adding extra parameters can always improve the likelihood of the data under the model.

Within logical Bayesianism, however, an *ad-hoc* hypothesis that is too closely tailored to fit the arbitrary details of the data incurs a low prior probability via Occam factors that arise automatically from correctly applying probability theory. These Occam factors keep us from favoring an overly complex hypothesis compared to a simpler hypothesis that adequately explains the data.

Recall that generally speaking, an *ad-hoc* hypothesis is properly regarded as one member of a family of hypotheses characterized by multiple parameters that take on different, but equally arbitrary values. To restate this point in slightly different terms, an *ad-hoc* hypothesis emerges from a model with multiple parameters that *a priori* could have taken on a large range of different values. As a model becomes more complex, its prior probability becomes spread out over a larger parameter space, and the posterior odds are reduced to the extent that this parameter space must be fine-tuned to fit the observed data. Similarly, whenever we include another parameter in the model and find that the range of values it must assume to account for the data is much narrower than the prior range of values deemed feasible given the background information alone, the model receives an Occam penalty.

Whether the posterior odds favor a more complex model relative to a simpler model depends on whether the complex model fits the data sufficiently better to overcome its Occam penalty. Compared to complex models, simpler models are generally ruled out more easily, because they are less able to explain a diversity of possible outcomes. On the other hand, Bayes' theorem rewards the simpler model for sticking its neck out and making less flexible predictions if those predictions come true. Bayesian analysis therefore helps find the signal without over-fitting the noise.

To see how Occam factors emerge from the mathematics of Bayesian probability, we reconsider the card-draw example presented in Section 3.4, where we draw the six of spades from a deck held by a stranger at a party. We are interested in comparing two hypotheses: H_R = *The card was arbitrarily selected from a randomly shuffled deck*, and an *ad-hoc* rival, $H_{6\spadesuit}$ = *The stranger is a professional magician with a trick deck that forces the six of spades*. The first step is to recognize that $H_{6\spadesuit}$ is one member of a family of 52 equally plausible related hypotheses, H_M = $H_M\, c_1$ *or* $H_M\, c_2$ *or ... or* $H_M\, c_{52}$, where $H_M\, c_k$ = *the magic trick forces card* c_k. In other words, we must compare H_R against H_M, a more complex model with a parameter c_k that can be adjusted to fit the data. We wish to calculate the posterior odds:

$$\frac{P(H_M|E\ I)}{P(H_R|E\ I)} = \frac{P(H_M|\ I)}{P(H_R|\ I)} \times \frac{P(E|H_M\ I)}{P(E|H_R\ I)} \qquad (A1)$$

Expanding the numerator of the likelihood ratio (the right-most term in $A1$), we have:

$$\frac{P(E|H_M\ I)}{P(E|H_R\ I)} = \frac{\sum P(c_k|H_M\ I)\ P(E|H_M\ c_k\ I)}{P(E|H_R\ I)}, \qquad (A2)$$

where we have used the law of total probability to introduce a sum over all 52 possible values of the card parameter c. In essence, we are averaging the likelihoods under each sub-hypothesis in the magic-trick family, weighted by the prior probability that the card parameter takes a particular value. When we plug in the 6 of spades for the evidence E, the sum in the numerator picks out that single value for the parameter c, because the likelihood of E=6♠ is zero for every sub-hypothesis except for that which forces the 6 of spades:

$$\frac{P(E|H_M\ I)}{P(E|H_R\ I)} = \frac{(\frac{1}{52}\times 0 + \frac{1}{52}\times 0 + \cdots + \mathbf{\frac{1}{52}}\times \mathbf{1} + \frac{1}{52}\times 0 + \cdots)}{(\mathbf{\frac{1}{52}})}. \qquad (A3)$$

In the denominator above, we have used the fact that the likelihood of E=6♠ under the random draw hypothesis is 1/52. Substituting ($A3$) into ($A1$), we can now rewrite the posterior odds ratio as the product of three factors:

$$\frac{P(H_M|E\ I)}{P(H_R|E\ I)} = \frac{P(H_M|\ I)}{P(H_R|\ I)} \times \frac{\left(\frac{1}{52}\right)}{1} \times \frac{1}{(\frac{1}{52})}. \qquad (A4)$$

These three factors on the right-hand side of ($A1$) are the model-level prior, the Occam penalty—a factor of 1/52 in the numerator, and the "fitted likelihood"—a factor of 1/52 in the denominator. The model-level prior remains to be assessed, using any salient background information about the chances that the stranger is a skilled magician as opposed to an ordinary partygoer with a randomly shuffled deck. The Occam penalty arises from the prior probability that a magic trick would favor the six of spades. The fitted likelihood, $P(E|H_M\ 6\spadesuit\ I)\,/\,P(E|H_R\ I)$, assesses how surprising or expected our evidence is under $H_{6\spadesuit}$ relative to H_R once we have chosen the six of spades as the parameter value for the magician model.

In essence, the more complex model H_M receives an Occam penalty when the data obtained rules out all but one of the 52 parameter values that were plausible *a priori.* This Occam factor keeps us from favoring the *ad-hoc* six of spades hypothesis, which on its own makes the card we chose much more likely than the random-draw hypothesis. Note that in general, the Occam factor will not exactly cancel the fitted likelihood; that effect is a special feature of this example. It is also important to emphasize that the posterior odds could end up favoring the more complex model, if the fitted likelihood is good enough to overcome the Occam factor. Accordingly, logical Bayesianism does not always favor simplicity—it balances simplicity against explanatory power.

A second example illustrates how Occam factors can emerge in explicit Bayesian process tracing.[15] Suppose we have two plausible explanations for why the government of Gonduria, a

[15] For the sake of illustration, we are explicitly identifying and evaluating an Occam penalty, but Occam factors arise automatically if Bayesian analysis is correctly employed. In actual practice, we need not think about Occam factors as a separate step in Bayesian analysis.

developing country on the Pandor continent, expanded social programs to reach a larger proportion of the poor:

> H_{WB} =*Expanding social programs was a condition for a World Bank loan;*
> H_R =*The government designed these measures to improve its approval ratings after the latter dropped below a critical threshold,* r_c.

H_R denotes a family of hypotheses, where r_c could take on many different values. A priori, it would be reasonable to assume that the threshold rating r_c falls between 25% and 50%. Regarding the upper limit, we reason that democratic governments tend to become concerned once approval ratings drop below 50%. We set the lower limit drawing on background information that approval ratings in Pandorian democracies generally have not dropped below 25% during periods of normal politics. We wish to calculate the posterior odds ratio (equation 3) for the two hypotheses in light of evidence E_0=*The government's approval rating at the time,* r^**, was 44%.*

We begin by evaluating the likelihood of the evidence under H_R:

$$P(E_0|H_R\, I)=\sum P(r_c|H_R\, I)\times P(E_0|r_c\, H_R\, I) \qquad (7)$$

where as in the previous example, we have used the law of total probability to introduce a sum over all possible values of the critical threshold (recall that each value of r_c defines a specific hypothesis in the H_R family); for simplicity we sum over integers instead of integrating over a continuum.[16] When r_c >50% or <25%, we have $P(r_c|H_R\, I)$=0. We take the prior likelihood of the threshold parameter to be uniform over the range of 25%–50%, such that $P(r_c|H_R\, I)$=1/25. Denoting evidence E_0 as r^*=44%, we have:

$$P(E_0|H_R\, I)=(1/25)\sum P(r^*=44\%|25\%\leq r_c\leq 50\%\; H_R\, I) \qquad (8)$$

The summand vanishes unless $r_c \geq$44%; otherwise the threshold hypothesis would be contradicted. For $r_c \geq$44%, we take all values of $P(r^*=44\%|r_c\; H_R\, I)$ to be equal, assuming that approval ratings at the time the government expanded social spending are independent of the critical threshold.[17] We can then replace the sum in equation (8) with a factor of 7:

$$P(E_0|H_R\, I)=(7/25)P(r^*=44\%|44\%\leq r_c\leq 50\%\; H_R\, I) \qquad (9)$$

More generally, for evidence E that includes r^*=44% along with other salient observations, we have:

$$P(E|H_R\, I)=(7/25)\times P(E|44\%\leq r_c\leq 50\%\; H_R\, I) \qquad (10)$$

We can now calculate the posterior odds ratio for H_R vs. H_{WB}:

$$\frac{P(H_R|E\, I)}{P(H_{WB}|E\, I)} = \frac{(7/25)\times P(H_R|I)\times P(E|44\%\leq r_c\leq 50\%\; H_R\, I)}{P(H_{WB}|I)\times P(E|H_{WB}\, I)} \qquad (11)$$

We find that H_R is penalized relative to H_{WB} by an Occam factor of 7/25, regardless of how plausible we find the family of hypotheses H_R relative to the World Bank hypothesis. This moderate penalty arises because the data r^*=44% rules out a moderate portion of the parameter space judged feasible given the background information. Had the value of r^* been lower, the Occam penalty would have been less significant. If the government's approval ratings at the

[16] We would not expect arbitrarily close values to be observationally distinguishable so this approximation seems reasonable.
[17] This assumption is an oversimplification—there could be many dependencies.

time fell below 25%, this evidence would be consistent with any value of the threshold between 25–50%, and H_R would not incur an Occam penalty relative to H_{WB}.