Reliability of Inference: Analogs of Replication in Qualitative Research

Tasha Fairfield
Andrew Charman

Forthcoming in Eds. Colin Elman, John Gerring, and James Mahoney, *The Production of Knowledge: Enhancing Progress in Social Science,* New York: Cambridge University Press.

How do issues related to replication translate into the context of qualitative research? As Freese and Peterson forewarn, their discussion of replication in quantitative social science cannot be directly transposed into this realm. However, we can identify analogs for the various combinations of same vs. new data, same vs. different procedures scrutiny that these authors discuss. While many of these analogs share essentially the same overarching definitions and import as their quantitative relatives, others diverge more significantly. The differences in these instances arise in large part from distinctions between frequentism, which underpins orthodox statistics, and Bayesianism, which a growing body of research identifies as the methodological foundation for inference in qualitative research.

In identifying and discussing these analogs, we will refer specifically to qualitative research that is informed by Bayesian reasoning. This approach is motivated by two considerations. First, we agree with Freese and Peterson in this volume that issues related to replication inevitably involve "a welter of complications and tradeoffs that will likely prompt different paths" for different epistemological communities, particularly when such communities espouse different understandings of inference and causation, as is the case in qualitative methods literature. Second, in our view, much of the best qualitative research that draws on process tracing and comparative historical analysis is implicitly, if not consciously, informed by Bayesian reasoning. 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 & Charman (2017) in espousing Bayesianism as the most appropriate logic of inference for qualitative research. We will therefore leave the question of how replication or analogs thereof might apply in non-Bayesian qualitative research open for other scholars to address.

By way of introduction, Bayesian inference proceeds by assigning "prior" probabilities to a set of plausible rival hypotheses given the (limited) information we possess. These prior probabilities represent our degree of confidence in (or vice versa, our degree of uncertainty about) the truth of each hypothesis taking into account salient knowledge from previous studies and/or experience. We then consider evidence obtained during the investigation at hand. 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. Bayesianism provides an especially appropriate framework for qualitative research given the following considerations: (1) Bayesianism is well-suited for explaining unique historical or sociopolitical events and allows us to conduct inference with a small number of cases and/or limited amounts of data; (2)

Bayesianism can handle non-stochastic data that cannot naturally be considered to arise from a randomized sampling procedure or experiment (e.g., interviews with expert informants and evidence from archival sources); and (3) Bayesianism mandates an iterative process of theory development, data collection, and analysis, which is how qualitative research almost always proceeds in practice (Fairfield & Charman 2019).

In this chapter, we will advance two positions that we believe could help promote greater consensus and common ground among quantitative and qualitative scholars. First, we advocate restricting the use of the term *replication* to a narrowly-defined set of new-data, same-procedures scrutiny that applies to orthodox statistical analysis and experimental research, both for the sake of clarity and to avoid the perception that norms from dominant subfields are being imposed inappropriately on qualitative research.[1] Second and relatedly, we argue that the overarching concern in all scientific inquiry—both quantitative and qualitative—is *reliability* of inference: how much confidence we can justifiably hold in our conclusions. Reliability encompasses but extends beyond the notions of replication and reproducibility. As Goodman, Fanelli, and Ioannidis (2016:1) observe: "The fundamental concern ... is ... not reproduciility per se, but whether scientific claims based on scientific results are true." Our discussion therefore focuses on practices that could help improve how we assess evidence, build consensus among scholars, and promote knowledge accumulation in qualitative research within a Bayesian framework, which provides a natural language for evaluating uncertainty about the truth of hypotheses.

Section 2 presents our understanding of replication and reliability as applicable to different types of research, characterized by the data (quantitative vs. qualitative) and the methodological framework (frequentist vs. Bayesian). Here we offer some suggestions for effectively conducting new-data scrutiny of qualitative research, although our focus will be on same-data scrutiny (Section 3), which we believe could have significant payoffs for improving reliability of inference. Accordingly, Section 3 elaborates Bayesian rules for same-data assessment and illustrates how they can be applied using published exemplars of process-tracing research and comparative historical analysis. In broad terms Bayesianism directs us to ask whether scholars have overstated the weight of evidence in support of the advocated argument by neglecting to assess how likely that evidence would be if a rival hypothesis were true, whether the hypotheses under consideration have been articulated clearly enough to assess how likely the evidence would be under a given explanation relative to rivals, and whether the background knowledge that scholars discuss justifies an initial preference for a particular hypothesis. We contend that Bayesianism provides a clear framework for scrutinizing analysis that can help build greater consensus among scholars and facilitate knowledge accumulation.

Analogs of replication in qualitative research

In their elucidating discussion of different and wide-ranging understandings of replication in quantitative social science, Freese and Peterson distinguish between using (1) the same vs. new data, and (2) the same vs. different procedures *vis a vis* the study in question. The latter dimension in their view correlates with whether the goal entails scrutinizing the study's specific results vs. the broader conclusions its draws. We likewise structure our discussion of reliability in qualitative research around these useful dimensions of data and procedures, which form the four rows in Figure 1.

[1] Such perceptions have arisen in the debate surrounding APSA's DART initiative; see for example the Qualitative Transparency Deliberation blog.

However, rather than labeling each of the four rows as different types of replication (e.g. same-data/same-procedures replication, same-data/different-procedures replication, etc.) we use terms that are intended to reflect the specific scrutinizing and/or knowledge-accumulating activities that fall within each row. Our intent is to sidestep the confusion that has plagued debates about replication and to avoid conceptual stretching. Freese & Peterson (p. 13, 3) emphasize that: "the term [replication] itself is used to mean such different things that discussions across fields go easily astray," and recount that they have observed "conversations about 'replication' in which parties have gone on for a disconcertingly long time before realizing they were using the term to mean different things." From a conceptual perspective, as Freese and Petersen point out, many scholars view the activities that fall under "same data, same procedures" as entailing something more trivial (although still important) compared to "new data, same procedures," which is the classic domain of replication in frequentist statistics.[2] Accordingly, for the former category we adopt Freese and Peterson's term *verification*—checking that "reported results can be generated from the data used."[3] Likewise, activities involving "new data, new procedures" in our view go beyond the scope of replication, as other scholars have also contended (see for example Freese and Peterson's discussion of "conceptual replication" in experiments). For this row, we instead coin the term *extended research*—assessing theory and generalizing findings to other contexts. Activities entailing "same data, different procedures" fall closer to the realm of replication and in our view are central to the question of whether reported results reliably follow from the evidence presented. However we prefer the term *sensitivity analysis*, which we feel more accurately captures the goal at hand.

Turning to the columns of Figure 1, we divide quantitative research on the left into two categories: frequentist vs. Bayesian, with qualitative Bayesian research on the right. In a given row, we use the same term across columns for activities that are closely analogous in essence—for example, *verification* and *extended research* in all quantitative and qualitative research.[4] In some instances, however, there are important distinctions to be made between frequentist quantitative analysis vs. Bayesian quantitative analysis, and/or quantitative Bayesian analysis vs. qualitative Bayesian analysis. The former distinctions arise from the epistemological differences between frequentism and Bayesianism, which we will explicate in Section 2.3. The latter distinctions arise from the greater degree of subjectivity inherent in analyzing the inferential import of qualitative evidence.

The most significant differences across columns arise in the new-data/same-procedures row. Here we apply a narrow definition of *replication* that is restricted to frequentist quantitative analysis and experimental research. For the more distant qualitative Bayesian analog, we devise the term *continued research*—gathering additional evidence to combine with existing evidence toward strengthening inferences within the same research design—which is also salient for quantitative research that employs a Bayesian framework. These terminological choices are again motivated by our concerns regarding conceptual stretching of "replication" beyond its natural epistemological context. As elaborated in Section 2.3, applying identical procedures to analyze new data makes sense for controlled experiments or observational studies where a different random sample can be taken from the same population (e.g., third-year east-coast college students, or light from a distant galaxy—as long as salient characteristics of the data sources do not change over the intervening time period), but in qualitative research, evidence (e.g., interviews with expert informants) can rarely be treated as a random sample, and in accord with a Bayesian logic, the goal of collecting more data is

[2] As reflected in terms used to distinguish new data/ same procedures from the other categories, which include "direct replication," "mechanical replication," and "close replication" (Freese & Peterson).

[3] Clemens (2017:327) uses the term verification similarly.

[4] Verification and extended research should also be directly applicable to non-Bayesian qualitative research.

to accumulate information and thereby strengthen inferences, rather than to validate findings by independently reproducing results with a new sample or a new run of the experiment. It is worth stressing that our distinction between *replication* and *continued research* is not intended in any way to downplay the importance of "new-data/same-procedures" assessment in qualitative research.

The following sub-sections explicate the qualitative analogs in each row of Figure 1 in more detail. Section 3 will then focus on same-data scrutiny of qualitative research (the first two rows of Figure 1), which we find most salient to discussions of replication—or as we prefer, reliability—in qualitative research.

Figure 1 here

Same Data—Same Procedures

The aim in this row entails examining existing research to ascertain whether the findings justifiably follow from the data and analytical procedures employed. Focusing first on the qualitative research column, an initial task entails assessing coding and measurement—asking whether concepts are well defined, operationalized, and scored in the study. A large body of literature provides guidelines to this end, so we will not dwell upon it here.[5] Checking that evidence has been accurately quoted or summarized from the original sources is another basic activity that we would include here (see Reiter in this volume on "measurement replication").

A second task entails verification. In a Bayesian framework, this involves checking whether the mechanics of probability theory have been correctly applied, as explicated in Section 3.[6] We view this question as an objective, technical matter, with the caveat that operationalizing Bayesian reasoning in qualitative research remains an active methodological frontier. As such, the literature contains different understandings of Bayesianism, different approaches to adopting Bayesianism in qualitative and multi-method research, and different levels of technical sophistication. For purely qualitative case research, we advocate the "logical Bayesian" approach presented by Fairfield & Charman (2017) (discussed further in Section 3).

For qualitative research, a third task entails scrutinizing the inferential weight of the evidence. This task inevitably involves an element of subjective judgment. In logical Bayesianism, probability represents the *rational degree of belief* we should hold in some proposition, such as a causal hypothesis, in light of all relevant background information we possess and all evidence collected as part of the study—independently of subjective opinion, personal predilections, or subconscious desires. However, logical Bayesiainism is an aspirational ideal that usually cannot be fully realized without approximations—all the more so in qualitative social science where there are no clear mathematical procedures for objectively translating complex, narrative-based, non-reproducible, highly contextual qualitative information into precise probability statements. Additional challenges arise when experts aim to independently assess the weight of evidence, as determined by the *likelihood ratio* (Section 3.1), which entails asking how much more plausible the evidence is under a given hypothesis relative to a rival. Background information shapes how we interpret evidence, yet scholars will bring very different knowledge to the table. Authors should explicitly invoke and discuss those elements of their background information that matter most for weighing the evidence,

[5] E.g., Collier, Mahoney, Waldner.

[6] E.g, has the author correctly defined the weight of evidence, does Bayes' rule yield the reported posterior odds on the hypotheses given the author's stated prior odds and weight of evidence.

but it is impossible to systematically list all salient background information that informs our analysis. Accordingly, the goal of scrutiny should not be to exactly reproduce numerical probabilities, but instead to promote discussion and foster a reasonable level of consensus on the inferential weight of evidence. As part of this process, elements of the author's background information that were previously used implicitly or sub-consciously may come to light and help to resolve disagreements.

Turning to quantitative research, both tasks—assessing coding and measurement, and verification—are of course relevant as well. The first is an important but to our knowledge much less widely discussed issue compared to verification, which would include double-checking mathematical procedures and computer codes for errors.

Assessing measurement and coding can be fairly straightforward in some contexts—for example, running inter-coder reliability checks when datasets have been generated by coding mentions in documents or open-ended responses on surveys. In other contexts, it may be quite challenging and time consuming, particularly when large amounts of qualitative information have been condensed into quantitative scores. Mark Beissinger's comments on the Qualitative Transparency Deliberation blog are highly salient in this regard:

> Having created several large-n datasets, I can attest to the enormous amount of research that necessarily must go into the coding of each individual variable for each observation and the judgments that are made by researchers in making those codings. If we are talking about real transparency in research, large-n researchers would need to provide extensive documentation on every single coding in their datasets.[7]

Despite the practical challenges, the importance of assessing measurement and coding in dataset generation should not be underestimated. For example, Haggard and Kaufman (2012:501) point out that their study on inequality and regime change:

> ...raises serious questions about the validity of the coding of democratic transitions in these two major datasets and, as a result, casts doubt on the inferences that have been drawn in the quantitative work that employs them. Only 55.4% of the CGV [Cheibub, Ghandi, & Vreeland] transitions are also Polity [IV] cases, and 21 of the 65 CGV transitions had Polity scores of less than 6. Even where the two datasets are in agreement, moreover, our examination of the cases raises questions about the validity of the coding process.

Our point here is not only to emphasize that dataset generation should be included more centrally in discussions about reliability, replication, and research transparency, but also to emphasize that the subjective judgments that necessarily complicate qualitative research are not absent from quantitative research; they simply enter at a stage that tends to receive less attention in these debates.[8]

Same Data—Different Procedures

Further assessments of reliability of inference can be conducted by reanalyzing a study's data with slightly altered procedures. We refer to such endeavors as *sensitivity analysis*, whether the data is quantitative or qualitative, with an additional related but distinct task that we call *consistency checks* in qualitative Bayesian research.

---

[7] www.qualtd.net/viewtopic.php?f=13&t=157 Beissinger goes on to warn that such requirements would be infeasible: "This is simply not being asked because it is not practicable—even though the real instances of fraud that we are aware of have come from falsified codings in large-n data sets."

[8] For excellent work on assessing conceptualization and measurement in large-N datasets, see Kurtz and Schrank's (2007) critique of the World Bank Government Effectiveness Indicators and Coppedge et. al's (2011) critique of Freedom House and other democracy indicators.

Importantly, our understanding of what "different procedures" entail necessarily changes when moving from frequentistism to Bayesianism. In the former world, different procedures might include altering model specifications or choosing different test statistics or estimators. In Bayesian analysis, "different procedures" takes on a much narrower meaning. Within a logical Bayesian framework, we must make judgments about which hypotheses to consider, where and how to acquire data, and how to interpret qualitative evidence. However, the underlying inferential procedure remains the same: apply the rules of probability to update initial beliefs regarding the plausibility of rival hypotheses in light of the evidence. As elaborated in Section 3, the analysis always involves assessing prior probabilities, assessing likelihood ratios, and updating probabilities in accord with Bayes' rule. Unlike frequentist analysis, there is no need to choose among sampling procedures, stopping rules, estimators, test statistics, or significance levels.

However, there is scope within Bayesianism for sensitivity analysis that entails assessing how much conclusions are affected by different choices of prior probabilities for the hypotheses—this is common practice in Bayesian statistics (e.g., Berger and Berry 1988:162, Greenland 2006:766). When working with qualitative information, where subjective judgment enters more strongly, additional sensitivity assessments can be carried out by varying the inferential weight attributed to the most salient pieces of evidence. This practice is particularly useful if the interpretation of key pieces of evidence is sensitive to background assumptions; for example, our level of trust in informants and/or the instrumental incentives we attribute to them.

Logical Bayesianism also provides opportunities to carry out various *consistency checks* on our inferential reasoning when we are working with qualitative evidence. Most importantly, rational reasoning requires that we must arrive at the same inference if we incorporate distinct pieces of evidence into our analysis in a different order, or if we parse the overall body of evidence into either more finely grained, or more coarsely-aggregated pieces of information. These consistency checks in essence entail "solving the problem" in different but logically equivalent ways. Section 3.5 will explicate these consistency checks in greater detail.

New Data—Same Procedures / Specific Results

When we arrive at the new-data/same procedures row, we find more significant differences as we move from quantitative frequentist analysis on the left-hand side of Figure 1 to qualitative Bayesian analysis on the right-hand side.

In quantitative frequentist research, applying identical procedures to new data is the classic realm of replication in the narrow sense. This endeavor entails using the original study's designated stochastic sampling and analysis procedures to obtain new data from the same population.[9] The centrality of repeated random sampling (in principle, if not always in practice) is built into frequentist inference on a foundational level. Probability itself is understood as the limiting proportion of some particular type of event in an infinite sequence of repeated random trials; p-values/significance levels, statistical power, and confidence levels are all defined in terms of long-run relative frequencies under repetition of the same experiment or sampling process, and randomization is meant to ensure balance between groups (only) in the long run—that is, under indefinite repetitions of the procedure.

Bayesianism, in contrast, focuses on making the best conclusions possible from all available data, whether generated through a stochastic process or not; repetition and long-run frequencies play much less central roles in inference, because within a Bayesian framework, what could have but did not happen, or what might but has not yet happened, is irrelevant for drawing inferences from

[9] Our narrow definition of replication corresponds to Clemens' (2017:327) "reproduction test," which entails "sampling precisely the same population but otherwise using identical methods to the original study."

the actual data in hand. The salient "new-data/same-procedures" endeavor is best described as *continued research;* namely, collecting more data (according to the parameters of the original research design) that will contribute to the final reported inferences, hopefully by reducing our uncertainty regarding which hypothesis is correct and/or narrowing posterior error bars for parameter estimates.

We emphasize that the key distinction between frequentist replication studies and "continued research" within a Bayesian framework is that the new data produced naturally contribute to the inferences *in combination* with the evidence that was already analyzed, not independently of the original data. Frequentist statistical theory is ill-suited for synthesizing information from multiple studies regarding the same hypotheses, in that it provides no universal rules for aggregating p-values from multiple null hypothesis significance tests into an overall measure of support, nor for assessing and combining the systematic (non-random) components of estimation errors. Therefore, in frequentist-based inference, replication and meta-analysis (i.e., combining the results of many studies) traditionally have been considered as separate endeavors.[10] Alternatively, treating repetitions as intermediate steps in a so-called sequential analysis may also be problematic, since in principle all of the sampling and stopping rules must be specified in advance, and the difficulty of the analysis tends to grow with the complexities and contingencies of these rules. Given these challenges, scholars are increasingly using Bayesian tools to conduct meta-analysis of individual studies that employ orthodox statistics (e.g., Pereira & Ioannidis 2011, Ionatis et al. 2016, van Aert & van Assen 2017). However, this epistemological mismatch leads to awkwardness, as scholars must make Bayesian sense of frequentist notions such p-values, power analysis, and estimators.

In contrast to frequentism, Bayesianism provides a unified procedure for combining evidence and learning from accumulated information. Bayesianism naturally accommodates contingent data gathering or follow-up data collection without need to distinguish preliminary from subsequent stages of analysis. And when Bayesian methods are employed at the stage of primary research and reporting, any subsequent meta-analysis simply proceeds via the same inferential framework. Probability theory itself is the mathematical expression of this updating process. Learning in the Bayesian framework occurs by virtue of the fact that (a) probability is understood as a logical concept that represents a rational degree of belief given the limited information we possess, not a long-run frequency, and (b) all probabilities within Bayesianism are necessarily *conditional* probabilities: confidence in one proposition depends on what else we know and generally changes when we make new observations. After completing an initial round of research, the posterior probabilities on our hypotheses, which take into account all evidence known so far, become the prior probabilities when we move forward to analyze additional evidence.

Stated in more concrete terms, the inferential weight of evidence in a Bayesian framework is additive. If evidence from a second round of research (or subsequent study following the same research design) runs counter to the inference drawn from a first round of research (or previous study), we would not necessarily conclude that the inference from the first round of research is "invalid" (unless scrutiny along the lines described in Sections 2.1. and 2.2 reveals problems with the analysis.) Instead, we aggregate the evidence from both rounds and examine the combined weight of evidence. Suppose the initial research concluded that hypothesis H_1 is substantially more plausible than rival hypothesis H_2, whereas the aggregate evidence from both rounds now leads to the conclusion that H_2 is slightly more plausible than H_1. We would not say that the findings from the first round "failed to replicate;" instead, we would assert that the tentative conclusions inferred from the earlier evidence no longer hold in light of the larger body of evidence now available. At any stage of research, our results are always conditional on the hypotheses we are comparing and the

[10] Some countervailing recommendations are beginning to emerge.

information we currently possess, and inferences are always subject to change in light of new information.

One example in which a Bayesian might conduct something more akin to a narrow-sense replication study intended to produce an independent data set for comparison would be if there are concerns about whether an experimental apparatus is functioning correctly (e.g., Could the detector be miswired? Did the laser fire at the right intervals?). But if the new results suggest that the apparatus did indeed work properly in the original run, any data generated in the second round (conducted under conditions as close as possible to the original run) could be combined with the data from the previous round via Bayes' rule to produce a single cumulative inference, and we would hence return to the realm of "continued research" rather than multiple replications of the experiment per se. In sum, "viewed through this [Bayesian] lens, the aim of repeated experimentation is to increase the amount of evidence, measured on a continuous scale, either for or against the original claim," (Goodman et al. 2016:4).

The distinctions between frequentist replication studies and Bayesian "continued research" become more acute when we move from quantitative to qualitative research, where evidence can rarely be considered to arise from a stochastic sampling process or randomized controlled experiment. Consider interviews. In contrast to large-N survey research where random sampling is the norm, qualitative scholars purposively seek out key informants based on prior expectations of their knowledge about the topic of investigation, in accord with the Bayesian principle of research design via maximizing expected learning. In some studies those interviewed may essentially exhaust the population of expert informants.[11] Similarly, qualitative scholars often aim to scrutinize all relevant documents (e.g. legislative records, news articles, reputable historical accounts)—not some sample thereof—that might provide salient evidence.[12] Cases themselves (whether countries, policymaking episodes, or electoral campaigns) are often selected based on expectations that they will be rich in data and/or will facilitate strong tests of competing theories, which again conforms to the Bayesian principle of maximizing expected learning (Fairfield & Charman 2018).[13] Even if random sampling of cases were desirable (e.g., for lack of any better criteria), it is very often impossible to define or delineate in advance all members of the population from which the sample would be drawn. For example, scholars often discover new cases while conducting fieldwork.[14] Case selection strategies in qualitative research are sometimes presented as "replicable," (see Kapiszewski 2012:211); however, applying the designated procedures again would result in the same set of cases (or at least a very high proportion of the same cases), not an independent sample containing different cases that could be used to test the stability of the original findings.

Setting aside the issue of whether a stochastic data-generation process is fundamental to the notion of replication—our above discussion does not imply that "replication" in the literal sense of repeating exactly the same procedures to obtain new data is impossible in qualitative settings. In theory, a scholar could take a published study, follow the same case selection and data-generating procedures to the last detail described, visit (potentially the same) field sights, interview (potentially the same) informants, consult (potentially the same) archival sources, etc., and produce a separate analysis that tests the same hypotheses. We cannot imagine any scholar wanting to undertake such

[11] Fairfield's (2015) research on taxation would be an example; nearly all current and former Chilean Finance Ministry technocrats with first-hand knowledge about tax policy formulation in the 1990s and early 2000s were interviewed.

[12] Moreover, repeated interviews with the same informant or repeated consultation of the same documents cannot sensibly be regarded as an independent random sample of statements from those sources.

[13] See also Van Evera's (1997) notions of seeking "data rich cases" and/or "cases in which different theories make divergent predictions."

[14] See Fairfield (2015: Appendix 1.3) on case selection under such circumstances.

an endeavor, but more importantly, we do not see much value in such an approach—indeed it strikes us as the least productive way to improve reliability of inference and foster knowledge accumulation in qualitative research. These goals are better served by scrutinizing the evidence and analysis presented in the original study—via the various activities described in Sections 2.1 and 2.2—and then seeking salient new evidence from complementary data sources through continued research, and/or extended research that examines different settings and refines hypotheses in new contexts (Section 2.4 below).

To summarize our argument, we view "replication" in the narrow-sense as largely irrelevant within a Bayesian framework—with the exception of a few specific contexts that may arise in quantitative/experimental research—because Bayesian inference follows different epistemological principles from frequentist inference. The fact that "replication" would be impractical for qualitative research is a secondary matter—the fundamental issue is that frequentist inference, with its emphasis on random sampling and stochastic data, is not appropriate for qualitative social science. Quoting Jackman and Western (1994:413): "*frequentist inference is inapplicable to the nonstochastic setting.*"

As with quantitative Bayesian research, continued research in a qualitative Bayesian setting aims to gather more and different evidence that will improve inferences made from previously obtained information. Qualitative scholars regularly gather additional evidence after conducting an initial round of analysis and drawing tentative conclusions. Continued research might entail collecting new information from sources that were previously consulted—for example, following up on previous interviews by asking the informant new questions or seeking to clarify the meaning of a response from an earlier conversation. Equally well, continued research can seek new sources of information—e.g., an archive collection recently made public, an outgoing government official who now has time or political leeway to grant interviews, or an expert witness who can now comment on a recent court decision.

Continued research is typically carried out by the original author(s) in the context of refining or improving the study, either by gathering new information for the same cases or by including new cases that fall within the parameters of the original research design and the original scope of the hypotheses considered.[15] Independent scholars might conduct some work of this type, particularly if the case(s) in question are of intrinsic substantive interest and/or carry special theoretical import, and if they deem that gathering additional evidence could significantly strengthen or alter the inferences. More typically however, independent scholars would engage in *extended research* as described in the following section.

Our recommendations for conducting productive continued research are both intuitive and practical in nature:

- *Carry out preliminary analysis of evidence periodically.*

  This practice helps us take stock of what has been learned so far and what kinds of evidence or clues would be most valuable moving forward.[16] In a Bayesian framework, the most decisive pieces of evidence are those that fit much better with a given hypothesis compared to the rival. Accordingly, we should think about where the rival hypotheses under consideration would most tend to disagree, identify divergent predictions, and seek additional evidence accordingly.

[15] It is worth noting that a "research design" in qualitative contexts tends to be much looser, more flexible, and less detailed than what one would find in experiments or in frequentist hypothesis-testing contexts, where all procedures must be specified from the outset.

[16] Kapiszewski et al. (2015: Chapter 10) provide useful general guidance on conducting analysis while gathering data in the field.

- *Revisit key informants.*

  This strategy allows the scholar to gather new information as new hypotheses and new questions arise over time. For example, Fairfield (2015:23) conducted follow-up interviews with the informants who possessed the most extensive, first hand knowledge of tax reforms over the course of both primary and follow-up fieldwork in order to pursue new lines of inquiry and to dig deeper into political processes in light of conflicting or unexpected accounts from other sources.

- *Keep an eye open for not only new sources of evidence, but also informative new cases that emerge as time progresses.*

  Just as we are free to gather new evidence and combine it with previously-acquired knowledge, within a Bayesian framework we can include additional cases that provide fruitful grounds for testing rival hypotheses (Fairfield & Charman 2018). Boas (2016) provides an excellent example (although he does not cast his research design in Bayesian terms). After conducting extensive research on presidential campaigns that took place from the late 1980s through 2006 in Chile, Brazil, and Peru and presenting findings based on the evidence compiled (Boas 2010), Boas (2015: 29, 32-34) included three additional presidential campaigns that took place in those countries from 2009–2011, along with a set of elections from other countries that he deemed to fall within the scope conditions of his "success-contagion" theory.[17]

- *Arrange follow-up trips to field sites.*

  If resources permit, this strategy allows the investigator to gain perspective over time and seek out specific sources and clues that can best contribute to filling in gaps and strengthening the weight of evidence where needed. For example, Garay (2016) returned to Argentina after conducting primary fieldwork and successfully obtained more decisive evidence in favor of her argument that electoral competition in presidential elections is critical for expansion of social programs, as opposed to competition in legislative elections, through interviews with participants in the 2009 midterm elections, which resulted in the government's loss of its absolute majority. Similarly, Fairfield (2015) was able to secure a critical interview with a former Chilean finance minister in 2007 following primary fieldwork in 2005; the interview significantly strengthened her conclusion that business's instrumental (political) power, not structural (investment) power, explained Chile's meager progress toward increasing progressive taxation during Ricardo Lagos' presidency.

  While costly, follow-up fieldwork can occasionally be strategically planned in advance. For example, Garay (2016) anticipated the need to revisit Mexico after the 2006 election, given that additional research on social policy innovations would be highly valuable regardless of which candidate prevailed; her SSRC grant allowed here to conduct several months of research prior to the elections and several month of research after the new government took power later in 2006 (Garay 2018, private communication).

By and large, we view good continued research as simply good research. Our overarching recommendation for improving continued research is to become familiar with the basic principles of

[17] Likewise, after conducting an initial round of research from 2006–2008 on tax reforms in Argentina, Bolivia, and Chile (Fairfield 2010, 2011, 2013), Fairfield (2015) added Chile's 2012 corporate tax increase. This case illustrates the phenomenon of popular mobilization counteracting business power—previously observed in Bolivia's 2005 hydrocarbons royalty—in a different context, thereby adding further support for the theory.

Bayesian probability and to implement Bayesian-inspired best practices for assessing the inferential weight of evidence (Fairfield & Charman 2017, 2018).

New Data—Different Procedures / Broader Findings

Whereas *continued research* entails gathering more data within the parameters of the original research design, *extended research*—the final row in Figure 1—pushes beyond previous work by testing theories in new ways, or assessing and/or refining theory in new contexts or domains. Examples for quantitative scholarship might include asking how well findings generalize to different populations—e.g., do African voters respond to information about corruption in the same way as Brazilian voters? Do similar results hold for adults with college degrees as for third-year undergraduates? In qualitative research, scholars regularly include preliminary discussions of how findings and hypotheses might apply elsewhere—whether in different regions or countries, different time-periods, or different policy areas.[18] Literature on well-developed research agendas—for example, state-building or welfare provision—commonly assesses hypotheses from foregoing studies with new data from different contexts. Meanwhile, research on new or less-studied phenomena regularly draws on theories from other domains and adapts or extends them to address the questions at hand.[19]

Scholarship on state-building and institutional development exemplifies a long-term trajectory of extended research. Pioneering work by Tilly (1992), Ertman (1997), and others on early modern Europe hypothesized that warfare drove the formation of strong and effective state institutions. Subsequent authors evaluated that theory and refined it to craft explanations for state-building outcomes in Africa (Herbst 2000), Latin America (Centeno 2002), and China (Hui 2010). Similarly, research on the resource curse in paradigmatic cases such as Venezuela (Karl 1997) stimulated continual reassessment and refinement of that theory in light of data on other cases, in Latin America and far beyond. New studies regularly assess the warfare hypothesis and/or the resource-curse hypothesis, proposing amendments or new explanations where they fall short (Slater 2010, Kurtz 2013, Soifer 2015).

Within this literature, we find several excellent individual illustrations of "extended research." Kurtz (2013) carefully scrutinizes assumptions underpinning previous work on the warfare and resource-curse hypotheses; for example, he argues that at their core, these explanations are essentially functionalist and lack well-specified causal mechanisms (Kurtz 2009:483). The alternative theory of state-building that he develops, which focuses on labor-repressive agriculture as the key impediment to institutional development, is inspired by Barrington Moore's classic work on democratization. In essence Kurtz adapts and extends Moore's theoretical insights to a different domain. Soifer (2015) in turn aims to construct a broader theory that accounts for state weakness not just as a result of no efforts at state building, but also as a result of failed efforts at state-building. Among other countries, Soifer reconsiders Peru. On the basis of distinct evidence about development of education and infrastructure, he agues that Kurtz mischaracterizes the period prior to the 1890s as a case in which state building efforts never emerged, whereas it actually exemplifies failed efforts at state-building (Soifer 2015:19). Throughout the empirical chapters Soifer reconsiders the warfare and resource curse hypotheses and argues that his theory, which focuses on the nature of local administrative institutions, better fits the evidence.

---

[18] Excellent examples include Boas (2016: Chapter 5) and Garay (2016: Chapter 8).

[19] For example, Fairfield (2015) takes the classic concepts of business's instrumental and structural power from literature on regulation and welfare in the United States (Vogel, Hacker & Pierson), refines them, and integrates them into a unified theoretical framework for explaining tax reform politics in Latin America, which had received little previous scholarly attention.

Our main recommendation for honing "extended research" in qualitative scholarship, so that it more effectively contributes to knowledge accumulation, is once again to carefully apply Bayesian reasoning. As discussed in Section 2.3, Bayesianism provides a natural framework for systematically aggregating inferences across multiple pieces of evidence—whether drawn from a single study, or pooled across studies. For well-established research agendas such as state-building or democratization, it would be especially useful to assess new theories against established rival arguments in light of the key pieces of evidence presented in prominent studies that originated those rival arguments—in addition to considering new and/or distinct evidence. For example, how well would Kurtz's (2009) novel labor-repressive agriculture hypothesis fare against the resource curse in light of evidence from Venezuela, a paradigmatic case that established the salience of resource wealth for institutional development (Karl 1997)? How well would Soifer's (2015) new administrative institutions hypothesis fare against Kurtz's (2009) labor-repressive agriculture hypothesis in light of the evidence that Kurtz presents to substantiate his theory? Such assessments would more consistently incorporate prior knowledge into the analysis, along with new empirical information. In the democratization literature, Collier and Mahoney (1997) informally follow this type of approach, revisiting the same case-study literature on which elite-based theories of transition were based to argue that a framework granting a more central role to labor mobilization better accounts for the political dynamics of democratization. An important related point is that knowledge accumulation depends on clearly specifying the hypotheses under consideration. New theories need to be clearly differentiating from existing theories so that we can assess how their empirical implications differ—namely, to what extent is a given piece of evidence is more or less likely under the new explanation relative to the plausible extant rivals.

Same-Data Assessments of Inference in Qualitative Bayesianism

Same-data assessment of qualitative research (Sections 2.1-2.2) is crucial for ascertaining to what extent reported results reliably follow from the evidence and analysis that scholars present. In this section, we aim to illustrate the different components of same-data assessment within a logical Bayesian framework and demonstrate how they could be implemented to improve the quality of inference and scholarly consensus in qualitative research. We begin by overviewing the basics of Bayesian inference and how Bayesian reasoning can be applied in qualitative research. We then elaborate rules for verification, scrutinizing weight of evidence, sensitivity analysis, and consistency checks. Throughout, we illustrate these Bayesian rules for good same-data assessment with concrete empirical examples that draw on published process-tracing research and comparative historical analysis.

Bayesian Inference in Brief

As noted previously, logical Bayesianism (Cox 1961, Jaynes 2003) views probability as the rational degree of belief we should hold in a hypothesis or some other proposition given all relevant information we possess, which will inevitably be limited. When probabilities take on the limiting values of one or zero, we are certain that the hypothesis is true or false, respectively. When probabilities take on intermediate values, we are in a situation of uncertainty regarding the truth of the proposition.

Bayes' rule, expressed in terms of conditional probabilities, states that:

$$P(H|EI) = P(H|I)\ P(E|HI) / P(E|I) \qquad (1)$$

The term on the left is the *posterior probability* in the truth of hypothesis H, given a body of evidence E as well as the salient background information I that we bring to bear on the problem. The first term on the right is the *prior probability* of the hypothesis given our background information alone. The second term on the right is the *likelihood* of the evidence: if we take H and I to be true, what is the probability of the evidence? The term in the denominator is the unconditional likelihood of the evidence. We can sidestep having to evaluate $P(E|I)$, which is usually very difficult, by directly comparing rival (mutually exclusive) hypotheses:

$$\frac{P(H_1|E\ I)}{P(H_2|E\ I)} = \frac{P(H_1|\ I)}{P(H_2|\ I)} \times \frac{P(E|H_1\ I)}{P(E|H_2\ I)}. \quad (2)$$

This relative odds-ratio form of Bayes' rule states that the posterior relative odds on H_1 vs. H_2 in light of the evidence must equal the prior odds multiplied by the *likelihood ratio*. The likelihood ratio can be thought of as the relative probability of observing the evidence E if we imagine living in a hypothetical world where H_1 is true, compared to the probability of observing E if we imagine living in an alternative hypothetical world in which H_2 is true (Fairfield & Charman 2017).

Assessing likelihood ratios is the central inferential step in Bayesian analysis that tells us how to update our odds on the hypotheses—does the evidence increase our confidence in H_1 or does it increase our confidence in H_2? In qualitative research, we must "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_j world" relative to the "H_k world," then that evidence increases the odds we place on H_j vs. H_k, and vice versa. In other words, we gain confidence in one hypothesis vs. another to the extent that it makes the evidence obtained more plausible.

Bayesian reasoning in qualitative research can be applied either heuristically—thinking qualitatively about probabilities without explicitly invoking the formal apparatus of Bayesian mathematics—or explicitly—by quantifying probabilities and using Bayes' rule. As an introduction to heuristic Bayesian reasoning, consider the following example that draws on Stokes' (2001) research on "neoliberalism by surprise" in Latin America during the debt crisis,[20] which is often used for explicating process tracing (Collier et. al 2004:257). Stokes assesses two hypotheses (assumed mutually exclusive) to explain why presidents who campaigned on protectionist economic platforms switched course once in office. The first is a representation hypothesis:

> H_{Rep} = Presidents violated protectionist policy mandates in order to represent voters' best interests.

The second is a rent-seeking hypothesis:

> H_{Rent} = Presidents violated those mandates in order to seek rents associated with neoliberal reforms (e.g. privatization).

For the case of Argentina's President Menem (1989-1999), Stokes presents the following evidence:

> E_M = In a 1993 interview with an Argentine journalist he [Menem] described his strategy during the 1989 campaign: "...If I had said 'I will privatize the telephones, the railways, and Aerolineas Argentinas,' the whole labor movement would have been against me. There was not yet a clear consciousness of what was required." (Stokes 2001:72)

As Stokes argues, E favors H_{Rep}, but a Bayesian analysis suggests that E_M only favors that hypothesis *weakly* relative to the rival. In a world where H_{Rep} is true, E_M is unsurprising; Menem himself has sketched out the logic behind the representation explanation (although much is left implicit), and he would have every reason at this time to disclose his noble motives and foresight in choosing policies

[20] See Fairfield & Garay (2017: Appendix) for additional examples of heuristic Bayesian reasoning.

that successfully stabilized the economy. However, Menem might also tell this same story to justify his behavior if H_{Rent} is true, given that he would be highly unlikely to admit rent-seeking motives, and he would have similar incentives in this world to claim credit for fixing the economy. E_M is nevertheless somewhat less likely in the H_{Rent} world, since there are other cover stories and responses Menem could provide when interviewed about his policy decisions (e.g., 'I had no choice given constraints imposed by the IMF,' 'I deny campaigning on an anti-neoliberal platform', 'I have no comment'). In sum, even though E fits very well with H_{Rep}, this evidence is only somewhat more probable under H_{Rep} relative to H_{Rent}. Learning E_M therefore at most modestly increases the odds in favor of the representation hypothesis.

When quantifying probabilities for explicit Bayesian analysis, we advocate working with a logarithmic scale, which mirrors how the brain processes information, in conjunction with an analogy to sound, such that relative probabilities are expressed in decibels (Fairfield & Charman 2017). Equation 2 then becomes simply:

$$\textit{posterior log-odds (dB)} = \textit{prior log-odds (dB)} + \textit{weight of evidence (dB)}, \quad (3)$$

where the weight of evidence, $WoE\ (H_1 : H_2)$, is proportional to the log of the likelihood ratio.

Keeping in mind that 3 dB corresponds to the minimal difference an adult with good hearing can detect, 5dB is clearly noticeable, 10 dB sounds about twice as loud, 20 dB sounds four times louder, and 30 dB sounds eight times louder—roughly corresponding to the difference between a normal conversation and a passing motorcycle—we can now express our assessment of weight of evidence in the above example more precisely: $WoE_M\ (H_{\text{Rep}} : H_{\text{Rent}}) = \sim 5$ dB.

Verification: Checking for technical errors

In many ways Bayesian analysis mirrors how we intuitively reason in qualitative research; however, in some regards correctly applying Bayesian reasoning changes the way we use our intuition and focuses attention to issues that might otherwise be overlooked when writing case-study narratives. We highlight here a few of the most relevant points that should be checked when scrutinizing Bayesian analysis in qualitative research, in accord with the guidelines discussed in Fairfield & Charman (2017).

As emphasized above, assessing the weight of evidence or the likelihood ratio—the key inferential step in Bayesian analysis—entails evaluating whether the evidence would be more or less likely under one hypothesis relative to a rival hypothesis. In other words, inference always involves comparing well-specified rival explanations. One common practice we have observed in the nascent Bayesian process tracing literature is to compare a working hypothesis H against its logical negation $\sim H$. This approach is best avoided, because in general, $\sim H$ will be poorly specified—there are multiple ways that a specific working hypothesis could fail to hold, each of which may correspond to a very different "world." As such, assessing the likelihood of the evidence conditional on $\sim H$ may be impossible without first asking what concrete possibilities are contained within $\sim H$. The background information *I,* on which we condition all probabilities, must in practice restrict our attention to a finite set of mutually exclusive hypotheses, such that we can concretely specify $\sim H$ as $\{H_1 \textit{ or } H_2 \textit{ or} ... H_N\}$. It is worth noting that this Bayesian approach departs sharply from frequentist null-hypothesis testing. For an example illustrating the importance of explicitly considering rival explanations rather than attempting to compare H directly against $\sim H$, see Fairfield & Charman (2017, Appendix A: §A4.3).

A second and related caveat is that simply "tracing the process" corresponding to the causal mechanism of the working hypothesis is not adequate for inference.[21] The key question in Bayesian analysis is not whether the evidence fits with a hypothesis, or whether deductive predictions of a hypothesis are born out, but whether the evidence obtained fits *better* with that hypothesis as compared to a plausible rival explanation. In the Stokes (2001) example above, if we had failed to ask how well E_M fits with the rent-seeking hypothesis, we would have significantly overestimated the degree of support that this evidence lends to the representation hypothesis. In some instances, a causal mechanism may be so unique to the working hypothesis that evidence of that mechanism may be largely incompatible with the alternative explanations; nevertheless, correctly assessing the inferential weight of evidence requires explicitly considering how plausible the evidence would be under the rival explanation(s). Likewise, it is important to recognize that a given piece of evidence need not speak directly to the causal mechanism of the working hypothesis in order to boost our confidence in that hypothesis—if this evidence is less likely under the plausible alternative explanations, that evidence favors the working hypothesis.

For the same reasons, stating that a certain collection of observations fits well with the working hypothesis while another set of observations fits poorly with a rival hypothesis departs from Bayesian logic. Our preliminary readings suggest that approach may be fairly common in qualitative and multi-methods research. It is certainly possible that a few key pieces of evidence may dramatically increase the odds in favor of the working hypothesis while one or two other pieces of evidence essentially eliminate a rival explanation; again, however, correctly assessing the inferential weight of any particular piece of evidence requires explicitly asking how likely that same evidence would be in the world of the rival hypothesis. In other words, Bayesian analysis does not entail sorting the evidence into observations consistent with or supportive of one hypothesis and observations consistent with or supportive of another hypothesis. Instead, it entails assessing to what degree each piece of evidence fits with one hypothesis vs. another.

Scrutinizing weight of evidence

Even if a study has correctly applied the Bayesian concept of weight of evidence, scholars may disagree on how strongly the evidence favors a given explanation over rivals. Bayesian analysis provides an excellent framework for scrutinizing evidence that can facilitate consensus building or at least clarifying the locus of disagreements. Open discussion and debate on the weight of evidence can help authors to clarify and communicate their reasoning more clearly, highlight their evidence more precisely, shed light on important elements of context and background information that matter for inferential judgments, and even identify and clarify ambiguities in the hypotheses themselves.

Scrutinizing weight of evidence can be carried out in multiple settings, from informal discussions and research presentations to peer review and published commentary, as well as in the classroom. We are developing teaching exercises to encourage this practice. A first task entails explicitly identifying specific pieces of evidence that authors present, as well as the hypotheses they evaluate. A second task entails assessing and debating the weight of evidence. Students are asked to first analyze the evidence on their own and quantify their assessment of how strongly it favors one hypothesis over another using the decibel scale described in Section 3.1 above. They then discuss and debate their assessments in groups and aim to reach a consensus on the weight of evidence. The discussions we have led illustrate how widely background knowledge varies among scholars and how differently hypotheses and evidence can be evaluated unless both are very clearly specified.

[21] If, however, we have already tentatively accepted a theory as correct and wish to use that theory to explain phenomena that we observe, explicit comparison of rival hypotheses becomes less salient. We suspect that most qualitative research involves a combination of theory comparison and explanation.

These exercises also illustrate that learning how to effectively "mentally inhabit the world" of each hypothesis requires significant practice.

We have already in essence illustrated the process of scrutinizing weight of evidence with the Stokes (2001) example discussed in Section 3.1. By way of further illustration, let us now consider additional information that she presents in her discussion of the Menem case and scrutinize her analysis of its inferential weight.

> E_D = Menem's Minister of Public Works, Roberto Dromi, told Stokes in a 1994 interview: "In this country, 10% of the labor force were government employees. We knew that if we talked of privatizing Aerolíneas Argentinas, we would have the airline workers on our backs, if we talked of privatization, we would have those workers on our backs. ...We knew that Argentines would disapprove of the reforms we planned, but would come to see that they were good."

Stokes (2001:74-75) characterizes this interview information as "strong evidence that policy switchers were sometimes motivated by the belief that the economy would perform much better under efficiency-oriented than under security-oriented policies"—in other words, strong evidence in favor of the representation hypothesis. Stokes' presentation of the evidence in its full and original form (which we have abridged above) and her subsequent clear and careful articulation of how and why it fits with the representation hypothesis are exemplary. Yet a Bayesian scrutiny reveals that the extent to which this evidence supports the representation hypothesis is overstated, because no similar assessment has been conducted regarding how well this evidence fits with the rival rent-seeking hypothesis. For the same reasons discussed regarding the Menem interview (E_M) in Section 2.1, while E_D fits very well with the representation hypothesis, is it also plausible under the rent-seeking hypothesis, because in that world, Dromi might tell a similar story rather than openly revealing the true ignoble motives that drove the policy switch. As with E_M, a Bayesian scrutiny reveals E_D to weakly or perhaps moderately, depending on what background information we possess regarding Dromi's probity—but not strongly—favor the representation hypothesis.[22]

Sensitivity analysis

In qualitative research, subjectivity tends to intrude most strongly at the stage of assessing priors odds on hypotheses, since it is impossible in social science to exhaustively list and carefully consider all of the background information that influences our beliefs about the plausibility of hypotheses. Assessing how much conclusions depend on the choice of priors can therefore be a useful exercise. While weight of evidence will tend to be less vulnerable to subjectivity than priors—because evidence is concrete, specific, and directly observable, whereas hypotheses are abstractions—disagreements over weight of evidence can still be substantial, as discussed above. Sensitivity checks that examine how much inferences change when key pieces of evidence are treated as more or less decisive can therefore be valuable as well.

Fairfield & Charman (2017: Appendix A) include detailed illustrations of both types of sensitivity checks, drawing on Fairfield's (2013, 2015) case study of a Chilean reform that revoked a regressive tax subsidy. Fairfield argues that an "equity appeal," made during a presidential race in which inequality had assumed unusually high issue salience, drove the right-wing opposition coalition to approve the reform in congress in order to avoid electoral punishment. During the campaign, the opposition's presidential candidate blamed Chile's persistent inequality on the governing center-left coalition. The incumbent president responded by linking the tax subsidy to inequality and publicly challenging the opposition to support the reform.

[22] Dromi is later revealed to have been implicated in corruption scandals involving privatizations.

Fairfield & Charman's Bayesian analysis of this case study compares the equity-appeal hypothesis against three rival explanations, which include a basic median-voter hypothesis, using three different prior probability distributions:

> (1) equal odds on each hypothesis, which avoids any initial bias,
>
> (2) a 50 dB penalty on the median-voter hypothesis, in light of extensive literature that questions its logic (e.g. Hacker and Pierson 2010, Kaufman 2009), with equal probabilities on the other three hypotheses, and
>
> (3) a 50 dB penalty on Fairfield's explanation with equal priors for the remaining three.

The last of these prior distributions is particularly salient, given that an earlier version of Fairfield (2013) received comments from a reviewer who deemed the equity appeal explanation implausible in light of background knowledge about the ineffectiveness of presidential appeals in US politics. The analysis in Fairfield & Charman (2017: Appendix A) illustrates that even with this heavy 50 dB penalty on the equity appeal hypothesis, it emerges as the best explanation in light of the evidence by at least 65 dB. Of course, skeptical readers might also question the assessment of the weights of evidence. The final section of Appendix A shows that even if we substantially reduce the extent to which we judge each piece of evidence to favor the equity appeal explanation over the rivals, the aggregate inference still weighs in favor of this hypothesis.

To give an example of how and when sensitivity analysis with respect to weight of evidence may be useful, consider the following piece of evidence (Fairfield 2013:49):

> E_{FM} = A finance ministry official told the investigator that the tax subsidy "was a pure transfer of resources to rich people... It was not possible for the right [coalition] to oppose the reform after making that argument about inequality."

As we will see, the weight we attribute to this evidence depends significantly on our background information. First consider the world of the equity-appeal hypothesis. E_{FM} fits well with the causal mechanism of this hypothesis in highlighting the importance of the exchange between the opposition candidate and the president that led to the equity appeal. Because E_{FM} makes the government appear savvy and effective at achieving socially desirable goals while insinuating that the opposition would otherwise have resisted redistribution, we see little reason for the government to conceal this information if the equity-appeal hypothesis is correct. Next consider an alternative world where a basic median-voter logic operates—the opposition coalition accepted the reform because electoral competition drives politicians to converge on policies that promote the median voter's material interests,[23] and the equity appeal played no relevant causal role. While E_{FM} is less expected in this world relative to the equity-appeal world, it is not overly surprising. Government informants could have incentives to incorrectly attribute the opposition's support for the reform to the equity appeal, because this story portrays the government in a positive light and the opposition in a negative light. Yet Fairfield's background information includes significant confidence in this informant's expertise, analytical judgment, and sincerity, based on multiple interactions during fieldwork. Balancing these considerations, we judge E to favor the equity-appeal hypothesis by 10 dB. However, a skeptical reader who possesses a very different body of background knowledge might be reluctant to share the author's confidence in the informant. To acknowledge this possibility, we can also report the overall weight of evidence presented in the case study[24] if we take E_{FM} to support the equity-appeal hypothesis over the median-voter hypothesis by a more conservative 5 dB (the same weight we attributed to Menem's interview in section 3.1 above).

---

[23] E.g., Meltzer and Richard (1981).

[24] The case study includes five additional key pieces of evidence.

Consistency checks

As emphasized previously, the key inferential step that tells us how to update the prior odds on our hypotheses entails evaluating the weight of evidence, where E stands for all relevant information obtained over the course of the study. The mathematics of logical Bayesianism allow for a number of consistency checks on our reasoning about the overall weight of this evidence. For instance, we can directly assess the likelihood of the joint evidence E, or we can instead decompose E into any number of constituent observations. If we opt for the latter approach, we are free to incorporate the constituent observations into our analysis in any order. To see why, we need only apply the product rule of probability and the commutativity of conjunction:

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

Equation (4) can be easily generalized if we wish to decompose E_1 and/or E_2 into additional constituent observations. It follows that likelihood ratios can be factored into products:

$$\frac{P(E|H_i\ I)}{P(E|H_j\ I)} = \frac{P(E_1|E_2\ H_i\ I)}{P(E_1|E_2\ H_j\ I)} \times \frac{P(E_2|H_i\ I)}{P(E_2|H_j\ I)} = \frac{P(E_2|E_1\ H_i\ I)}{P(E_2|E_1\ H_j\ I)} \times \frac{P(E_1|H_i\ I)}{P(E_1|H_j\ I)}, \quad (5)$$

and weights of evidence are additive:

$$\begin{aligned} WoE\ (H_j : H_k) &= WoE_1\ (H_j : H_k\ ,\ E_2) + WoE_2\ (H_j : H_k) \\ &= WoE_2\ (H_j : H_k\ ,\ E_1) + WoE_1\ (H_j : H_k)\ , \end{aligned} \quad (6)$$

with the important caveat that each piece of evidence that has already been incorporated into our analysis becomes part of the background information that we use to assess the weight of evidence for the next piece of evidence we consider.

Consistency checks that involve analyzing pieces of evidence in different orders are salient if there are potential logical dependencies among the evidence under a given hypothesis. Notice that assessing $P(E_1|E_2HI)$ entails asking: given the hypothesis and the background information, is E_1 any more or less likely given that we also know E_2? In other words, we must condition on previously incorporated information,[25] which can be a very difficult task. Given the technical subtleties and analytical challenges involved, we refer interested readers to Fairfield & Charman (2017: §3.3.3-4); Appendix A (§A6) of that article provides a fully worked example.

As an example of consistency checks that entail parsing the evidence at different levels of aggregation, consider the following example based on Kurtz's (2013) research on state building in Latin America. We wish to compare a resource curse hypothesis H_R against a warfare hypothesis H_W (assumed mutually exclusive), in light of three salient observations $E=E_{WS}\ E_{ME}\ E_{IC}$ about the case of Peru:

> H_R =Abundant mineral rents hinder institutional development. Easy money undermines administrative capacity by precluding the need to collect taxes, and public resources are directed toward inefficient industries, consumer subsidies, and patronage networks that sustain elites in power.
>
> H_W =Absence of warfare hinders 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.

[25] Previously incorporated information need not be collected at an earlier point in time relative to subsequently incorporated information.

E_{WS} = Peru never developed an effective state.

E_{ME} =Peru's economy has been dominated by mineral exports since colonial days.

E_{IC} = Peru was among the Latin American countries most consistently threatened by international military conflict.

One approach would be to directly assess the overall weight of evidence E. Taken together, the three observations strongly favor H_R over H_W. In a world where H_R is the correct hypothesis, mineral dependence in conjunction with weak state capacity are 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, would have to 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. Accordingly, we might assign $WoE\ (H_R : H_W) = 50$ dB, roughly corresponding to the difference between a whisper and a noisy restaurant.

A second approach could entail assessing the three pieces of evidence separately:

$$WoE\ (H_R : H_W) = WoE_{WS}\ (H_R : H_W) + WoE_{ME}\ (H_R : H_W,\ E_{WS}) + WoE_{IC}\ (H_R : H_W,\ E_{WS}E_{ME})$$

We evaluate each term in turn. First, we have $WoE_{WS}\ (H_R : H_W) = 0$. Assuming we have no salient background information about the Peruvian case from the outset, a weak state is no more or less surprising under either hypothesis. Second, we judge $WoE_{ME}\ (H_R : H_W,\ E_{WS}) = 12$ dB. Mineral export dependence is exactly what we would expect under H_R given state weakness. In contrast, E_{ME} is neither surprising nor expected under H_W; the warfare hypothesis makes no predictions about Peru's economy in the absence of any relevant background information about this case. Accordingly, this evidence moderately favors the resource-curse hypothesis. Third, we deem $WoE_{IC}\ (H_R : H_W,\ E_{WS}\,E_{ME}) = 45$ dB. In a world where H_W is correct and we have a weak state, international military conflict would be highly surprising; it contradicts the expectations of the theory, which would be absence of warfare. In contrast, this evidence is moderately expected under the rival hypotheses, because a weak state with natural resources would be an easy and attractive target for invaders. The evidence therefore very strongly favors H_R over H_W. Adding the three weights of evidence, we find $WoE\ (H_R : H_W) = 57$ dB, roughly corresponding to the difference between a typical conversation and a live rock concert.

The difference between the weight of evidence we assigned via these two approaches is small, in both absolute and relative terms. However, it is a noticeable amount—7 dB. In this case, we would judge the second approach, where we considered each of the three pieces of evidence in turn, to be more reliable. If we found a more substantial difference between the overall weight of evidence assessed through our two logically equivalent approaches, we would need to reconsider our analysis and look for the inconsistency in our reasoning.

A third type of consistency check can be relevant if we are considering three or more rival hypotheses. For three hypotheses, any two weights of evidence determine the third weight of evidence, since:

$$WoE\ (H_1 : H_2) + WoE\ (H_2 : H_3) + WoE\ (H_3 : H_1) = 0\ , \qquad (7)$$

where our sign conventions dictate assigning negative decibels to $WoE\ (H_i : H_j)$ if the evidence favor H_j over H_i.[26] Equation (6) follows directly from the fact that:

[26] Equation 6 is easily generalized for additional hypotheses.

$$\frac{P(E|H_1\ I)}{P(E|H_2\ I)} \times \frac{P(E|H_2\ I)}{P(E|\ H_3\ I)} \times \frac{P(E|H_3\ I)}{P(E|H_1\ I)} = 1 \quad (8)$$

As an illustration of the cross-checks this relationship facilitates, consider adding a third rival hypothesis to be considered alongside H_R and H_W in our previous example:

> H_{LRA} = Labor-repressive agriculture hinders 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 enforce "the strict local social control and labor coercion upon which the agrarian political economy, and potentially their physical survival, depends." Instead, elites seek to maintain their own security forces with which to repress their local labor forces. (Kurtz 2013)

We will focus on E_{IC}. We have already judged that WoE_{IC} ($H_R : H_W, E_{WS}E_{ME}$) = 45 dB. By essentially identical logic, we would also judge that WoE_{IC} ($H_{LRA} : H_W, E_{WS}E_{ME}$) = 45 dB—while international conflict is highly surprising under H_W, the evidence is moderately expected under H_{LRA}, because a weak state with natural resources would be an easy and attractive target for invaders. By equation (6), we must then have WoE_{IC} ($H_{LRA} : H_R, E_{WS}E_{ME}$) = 0 dB. As a cross-check, we can mentally inhabit the respective worlds of H_{LRA} and H_W ask whether this last probability assignment makes intuitive sense. Indeed it does, since our same logic implies that regardless of whether labor repressive agriculture or the resource curse gave rise to state weakness, we are not surprised that a weak state with mineral resources would experience military threats.

Suggestions for Moving Forward

How can we promote more reliable inference and better same-data scrutiny, continued research, and extended research in qualitative scholarship? In this volume, Jacobs advocates pre-registration as a useful tool for improving qualitative research; we take the opposite view and argue that pre-registration is ill-suited to qualitative research that follows the principles of Bayesian inference (Fairfield & Charman 2019). Pre-registration can play an important role in frequentist inference, which 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. In contrast, logical Bayesianism imposes built-in safeguards against confirmation bias—by requiring us to consider how well the evidence fits not just with our favored working hypothesis, but also with rival hypotheses—and against *ad-hoc* theorizing—via "Occam factors" that automatically penalize the prior probability of overly-complex hypotheses if they do not add sufficient explanatory leverage relative to simpler alternatives. In lieu of pre-registration, this final section offers some suggestions for improving reliability in qualitative research that focus on disciplinary norms and training.

Disciplinary norms

We envision two key fronts on which disciplinary norms could be usefully adjusted for the sake of promoting more reliable inference. First, altering publication norms regarding negative results and requisite levels of confidence in findings could go a long way toward mitigating incentives for falsely bolstering results. In our view, a replication crisis can arise only to the extent that scholars overestimate or mischaracterize the degree of confidence that is justified in light of the evidence. 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"—which is firmly grounded in Bayesian reasoning—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 help signal integrity and encourage critical thinking. In the messy world of social science, where causal complexity is the rule, it is implausible that a single hypothesis will outperform every considered alternative with respect to every piece of evidence. If no countervailing facts are mentioned, concern is warranted that the author has omitted or not looked hard enough for contrary evidence.

Second, we concur with others in this volume (Freese & Peterson, Elman, Gerring, & Mahoney) that measures should be taken to counteract publication bias toward counterintuitive findings. Within a Bayesian framework, a counterintuitive explanation should be penalized from the outset with a low prior probability; extraordinary claims require extraordinary evidence. Confirming accepted theories or validating common-sense expectations can make a positive contribution to scholarship, although editorial judgment is still required in distinguishing useful solidification of knowledge vs. rehashing known facts or restating the obvious.

Bayesian Training

In reflecting on the replication crisis, Andrew Gelman has in part blamed conventional statistical discourse for fostering the misconception that the purpose of statistical testing and modeling is to somehow distill certainty from uncertainty. Instead, the goal is to confront, assess, manage, and communicate the uncertainty that inevitably remains after evidence is collected and analyzed. Bayesianism, which is becoming more widely used in quantitative as well as qualitative research, provides a natural language for talking about uncertainty in all its guises. Whenever we assess the posterior probability that a hypothesis is true in light of the evidence, we are effectively communicating how much uncertainty we believe surrounds the inference.

Accordingly, training in Bayesian probability could be highly beneficial for qualitative research. On the one hand, learning the basics of Bayesian inference can help improve and leverage intuition. Bayesianism mirrors the way we naturally approach inference in qualitative research, yet it also helps to avert cognitive biases that might unwittingly lead to faulty conclusions (Fairfield & Charman 2019). On the other hand, Bayesianism provides a clear framework for scrutinizing inference, pinpointing the sources of disagreement among scholars, building consensus on the strength of inferences, and learning from accumulated knowledge. Bayesianism highlights the importance of, and provides concrete rules for scrutinizing the weight of evidence, assessing whether authors' (implicit) priors are well justified by the background information presented and ascertaining how sensitive conclusions are to the choice of priors, and conducting consistency checks to see if the conclusions follow when analyzing evidence in different but logically equivalent ways. Bayesian probability need not necessarily be applied explicitly and formally in qualitative research. Informal or heuristic Bayesian reasoning (especially with regard to the weight of evidence) and/or more formal Bayesian scrutiny of selected case examples and/or critical pieces of evidence alongside traditional case-study narratives can still have substantial beneficial effects on the quality, transparency, and reliability of inference.

References

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. 2015. "Disciplining Our Conjectures: Systematizing Process Tracing with Bayesian Analysis." In Andrew Bennett and Jeffrey Checkel, eds, *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool.* Cambridge University Press, 276–98.

Berger, James, and Donald Berry. 1988. "Statistical Analysis and the Illusion of Objectivity," *American Scientist* (March-April):159-165.

Boas, Taylor. 2016. *Presidential Campaigns in Latin America: Electoral Strategies and Success Contagion.* New York: Cambridge University Press.

Centeno, Miguel. 2002. *Blood and Debt: War and the Nation State in Latin America.* University Park, PA: Pennsylvania State University Press.

Collier, Ruth Berins, and James Mahoney. 1997. "Adding Collective Actors to Collective Outcomes: Labor and Recent Democratization in South America and Southern Europe." *Comparative Politics* 29(3):285-303.

Collier, David, Henry Brady, and Jason Seawright. 2004. "Sources of leverage in causal inference: Toward an alternative view of methodology," in Henry Brady and David Collier, ed.s, Rethinking Social Inquiry: Diverse Tools, Shared Standards, Rowman & Littlefield Publishers.

Coppedge, Michael, and John Gerring, with David Altman, Michael Bernhard, Steven Fish, Allen Hicken, Matthew Kroenig, Staffan I. Lindberg, Kelly McMann, Pamela Paxton, Holli A. Semetko, Svend-Erik Skaaning, Jeffrey Staton, and Jan Teorell. 2011. "Conceptualizing and Measuring Democracy: A New Approach," *Perspectives on Politics* 9, (2, June):247-267.

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

Elman, Colin, John Gerring, and Jim Mahoney. Introduction to The Production of Knowledge: Enhancing Progress in Social Science.

Ertman, T. 1997. "The birth of the Leviathan." New York: Cambridge University Press.

Etz A, Vandekerckhove J. 2016. "A Bayesian Perspective on the Reproducibility Project: Psychology." PLoS ONE 11(2): e0149794. doi:10.1371/journal.pone.0149794

Fairfield, Tasha. 2013. "Going Where the Money Is: Strategies for Taxing Economic Elites in Unequal Democracies." *World Development* 47: 42–57.

———. 2015. *Private Wealth and Public Revenue in Latin America: Business Power and Tax Politics.* Cambridge University Press.

Fairfield, Tasha, and Andrew Charman. 2017. "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats." *Political Analysis* 25 (3): 363-380.

Fairfield, Tasha, and Andrew Charman. 2018. "Bayesian Perspectives on Case Selection and Generalization." Presented at the American Political Science Association Annual Meeting, Boston, MA, August 30–September 2.

Fairfield, Tasha, and Andrew Charman. 2019. "A Dialogue with the Data: The Bayesian Foundations of Iterative Research in Qualitative Social Science." *Perspectives on Politics* 17(1):154-167.

Fairfield, Tasha, and Candelaria Garay. 2017. "Redistribution under the Right in Latin America: Electoral Competition and Organized Actors in Policymaking," *Comparative Political Studies,* 50 (14): 1871-1906.

Freese, Jeremy, and David Peterson. "Replication." *In this volume.*

Freese, Jeremy and David Peterson. (2016) The Emergence of Forensic Objectivity. osf.io/2ft8x.

Garay, Candelaria. 2016. *Social Policy Expansion in Latin America.* New York: Cambridge University Press.

Goodman, Steven, Daniele Fanelli, and John Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8(341):1-6.

Greenland, Sander. 2006. "Bayesian Perspectives for Epidemiological Research," *International Journal of Epidemiology* (35):765-775.

Hacker, Jacob, and Paul Pierson 2010. "Winner-Take-All-Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States." *Politics and Society* 38(2):152–204.

Haggard, Stephan, and Robert Kauffman. 2012. "Inequality and Regime Change: Democratic Transitions and the Stability of Democratic Rule," *American Political Science Review* (August):1-22.

Herbst, Jeffrey. 2000. *States and Power in Africa: Comparative Lessons in Authority and Control.* Princeton University Press.

Hui, Victoria. 2005. *War and State Formation in Ancient China and Early Modern Europe.* New York: Cambridge University Press.

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

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

Pereira, Tiago, and John P.A. Ioannidis 2011. "Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects," *Journal of Clinical Epidemiology* 64(10):1060-69.

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

Jacobs, Alan. "On Bias and Blind Selection: Assessing the Promise of Pre-registration and Results-Free Review," *this volume.*

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

Kapiszewski, Diana. 2012. *High Courts and Economic Governance in Argentina and Brazil.* New York: Cambridge University Press.

Karl, Terry. 1997. *The Paradox of Plenty.* Berkeley: University of California Press.

Kaufman, Robert. 2009. "The Political Effects of Inequality in Latin America: Some Inconvenient Facts." *Comparative Politics* 41(3):359–79.

Kurtz, Marcus. 2013. *Latin American State Building in Comparative Perspective.* New York: Cambridge University Press.

Meltzer, Allan, and Scott Richard. 1981. "A Rational Theory of the Size of Government." *Journal of Political Economy* 89(5):914–27.

Slater, Dan. 2010. *Ordering Power: Contentious Politics and Authoritarian Leviathans in Southeast Asia.* New York: Cambridge University Press.

Stokes, Susan. 2001. *Neoliberalism by Surprise: Mandates and Democracy in Latin America.* Cambridge University Press.

Tilly, Charles. 1992. *Coercion, Capital, and European States, AD 990–1992.* Oxford, UK: Blackwell.

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

Van Aert RCM, Van Assen MALM. 2017. "Bayesian Evaluation of Effect Size after Replicating an Original Study." PLoS ONE 12(4): e0175302. https://doi.org/10.1371/journal.pone.0175302

Figure 1. Reliability of Inference

		Quantitative Research		Qualitative Research
		Frequentist	***Bayesian***	***Bayesian***
Same Data	*Same procedures*	← **Assessing coding and measurement** →		
		← **Verification** →		
		Checking that reported results can be generated from the data used and frequentist procedures described	Checking that reported results can be generated from the data used and Bayesian procedures described	Checking that reported results follow from correct application of Bayesian reasoning **Scrutinizing weight of evidence** Assessing likelihood ratios, which determine how strongly the evidence favors one hypothesis relative to rivals
	Different procedures	← **Sensitivity analysis** →		
		Altering model specification and assumptions (e.g. error model) [a]	Altering priors and/or changing likelihood function	Altering priors and/or likelihood ratios **Consistency checks** Making sure the same conclusions follow when analyzing evidence in different but logically equivalent ways
New Data	*Same procedures, Specific results*	**Replication** →		
		Repeating the original random sampling procedure to generate new data from the same population [b]	Repeating the same experiment under conditions as close as possible to the original research design	
			Continued research →	
			Gathering additional evidence to combine with existing evidence using the same research design	
	Different procedures, Broader findings	← **Extended research** →		
		[c]	Assessing theory and generalizing findings to other populations, cases or contexts; Refining theory in new contexts	

Notes:
[a] Clemens' (2017:327) "reanalysis tests" fit here.
[b] Our definition of replication corresponds to Clemens' (2017:327) "reproduction test," which entails "sampling precisely the same population but otherwise using identical methods to the original study."
[c] We would place Clemens' (2017:327) "extension tests"—which use data "gathered on a sample representative of a different population,"—in this category. Although these tests use essentially the same statistical analysis as the original research, sampling from a different population changes the research design and aims to assess generalizability of findings.