Caveats." *Political Analysis* 25, no 3 (July): 363-380.

Hall, Peter A. 2003. "Aligning Ontology and Methodology in Comparative Politics." In *Comparative Historical Analysis in the Social Sciences*, edited by James Mahoney and Dietrich Rueschemeyer, 373–404. New York: Cambridge University Press.

Jackson, Patrick T. 2016. *The Conduct of Inquiry in International Relations: Philosophy of Science and Its Implications for the Study of World Politics*. 2nd ed. London: Routledge.

Mayne John. 2012. "Contribution Analysis: Coming of Age?" *Evaluation* 18, no. 3 (July): 270-280.

Møller, Jørgen, and Svend-Erik Skaaning. 2018. "The Ulysses Principle: A Criterial Framework for Reducing Bias When Enlisting the Work of Historians." *Sociological Methods & Research* 1-32 doi.org/10.1177/0049124118769107.

Schmitt, Johannes, and Derek Beach. 2015. "The Contribution of Process-Tracing to Evaluations of Budget Support Programs." *Evaluation* 21, no. 4 (October): 429-447.

Scriven, Michael. 2008. "A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research." *Journal of MultiDisciplinary Evaluation* 5, no. 9 (March): 11–24.

Tavory, Iddo, and Stefan Timmermans. 2014. *Abductive Analysis: Theorizing Qualitative Research*. Chicago: University of Chicago Press.

Qualitative and Multi-Method Research Fall 2020, Vol. 18, No. 2 https://doi.org/10.5281/zenodo.4046614

A Bayesian Perspective on Theory-Blind Data Collection

Tasha Fairfield
London School of Economics

Copestake, Goertz, and Haggard's (CGH) "Veil of ignorance Process Tracing" (VPT)—which in essence entails placing a firewall between data collection and hypothesis testing[1]—is an interesting addition to a growing list of proposals made in recent years that aim to address potential sources of bias in qualitative social science. Many of these proposals (e.g., pre-registration, time-logging whether evidence was discovered before or after a hypothesis was devised) import prescriptions from large-N, frequentist, statistical research that, from a Bayesian perspective, are not applicable to qualitative research. Bayesian reasoning provides its own safeguards against the problems of confirmation bias and ad hoc hypothesizing, without imposing procedural constraints that would interfere with the inherently iterative, dynamic, and interactive nature of case-study research—where we go back and forth between hypothesizing, data collection, and analysis (Fairfield and Charman 2019).

My comments begin by outlining the costs (which seem significant) and gains (limited, in my analysis) of firewalled data collection in qualitative research. I then discuss what I interpret as a fundamental shortcoming with the authors' approach that seems to undermine its core aim of separating data collection from hypothesis testing—namely, conflating evidence, evidentiary sources, and causal claims. Finally, I briefly outline my preferred approach for managing the problems of confirmation bias, ad hoc hypothesizing, and cherry-picking.

Scrutinizing the Costs and Benefits of Firewalled Data Collection

As with suggestions for pre-registration or time-logging evidence relative to hypothesis generation, firewalled data collection runs counter to the way that qualitative research is generally conducted. Instead of proceeding linearly from theory generation to data collection to theory testing, we naturally engage in a "dialogue with the data," (Bayesian astrophysicist Stephen Gull, quoted in Sivia (2006)) where we go back and forth between theory and evidence. We revise and refine theory in light of the data, and we revisit the evidence in light of new ideas and new theory, analyzing the information differently or more deeply, asking new questions, and deciding what kinds of additional data to collect.

Firewalled data collection would come at a significant cost of precluding an effective dialogue with the

1 As CGH (this issue) write: "The reference to 'veils of ignorance' arises from a division of labor that allows a research assistant to carry out key data selection and coding tasks without knowledge of the theories, hypotheses, and mechanisms being tested by the principal."

data, where scholars can adjust the research strategy along the way when the evidence uncovered suggests new hypotheses to investigate and new sources of information to pursue. If the research assistants (RAs) in charge of gathering evidence are completely blind to the initial theories under consideration, and potentially (as CGH suggest) even blind to the details of the research question itself, they can hardly make the informed analytical decisions that are needed while in the field, or visit the archives or scrutinize secondary literature, to be able to collect evidence that will bear substantial inferential weight when it comes time for theory evaluation. From a Bayesian perspective, strong evidence is information that discriminates between competing theories—in colloquial terms, we seek clues that fit much better with one hypothesis compared to a rival. If we keep alternative explanations in mind while we gather our evidence, we are able to look for the kinds of clues that we expect to be most informative, whereas if we deliberately ignore theory throughout the process, we may well end up with a sub-optimal dataset filled with weak or irrelevant information that does not effectively allow us to adjudicate between alternative explanations. It also bears emphasis that "selecting sources," which is one of CGH's central concerns, is only one component of searching for evidence. We must also work hard to extract useful information from the sources we consult, which entails asking informants the right questions, or knowing how to spot salient exchanges in congressional records or relevant details in news accounts. Again, the decisions we make while collecting or generating evidence should be guided by our evolving ideas about the plausible alternative explanations to be assessed.

Yet, while imposing potentially substantial costs in terms of the quality of evidence obtained, firewalled data collection does not go very far toward solving the potential problems that the authors aim to address. This approach does preclude confirmation bias *at the data collection stage*—RAs cannot seek out only evidence that supports a pet theory if they have no knowledge whatsoever regarding the theories that are to be evaluated. But confirmation bias can just as easily occur *after data has been collected*, when it comes time for data analysis and theory testing. In fact, a commonly discussed form of confirmation bias entails overestimating the extent to which a given piece of evidence supports a favored hypothesis, often by forgetting to ask whether that evidence might be equally or even more consistent with a rival hypothesis. Likewise, ad hoc hypothesizing entails over-fitting an explanation to the particular details of the evidence at hand; as such this problem usually arises *after* the data have been collected.

Furthermore, firewalled data collection in and of itself does not preclude "cherry-picking"—which I understand to mean deliberately ignoring evidence that does not support a favored hypothesis. Dishonest scholars can always find ways to be dishonest, regardless of whatever constraints are imposed by the discipline. As Ansell and Samuels observe regarding the related suggestion of results-blind review, it is always possible to "sweep dirt an author wants no one to see under a different corner of the publishing carpet" (Ansell and Samuels 2016, 1810). In the instance at hand, one could easily imagine scenarios where a research team purports to follow Veil of ignorance Process Tracing but finds ways to "cheat" that would be difficult for reviewers to uncover. For example, the principal researcher (PR) manages to subtly communicate a pet theory to the RAs or the PR "accidentally" deletes unfavorable evidence from the dataset after the fact while ensuring that the RAs are unaware or have incentives to stay quiet.

Moreover, ensuring that the PR consults a wide range of *sources* that have been selected without bias by theory-blind RAs does nothing to address what I view as the more serious problem of cherry-picking *from those sources* only those pieces of information that support the PR's pet theory.[2] Indeed, biased extraction of evidence from the sources consulted seems to be a more common problem than biased selection of the sources themselves.[3] It is often the case that a document or an informant provides some information that supports one hypothesis, but also other pieces of information that favor a rival hypothesis, so there is ample opportunity for dishonest or sloppy scholars to include the former while omitting the latter. Meanwhile, identifying bias in the sources consulted is a relatively straightforward matter—conscientious reviewers scrutinize bibliographies to check if important sources have been ignored or overlooked. If, for example, I had interviewed only informants from Chile's center-left government in my research on taxation, without talking to anyone from the right-wing opposition or the business sector, I expect that other scholars would have noticed these omissions and that my work would have been much less favorably received.

2 I thank Stephan Haggard for a useful email exchange related to these points.

3 Accordingly, many transparency advocates have focused on making interview transcripts (i.e., the content of what the sources have said) publicly available—which poses a distinct set of tradeoffs.

Conflating Evidence and Causal Claims

Setting aside the above critique regarding the substantial costs and limited benefits of VPT, the central point of this approach as I understand it is to separate data collection from hypothesis testing. Yet in my reading, the authors' discussion seems to conflate evidence with causal claims, data collection with data analysis, and theory building with theory testing—which in turn renders their approach problematic.

I first define my usage of some basic concepts. A *hypothesis* is a proposition that makes a causal claim about the way the world works. Well-stated hypotheses generally include some causal mechanism(s) that explain how and why the outcome of interest occurs. *Evidence* is any concrete information we learn that can be used to test our hypotheses. A *causal inference* is a conclusion that we draw about the truth of our hypotheses after analyzing them in light of this evidence.

The crux of the problem in CGH's discussion is the notion that data collection entails "extracting causal claims" from sources. Data collection in qualitative research entails gathering *evidence*, which in and of itself does not make "causal claims"—*hypotheses* articulate causal claims. As I will explain below, evidence can sometimes entail *a particular source asserting a causal claim*, but this is not at all the same as the causal claim itself, and moreover, many kinds of evidence do not fit this mold at all.[4]

My best effort to make sense of what it means for theory-blind RAs to "extract" and "code" various "causal claims" from sources is to re-interpret this task in terms of crowdsourcing hypotheses—the RAs talk to informants or examine written materials for possible explanations of whatever the PR is investigating. This would of course seem problematic considering that the RAs are supposed to have minimal (if any) knowledge of the research topic, but let us set that difficulty aside for now. The crowdsourcing interpretation seems to be substantiated by CGH's (this issue) instructions that the RAs should code, for example, whether a variable in a "causal claim" is necessary or sufficient, whether there are interaction terms or mediators, what is the strength of the proposed causal relationship, and so forth—these tasks all fall squarely in the realm of *hypothesis generation*.[5]

But what does the PR do with these crowd-sourced hypotheses (or using different language, the theories or causal mechanisms that emerge inductively from the sources)? Would the PR place these new hypotheses into the vault, and then proceed through another round of firewalled data collection?[6] This second round would have to involve something other than extracting "causal claims"—we need *evidence* with which to *test* the causal claims. The authors instead assert that "the end product...is a causal mechanism figure that synthesizes and makes sense of the causal claim data" (this issue). But I would emphasize that articulating a causal mechanism (whether verbally, graphically, or in combination) is an exercise in *theory generation,* not theory testing. Even within a Bayesian framework, which precludes any need for firewalls between theory building and theory testing, these are conceptually distinct stages of analysis, and inference does not entail simply proposing a theory that "makes sense" of the data—rival theories must be pitted against each other in light of all available evidence that speaks to their plausibility.

The authors nevertheless claim that they are simultaneously "evaluating" theory, but in my view, they do not articulate a sound methodology for that purpose. They suggest that theory evaluation proceeds by somehow examining the "incidence of causal claims" coded by the RAs (this issue)—perhaps thinking that the PR's hypothesis gains support commensurate with the proportion of "causal claims" that match the hypothesis. But this notion of inference violates a core principle of Bayesian reasoning—evidence is to be weighed, not counted. Weighing the evidence requires asking which of one or more rival hypotheses makes the evidence more expected. Evidence that is consistent with a given hypothesis does not necessarily support that hypothesis, because the evidence could be *even more compatible* with a rival hypothesis. Accordingly, it is not enough to simply "trace out"

4 One might ask whether the authors have *causal inferences* in mind rather than evidence when they speak of "causal claims," but then VPT would delegate not just data collection but also *data analysis* to RAs, who by design are not equipped to complete that task, for lack of any knowledge about theory or even the research question—so I set that possible interpretation aside.

5 This interpretation is further substantiated by CGH's discussion of Bennett and Checkel's (2015) best practices (which are in large measure Bayesian inspired). Bennett and Checkel call for scholars to cast the net widely for alternative *explanations* (e.g., causal hypotheses); CGH respond by discussing selection of *sources*—which then presumably supply the hypotheses. And in addressing Bennett and Checkel's call for inductive insights, CGH again seem to confirm my interpretation that their "causal claims" are indeed new hypotheses (i.e., features of causal mechanisms that were not thought of before)—not evidence with which to test hypotheses.

6 Alternatively, if VPT allows the PR to freely engage in a second round of data collection after a new theory has come to mind, without delegating once again to theory-blind RAs, then there would be little point in imposing a veil of ignorance during the first round of data collection. Any supposed benefits from an initial round of fire-walled data collection in terms of reducing bias would be easily undermined if the firewall is lifted during subsequent rounds of data collection.

a causal process—we must also ask whether any "causal process evidence" is more or less expected under an alternative hypothesis.

In sum, we are left with a procedure that at its core seems perplexing. Rather than collecting evidence, RAs appear to be crowd-sourcing hypotheses, which simply generate more theory that needs to be evaluated in light of actual evidence. Within a Bayesian framework, it is perfectly possible to devise new hypotheses inspired by the evidence and then use that same evidence to evaluate the hypotheses, as CGH seem to want to do, but my co-author and I advance the strong claim that Bayesianism is the only self-consistent inferential framework that can justify this practice (Fairfield and Charman 2019). The notions of theory evaluation that CGH seem to espouse depart sharply from Bayesianism in some critical regards.

Within a Bayesian framework, the key to resolving the underlying confusion about data and "causal claims" is to carefully handle *testimonial evidence*—information we receive from fallible human sources, who may have incomplete knowledge of the topic at hand as well as instrumental motives to exaggerate, obfuscate, or dissemble. When we interview an informant or read an account in a newspaper archive, the surface content (X) of what that particular source has said—which may well be a particular causal story—does *not* constitute the evidence (E) that we use to evaluate our hypotheses. Instead, the evidence must take the following form: E = *source S made statement X in context C*. In the example from Fairfield's research on tax reform that CGH quote, we have E = "Governing-coalition informants (i.e., *source S*) told Fairfield in an interview (i.e., *context C*) that '... the measure was ruled out as infeasible on every occasion due to resistance from the right-wing coalition' (i.e., *statement X*)" (Fairfield 2015, 122).

Formulating testimonial evidence in this manner is critical for assessing its inferential import, which in Bayesian terms comes from evaluating its likelihood under alternative hypotheses. In the tax reform example, we must ask whether it would be more expected for the government informants (S) to tell Fairfield this particular story (X) about the tax reform in question if Fairfield's "equity-appeal hypothesis" is correct, or whether it would be more expected to hear the informants tell Fairfield this story if the rival median-voter hypothesis is correct. As part of this reasoning process, we must assess the distinct incentives that the informants could have to reveal or distort the truth in the world of each respective hypothesis (see Fairfield and Charman 2017 for details). Critically, the inferential weight of this evidence does not come from treating a causal story that the informant has articulated (X) as an "instance" of the equity-appeal hypothesis, to be tallied up against distinct "causal claims" made by other informants.[7]

It is also important to emphasize that evidence does not consist exclusively of sources articulating their understandings of a causal process, nor need it fall into the more general testimonial category explained above (where a source S may make some other kind of statement X). Evidence also includes well-established facts that on their own do not express any kind of "causal claim" (e.g., *the reform did not pass until 2005*, or *suspect A was out of town when the murder was committed*). Nor does evidence necessarily constitute or suggest a "link" or component in some theorized causal mechanism or graph. The inferential import of these facts or observations emerges once again by asking whether they are more likely in the world of one hypothesis compared to a rival. For instance, if H_A = *suspect A acting alone killed the victim with an ice pick,* and H_B = *suspect B acting alone killed the victim with an ice pick*, then information that A was out of town at the time weighs very strongly in favor of H_B vs. H_A, without speaking at all to the manner in which suspect B committed the murder

Understanding these distinctions between evidence, sources, causal claims, and hypotheses should help clarify the importance of having well-informed scholars conduct data collection. Data collection is not simply a matter of selecting sources, nor does it entail simply coding "causal claims" that emerge from those sources. Soaking and poking is valuable—we do not need to have our theories completely formulated and engraved in stone beforehand. But whoever is in charge of data collection should be closely acquainted with the goals of the research, and they should be familiar with existing hunches about competing hypotheses so that they can recognize and search for strongly discriminating evidence during a dynamic data-gathering process, and thereby be in a position to effectively pursue new leads when new hypotheses or new sources of information come to mind.

7 Note also that our assessment of the truthfulness of the testimony may vary across the different hypothesized worlds. A statement made by a particular source may be truthful conditional on one hypothesis but may necessarily be mistaken or mendacious under a different hypothesis, so this is not something that could be coded by an RA who is ignorant of the hypotheses in question.

An Alternative Suggestion: Promote Bayesian Reasoning and Standards of Integrity

Given that firewalled data collection seems to entail significant costs with limited benefits, even if the problems with CGH's particular approach were redressed, how should we best seek to curtail the potential problems of confirmation bias, ad hoc hypothesizing, and cherry-picking in qualitative research? Following Fairfield and Charman (2019), my prescription is to apply the principles of Bayesian reasoning to address the first two problems and to emphasize standards of integrity and truth-seeking with respect to the third concern.

As noted previously, a prevalent form of confirmation bias arises from overestimating how strongly the evidence in hand supports the hypothesis we hope is true, by forgetting to ask whether the evidence would fit equally well or even better with a rival explanation. Correctly applying Bayesian reasoning automatically precludes this cognitive pitfall, because the key inferential step involves evaluating *likelihood ratios*—instead of asking how expected the evidence would be if the working hypothesis is true, we must ask whether the evidence would be more or less expected under that hypothesis *as compared to* a rival hypothesis. Bayesian inference simply cannot proceed without reference to a rival hypothesis, so there is no room to over-focus on a single hypothesis. It is of course possible that our hopes and desires might psychologically influence our reasoning about which hypothesis makes the evidence more expected. But we emphasize that research is not just a dialogue with the data. It is also a dialogue with a larger community of scholars seeking to identify and resolve disagreements about inferences and thereby accumulate knowledge. Any well-written scholarship should articulate the hypotheses under consideration, present the evidence, and explain the reasoning behind the analytical conclusions. Readers and reviewers can and should scrutinize the author's work for signs of sloppy thinking or motivated reasoning when assessing the evidence. Using the Bayesian framework we advocate, they can evaluate the author's hypotheses and evidence with their own independent brainpower and request revisions or clarifications as needed.

Ad hoc hypothesizing—or constructing "just-so stories" that are over-tailored to the details of the evidence in hand—is a distinct problem that is also readily addressed within a Bayesian framework. Whereas Bayesian likelihood ratios help to protect against confirmation bias, prior probabilities in Bayesian analysis help to protect against ad hoc hypothesizing. Bayes' rule in essence contains a built-in "Occam's razor" that mediates the tradeoff between parsimony and complexity (Fairfield and Charman forthcoming). Compared to simpler rivals, a more complex hypothesis incurs an Occam penalty via its prior probability. If the more complex hypothesis is in fact the best explanation, its posterior probability should win out thanks to the improved inferential leverage it provides compared to the simpler rivals. More precisely, the accumulated weight of evidence will overwhelm the initial Occam penalty. Bayesianism thus penalizes complex explanations if they do not provide enough additional explanatory power relative to simpler rivals, in line with Einstein's dictum that things should be as simple as possible, but no simpler.

To convey a sense of how the Bayesian "Occam effect" works, suppose a stranger at a party shuffles a deck of cards, and you draw the six of spades (Fairfield and Charman 2019, drawing on Jefferys 2003). One hypothesis holds that you arbitrarily selected that card from a randomly shuffled deck (H_R); a rival proposes that the stranger is a magician with a trick deck that forced you to draw the 6 of spades ($H_{6\spadesuit}$). Intuition suggests that $H_{6\spadesuit}$ is ad hoc. The reason it is indeed ad hoc is that we should treat $H_{6\spadesuit}$ as one member of a family of 52 related hypotheses, each of which proposes that the magic trick favors a different card in the deck. Without looking at the card you picked, each of these 52 hypotheses would be equally plausible, so however likely it is that the stranger has a trick deck, that probability must be spread out equally among the 52 different hypotheses in the magic-trick family, thereby reducing the prior probability of the particular possibility $H_{6\spadesuit}$ by a factor of 1/52. In essence, $H_{6\spadesuit}$ derives from a model with an adjustable parameter (the trick card) that has been fit to the data at hand, whereas H_R is a simpler explanation with no adjustable parameters.

Occam factors arise automatically in quantitative Bayesian model comparison. In qualitative research, there are no universal prescriptions for assessing whether a hypothesis is too complex or ad hoc. Our heuristic Bayesian recommendations are to (a) treat inductively-inspired hypotheses with healthy skepticism, (b) start with reasonably simple theories and add complexity incrementally as justified by the data, (c) scrutinize whether all of the causal factors in a hypothesis actually improve explanatory leverage compared to simpler rivals, and (d) ask if the hypothesis might apply more broadly. If a given hypothesis invokes many more causal factors or very specific or elaborate conjunctions of causal factors, good practice would entail penalizing its prior relative to the ri-

vals. If an author fails to treat an inductively inspired or especially complex or finely-tuned hypothesis with adequate prior skepticism, readers and reviewers should take notice and call attention to the problem.

Finally, we advocate a focus on disciplinary norms as the most sensible way to discourage dishonest practices such as deliberately cherry-picking evidence. First, we need to bolster academic commitments to truth-seeking and scientific integrity—quoting Van Evera's still sage advice: "Infusing social science professionals with high standards of honesty is the best solution" (Van Evera 1997, 46). Second, adjusting publication norms regarding requisite levels of confidence in findings would mitigate incentives for falsely bolstering results. For qualitative research, we should embrace Bennett and Checkel's (2015, 13) Bayesian-inspired dictum that "conclusive process tracing is good, but not all good process tracing is conclusive," and focus on providing an honest assessment of the uncertainty surrounding our inferences, rather than attempting to prove that a hypothesis is correct. An associated best practice entails explicitly addressing those pieces of evidence that on their own run most counter to the overall inference; we think that this kind of transparency could encourage critical thinking and signal integrity in a more meaningful way than either VPT or other alternatives like pre-registration.[8]

These suggestions are neither silver bullets nor quick fixes, but in the long term, promoting Bayesian reasoning and rethinking academic norms and practices could help us to do a better job of avoiding cognitive biases, recognizing and characterizing the uncertainty that surrounds our conclusions, and accumulating knowledge, without imposing burdensome straightjackets on qualitative research that would ultimately undermine the quality of its contributions.

References

Ansell, Ben, and David Samuels. 2016. "Journal Editors and 'Results-Free' Research: A Cautionary Note." *Comparative Political Studies* 49, no. 13 (November): 1809–15.

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

Fairfield, Tasha. 2015. *Private Wealth and Public Revenue in Latin America: Business Power and Tax Politics*. New York: Cambridge University Press.

Fairfield, Tasha, and Andrew Charman. 2017. "Explicit Bayesian Analysis for Process Tracing." *Political Analysis* 25, no. 3 (July): 363–80.

_____. 2019. "A Dialogue with the Data: The Bayesian Foundations of Iterative Research in Qualitative Social Science." *Perspectives on Politics* 17, no. 1 (March): 154-167.

_____. Forthcoming. *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*. New York: Cambridge University Press.

Jefferys, William. 2003. "Bayes' Theorem," *Journal of Scientific Exploration* 17(3:537-42).

Sivia, D.S., with J. Skilling. 2006. *Data Analysis: A Bayesian Tutorial,* 2nd Ed. New York: Oxford.

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

Qualitative and Multi-Method Research Fall 2020, Vol. 18, No. 2 https://doi.org/10.5281/zenodo.4046622

The Veils of Inequity, Impracticality, and Inaccessibility

Amy H. Liu
University of Texas at Austin

Two polarizing biases simultaneously plague process tracing—specifically, theory-testing process tracing—as a method. One is selection: We gravitate towards and choose certain pieces of evidence precisely because they corroborate our argument. The other bias is omission: We overlook—if not outright ignore, even if unintentionally—data that run counter to our theoretical priors. Without considering how these two biases affect the data we collect, the inferences we draw may be subject to doubt. To

8 Note that preregistration did nothing to prevent the most prominent recent example of scientific misconduct in political science—the LaCour-Green affair.