

Q

MMMR

Qualitative and Multi-Method Research

Fall 2020 | Volume 18 | Number 2

Inside:

Letter from the President - Jason Seawright

Letter from the Editor - Jennifer Cyr

Original article: Disrupted Fieldwork: Navigating Innovation, Redesign, and Ethics during an Ongoing Pandemic

Authors: Lauren M. MacLean, Nabila Rahman, Robin L. Turner, and Jack Corbett

Original article: Shadow Cases in Comparative Research

Author: Hillel Soifer

Symposium: Authors meet Critics—Veil of ignorance Process Tracing

- Contributors: James Copestake, Gary Goertz, Stephan Haggard, Derek Beach, Tasha Fairfield, Amy H. Liu, Lauren M. MacLean, James Mahoney, Sandrino Smeets

Introducing a New Web Resource for Digital Fieldwork

Longform APSA Awards (2020)

Table of Contents

Letter from the President.....	ii
Jason Seawright - DOI: 10.5281/zenodo.4046504	
Letter from the Editor	iii
Jennifer Cyr - DOI: 10.5281/zenodo.4046528	
Original Article: Disrupted Fieldwork: Navigating Innovation, Redesign, and Ethics during an Ongoing Pandemic	1
Lauren M. Maclean, Nabila Rahman, Robin L. Turner, Jack Corbett DOI: 10.5281/zenodo.4046546	
Original Article: Shadow Cases in Comparative Research	9
Hillel Soifer - DOI: 10.5281/zenodo.4046562	
Symposium: Authors meet Critics—Veil of ignorance Process Tracing	
The Veil of ignorance Process Tracing (VoiPT) Methodology	19
James Copestake, Gary Goertz, Stephan Haggard - DOI: 10.5281/zenodo.4046580	
VPT: A Bad Idea Whose Time Has Come?	26
Derek Beach - DOI: 10.5281/zenodo.4046609	
A Bayesian Perspective on Theory-Blind Data Collection.....	30
Tasha Fairfield - DOI: 10.5281/zenodo.4046614	
The Veils of Inequity, Impracticality, and Inaccessibility	35
Amy H. Liu - DOI: 10.5281/zenodo.4046622	
Unpacking the Assumptions of the Veil of Ignorance to Highlight the Politics of Knowledge Production	38
Lauren M. Maclean - DOI: 10.5281/zenodo.4046636	
Veil of ignorance Process Tracing and Contested Critical Observations	41
James Mahoney - DOI: 10.5281/zenodo.4046644	
The VPT Methodology: Some Practical Considerations and Limitations (to Decoupling).....	45
Sandrino Smeets - DOI: 10.5281/zenodo.4046652	
Lost in Translation: A Reprise on the Veil of ignorance Process Tracing Method.....	47
James Copestake, Gary Goertz, Stephan Haggard - DOI: 10.5281/zenodo.4046688	
Introducing a New Web Resource for Digital Fieldwork	55
DOI: 10.5281/zenodo.4428970	
Longform APSA Awards (2020).....	56
DOI: 10.5281/zenodo.4046696	

Letter from the President

Qualitative and Multi-Method Research

Fall 2020, Vol. 18, No. 2

<https://DOI: 10.5281/zenodo.4046504>**APSA-QMMR
Section Officers****President**Jason Seawright
*Northwestern University***Vice-President**Erica Simmons
*University of Wisconsin-
Madison***President-Elect**Alan Jacobs
*University of British
Columbia***Secretary-Treasurer**Colin Elman
*Syracuse University***QMMR Editor**Jennifer Cyr
*University of Arizona***Executive Committee
Members:**Sara Goodman
*University of California
Irvine*Joe Soss
*University of Minnesota*Layna Mosley
*Princeton University*Marcus Kurtz
*Ohio State University***2021 Nominating
Committee Members:**Taylor Boas
*Boston University*Jennifer Bussell
*University of California
Berkeley*Hillel Soifer
*Temple University***2021 Division Chair**Jody LaPorte
*University of Oxford*Jessica Rich
Marquette University

The past year has been exceptionally traumatic for all of us, along many dimensions. While our personal, social, political, and economic lives have been disrupted in profound ways, it is equally true that many modes of qualitative research have become impossible to carry out as usual in these pandemic times. All of us have faced delays and disruption to our research and scholarly agendas, but the weight of these problems falls on us unequally. Those of us who are established scholars must deal with frustration, but graduate students, junior faculty members, and others still in the career-building stage potentially face much more serious consequences. Some universities have extended graduation and tenure timelines, while many others have not. Even additional time may not fully restore the planned careers of many of our colleagues; we do not, after all, know the future course of the pandemic or related societal and political changes.

We might try to respond to this situation by calling for adjusted expectations and pandemic-specific standards for evaluating scholarly portfolios, but this seems unhelpful. For one thing, many universities and colleagues will not cooperate with such a change. Furthermore, such adjusted standards would likely create a lasting and counterproductive stigma against an entire cohort of scholars.

An alternative, and I think far more productive, approach is for us as a group of methodological scholars to dedicate work and reflection to reevaluating the forms of qualitative research that are compatible with the pandemic period. While fieldwork has long been the ideal (and perhaps idealized) mode for qualitative inquiry, we can and should ask: what are the advantages and disadvantages of in-depth online qualitative interviews? Can Zoom-based focus groups replicate some of the value of ethnographic immersion in a community or an institution, and would such digital modes of research potentially have intellectual advantages as well as drawbacks? How does qualitative inquiry via social media compare with in-person and archival practices?

Reflection and writing on these topics, now and in the coming years, will help us as a community to productively and sensibly evaluate the kinds of qualitative research that are being done now as well as the kinds that are likely to predominate for at least some time to come. Even after the pandemic, serious scholarship on these topics will remain valuable: the barriers to entry for qualitative research are diminished as distanced, digital modes of investigation enter our toolkit. Many scholars with limited budgets, young children, health problems, political situations limiting travel, and so forth can more readily engage with qualitative research techniques when digital tools are part of our menu.

By doing the methodological work to orient ourselves and guide others regarding the best practices and pitfalls to avoid in online qualitative research, we all have an opportunity not just to prepare our discipline to appropriately evaluate pandemic-era research, but also to expand our community moving forward. This is an exciting agenda that deserves far more intellectual prominence in journals and syllabi than most political scientists have given it to date. I hope that we as a section and an intellectual community can contribute in this regard.

I should conclude by emphasizing the ongoing strength of our section despite current challenges. Our participation in this year's all-digital APSA was noteworthy and valuable;

I would like to especially thank the many colleagues who led distanced short courses for their special contribution to the section and the discipline in this challenging time. Furthermore, Sara Niedzwiecki deserves special commendation for organizing our section's excellent program for this unusual event. As always, Colin Elman has worked tirelessly to keep the qualitative and multimethod community active and engaged through this crisis year; he and the rest of the IQMR team deserve special commendation for their efforts to create a partial workshop experience for advanced graduate students who may be unable to participate in future years. Jennifer Cyr has made an extraordinary effort to get the section publication up to date, and along the way has brought us a range of exceptional pieces. I look forward to working with her and the incoming editorial team as we explore the prospects for converting the publication into a peer-reviewed journal. Finally, thanks to all of you for the many small and large ways that you have kept our intellectual community alive and vibrant during these months of distance and isolation. May we meet healthily in person in 2021!

Jason Seawright
j-seawright@northwestern.edu

Letter from the Editor

Qualitative and Multi-Method Research

Fall 2020, Vol. 18, No. 2

[https://DOI: 10.5281/zenodo.4046528](https://doi.org/10.5281/zenodo.4046528)

I am perhaps unreasonably excited to introduce this issue of QMMR to the world. It is provocative, it is topical, it is enlightening—in short, a real page-turner. To be honest, I have come to expect nothing less from our community of methods scholars.

Take, for example, the opening article. The COVID-19 crisis has so many of us roiling, for so many reasons. It has been incredibly unsettling to our personal lives and also to our professional work. In “Disrupted Fieldwork: Navigating Innovation, Redesign, and Ethics during an Ongoing Pandemic,” a set of stellar scholars provide some useful and timely advice for undertaking data collection when circumstances—shaped by resource scarcity, personal obligations, or a global pandemic—make on-the-ground fieldwork near impossible. Along these lines, this issue also introduces and provides information about a new website on digital fieldwork—one that is “[b]uilt by researchers for researchers,” as the announcement toward the end of the issue explains.

Next, Hillel Soifer writes about a practice that many small-N comparativists adopt but for which they have little written guidance: ancillary cases. Shadow cases, in particular, are regularly used to evaluate the extent to which arguments, based on a small set of core cases, travel. The article offers a definition of the shadow case, explains its multiple uses for building from a theory, and offers initial advice on how to choose a case to perform this role.

The current issue also introduces a new technique for undertaking process tracing: the Veil of Ignorance Process Tracing (VPT) methodology. VPT is an approach to process tracing that separates its two primary tasks. A research assistant, who is deliberately “veiled” from knowledge about theories, hypotheses, and mechanisms, is responsible for data selection and coding. The researcher then tests the data collected vis-à-vis the working hypotheses. This division of labor protects against biases that may come when the two roles are fused. As a consequence, its proponents argue, this method yields more valid causal claims.

The practice of process tracing has received no small amount of attention in the past years. VPT, in other words, is not the only approach to undertaking process tracing, although it is arguably the most recent one. We have invited several authors and practitioners of process tracing to comment on VPT. As you will see, the approach is not without controversy. The vigorous and thoughtful give-and-take that unfolds in the pages below represents a master class in how to undertake a robust methodological debate.

I hope you enjoy these texts and learn from them as much as I have.

Wishing you good health and peace of mind in these strange and difficult times,

Jennifer Cyr
jmcyr@arizona.edu

Disrupted Fieldwork: Navigating Innovation, Redesign, and Ethics during an Ongoing Pandemic

Lauren M. MacLean
Indiana University-Bloomington

Nabila Rahman
Indiana University-Bloomington

Robin L. Turner
Butler University

Jack Corbett¹
University of Southampton

Many of us were conducting or planning fieldwork this spring 2020 when the COVID-19 pandemic spread across the globe. Some of us had optimistically hoped that the virus might be waning or somehow under control within a few months. While the situation has improved in some countries such that everyday life is shifting to a “new normal,” in other countries, the pandemic has worsened or reemerged after briefly subsiding.

All of this continues to disrupt fieldwork and make the future difficult to plan.

Researchers from across the social sciences and humanities engage in fieldwork, so this disruption has broad and potentially long-term impacts. Fieldwork is already marginalized in some disciplines due to the time and resources it requires relative to other forms of data collection. The COVID-19 pandemic ratchets up the pressure on established scholars who do this type of work regularly. We must consider other ways of conducting our research and even reevaluate the questions we are asking. A generation of graduate students might be discouraged from attempting fieldwork at all and never gain the experience.

We are persistently hopeful, but we are also profoundly unsure when things will genuinely open up. Any process of opening up will not be linear or unidirectional. Much more likely is a zig zag of ups and downs, or an opening and closing of windows that will be difficult to

predict, evaluate, or know with any certainty whether and how to proceed.

Meanwhile, we remain absolutely convinced of the indispensable value of fieldwork for understanding politics and the social world (Kapiszweski, MacLean, and Read 2015; Krause and Szekely 2020). In fact, it seems more important than ever.

So, how do we do fieldwork in the COVID-19 world? Many of us have been struggling to rethink how we approach field research under these constraints. This essay is an attempt to pull together thoughts from discussions with graduate students and faculty colleagues, various writing projects, published blogs, public webinars, and collaborative Google Docs. It is neither wholly original nor exhaustive, but, hopefully, it provides a touchstone in a time when we are all extremely disoriented.

Taking the Time to Acknowledge the Emotion of Disruption

When fieldwork is disrupted, whether from a global health crisis like COVID, or some other source (e.g., a dramatic change in the political dynamics in the field site; a change in your own or your family’s health), it is important to first acknowledge the emotional weight of the disruption. For many scholars, months and years of work have been dedicated to planning a project design and methodology. Lots of effort has been invested toward obtaining the expertise, funding, and time to carry out the fieldwork.

¹ This essay emerged from the discussion among Jack Corbett, Diana Kim, Lauren MacLean, Nabila Rahman and Robin L. Turner, as well as the participants who engaged in the IQMR Webinar, “Conducting Fieldwork Under COVID Constraints: Interpretive Approaches and Ethnography,” July 21, 2020, at https://www.maxwell.syr.edu/moynihan/cqrm/Summer_2020_Panels_and_Webinars/.

When fieldwork is disrupted, either in process or before it even begins, it can feel like a devastating loss. It can seem like our whole life's plan has been unmoored, and we have become aimless, without purpose. This loss can be heavy on its own. And, it can also remind us of other challenging times when we may have felt similar losses. Perhaps we remember another national crisis or international turning point, or when we have lost someone close to us. This remembering of other hard times can magnify the sadness that we feel.

In the U.S., the impact of COVID is amplified by systemic injustices, including racial, ethnic, and gender disparities. Not only are Black, Indigenous, and Latinx communities disproportionately affected by the pandemic (see, e.g., Lakhani 2020), they are further impacted by the devastation and repeated losses incurred through police violence (Males 2014). Asians, Asian-Americans, and Pacific Islanders have experienced aggravated stereotypes and scapegoating. In light of these inequalities, the loss of our planned fieldwork may not be at the forefront of our minds, particularly for scholars directly impacted by these disparities. Since COVID has magnified intersecting injustices around the world, scholars based outside the U.S. may also have similar experiences and emotions (see, e.g., Johnston, Mohammed, and van der Linden 2020; Public Health England 2020).

It is important to first recognize these intense and sometimes conflicting feelings. Preparing to do fieldwork and then not being able to carry it out is not akin to missing a conference panel or other professional meeting. It is much bigger. Whether fieldwork involves learning new languages, research methods, and literatures, or sifting through old newspapers and microfilm in archives, scholars were ready to move away from the comforts of home and immerse themselves in a new context. They had said their goodbyes and had begun to create new networks of colleagues and contacts in a new field site community. Getting ready to do fieldwork consumes extensive energy and time for both the intellectual and logistical preparations.

Similarly, leaving the field suddenly, or putting a trip on indefinite hold, also takes a lot of intellectual, logistical, and emotional time and energy. We have very little time to say goodbye to new colleagues and friends in the field. We have to cancel flights, hotels, transportation, apartment leases, etc., and try to fight for refunds where they are due. We may need to arrange new accommodation wherever we land or let our roommate or family members know to expect us back! If students, we might need to communicate quickly with our graduate

program to see if funding can be reinstated. Or, as faculty, we may need to try and work ourselves back into the teaching rotation on our campus because we have postponed our sabbatical leave for later. So many things to do at once right when we have no energy and feel little motivation. So many emails and correspondence to follow up on right when it is so difficult to focus on a single thought.

It's important to stop. Take a breath. Get some rest. Eat well. Exercise and get some fresh air. And feel the grief.

This really sucks.

There is no other way to say it.

But, then... we need to pick ourselves up, adapt, and keep moving. So how can we do that?

Some Options for Innovative Data Collection while Maintaining Social Distance

After initially being stunned by the disruption and not being able to move, we may then experience a wave of panic. Does the pandemic preclude qualitative or interpretive analysis when it is difficult to get close and talk to people?

While extremely difficult, we contend that it is not impossible. When we cannot travel or immerse ourselves in the field, we need to innovate to try and access the field remotely. We need to be creative about how to access evidence about how people are thinking and what politics means to them. We discuss below several options for innovation including alternate digital sources of evidence; revisiting previously known sources; remote interviewing via technology; and contracting out to field-based team members.

Identifying Alternate Digital Sources Already Available Online: First, we highlight a range of digital data sources that emerge from the field but are available remotely. Several possible sources of evidence that are available online include but are not limited to:

- Government websites that include policy documents, government regulations, press briefings, speeches, or other mission statements online;
- National, regional, state, or supranational legislatures that post their minutes of debate, proposed bills, and passed legislation online;
- Judicial records, court cases, and judicial opinions that are digitized and available on government websites or in archives;
- Colonial, national, and organizational archival

documents that have been digitized and are available online with some of these archives expanding open access temporarily due to the COVID-19 disruption;

- Social media platforms such as Twitter, Facebook, or WhatsApp, which may have pages or organizations or comments on the topic by the participants of interest, including public accounts by key political leaders;
- Traditional media outlets that may include letters to the editor, opinion pieces, features, or interviews with the participants of interest and may be available online or be digitized in library collections;
- Maps and geo-spatial data that may be available on government or library websites;
- Political party campaign ads, posters, or pamphlets that may be digitized and available through party or library websites;
- Local NGOs and think tanks that may have press briefings with policy documents and statements or analysis by scholars and citizens;
- Local artists including painters, sculptors, performance artists, cartoonists, videographers, photographers, etc., that may have Instagram and Facebook pages where they share images and commentary on their work;
- Local musicians that may post music videos or lyrics on political events or topics on Instagram and Facebook pages with commentary and responses by followers;
- Published autobiographies of political elites or travelers during earlier historical periods that may be available in libraries;
- Published secondary literature that often includes extended quotation of primary sources and can be cited and analyzed from a new theoretical perspective.

And there are many other examples of data sources that may be available publicly online. Some of these sources are highlighted in the “crowdsourced” thinking on disrupted fieldwork cited below (e.g. Lupton 2020).² There are other sources that may be available privately, collected previously by scholars who, if contacted, might be willing to share.

Revisiting Previously Known Sources: In addition to searching for alternate sources of digital evidence, it can

be fruitful to revisit or reanalyze old materials. These may be data we have collected ourselves or archived by other researchers. The key point of this secondary analysis is that revisiting this material may bring a fresh perspective that allows for new insights and conclusions.³ Moreover, new evidence and concepts that did not exist at the start of a project will have emerged and may shed new light on old data. In this way, a fresh analysis can make a novel contribution.

Remote Data Collection Using Voice or Video Technology: Beyond thinking creatively about sources of evidence that may be available online, technology such as Skype, WhatsApp, FaceTime, or Zoom, offer the possibility of doing remote data collection including in-depth, survey, or focus group interviews, or even ethnographic-style observation depending on the field site context. For example, in South Africa, government legislative hearings are now being shown online and permit remote observation.

On the one hand, interviews of any type may be more convenient to schedule and organize if they are done remotely and participants can choose their location. Remote observation of public events obscures the presence of the observing scholar to the participants, perhaps reducing their awareness or possible discomfort at being observed.

On the other hand, while it is possible to establish some rapport in a video or telephone meeting, this is much more difficult if the researcher and participant do not already have a “working relationship.”⁴ If researchers are working in a brand new field site, it will be essential to reach out, introduce themselves, and begin to build a network of interlocutors. In the past, this was often done via email or phone in advance of a field visit. What is challenging is that the researcher may not ever be able to visit in person and may be limited to a virtual encounter. Further, remote observation does not permit immersion in the context, which is usually understood to be a part of the ethnographic method (see , e.g., Schatz 2009).

Researchers will also need to consider participants’ privacy (or the lack thereof) during virtual interviews. It might make sense to seek consent from others who share the living or meeting space and may be present during remote interviews.

Contracting Out: Some researchers may also

² See also DeHart (2020), who draws on ethnography to rethink ideas of location and generate alternative sources or “traces of local sentiment.”

³ Thanks to Diana Kim for sharing this point.

⁴ According to Lee Ann Fujii (2017, 15): “A working relationship simply means that interviewer and interviewee arrive, explicitly or implicitly, at mutually agreeable terms for interacting, conversing, listening, and talking with one another.”

consider contracting out qualitative in-depth or focus group interviewing, survey interviewing, or even archival research to a local research firm or RA in the field site if they are unable to travel. In some situations, contractors will be able to conduct in-person interviews, and, in others, they will offer to facilitate phone interviews or short text-based surveys using their own call centers. In some instances, if a scholar has contacts in the field site already, they may be able to make a connection with a local graduate student who can conduct some archival or policy research on their behalf.

Of course, contracting out requires significant financial resources and professional contacts with reliable firms or assistants that scholars can collaborate with and trust to act in their stead. Researchers should actively work to decolonize knowledge production within these partnerships through collaborative methodologies and approaches (see, e.g., Firchow and Gellman, forthcoming; Asiamah, Awal, and MacLean forthcoming; Fransman and Newman 2019; Smith 2012).

The bottom line from the above discussion is that, even in a pandemic, and sometimes precisely because of the pandemic, researchers still have a variety of options for data collection from or in the field. Each of us may therefore choose a different combination of data sources given the changed reality, or may elect to put more weight on a data source that was previously more of a bonus item.

Project Redesign and the Value of Multiple Contingency Plans

After surveying the range of options, the hard part comes next: going back to the drawing table and engaging in a redesign of the project. To be honest, most field-based projects require some updating and redesign in the field under “normal” conditions, but the pandemic certainly necessitates major rethinking (see Kapiszewski, MacLean, and Read, n.d.; Boswell, Corbett, and Rhodes 2019).

The pandemic not only raises new logistical challenges with theoretical and intellectual implications. It also poses many new ethical dilemmas (discussed in more detail in the next section) that must be at the forefront of any redesign.⁵ The crisis of the pandemic does not grant researchers an “exception” to skimp on ethical requirements to complete our research; quite the reverse.⁶ All of the evaluations involved with a major redesign should require further consultation with a

diverse range of local experts and interlocutors from field sites in addition to one’s collaborators and mentors at the home institution.

Before deciding on changes to the research design, researchers should reexamine their research questions and the key concepts under study. The questions may be formulated in a way that was so heavily dependent on original field-based data that is now simply impossible to obtain, even with all of the creativity and innovation imaginable. Rarely does a project need to be scratched completely. Often the research question can be tweaked or revised so that it is both theoretically interesting and practically feasible to answer given the COVID constraints.

Once the question is clarified and confirmed, scholars should think carefully about the key concepts or hypothesized causal process and how they may be influenced by the pandemic. Are these concepts or processes likely to be independently and significantly shaped by the dramatic changes in politics, the economy, and society related to COVID? Even in normal times, many political scientists are cautious about planning large research projects around high-stakes elections. Or when some significant and unexpected political event surprises them in the field, they attempt to untangle the new complexity on the ground. No one can predict how long COVID will last, but at the time of this writing, it is not going away any time soon. Researchers must be mindful that the pandemic’s effects are experienced unequally across space and by different racial, ethnic, socioeconomic, and age groups. This COVID impact analysis might suggest a new approach to conceptualization and measurement, revised hypotheses, or a new case selection strategy.

Of course, any research project should be developed with contingencies in mind for various research design elements that are impossible to fully know and plan in advance.⁷ Scholars should highlight where they lack key information and think about what they might need to know and when to make decisions.

During a pandemic, we may need to think through a Plan A, B, and C, with more than one contingency plan depending on how the situation for travel and public health evolves. Plan A may be based on the best case scenario with full availability of travel and interaction; Plan B may be a mid-range scenario where accessibility to the field site is hindered or timing is delayed or reduced;

5 See Bond, Lake, and Parkinson (2020) for lessons drawn from conflict studies on situations of changing risk and vulnerability.

6 See London and Kimmelman (2020) for an argument against “pandemic exceptionalism” in biomedical research trials.

7 See Kapiszewski, MacLean, and Read (n.d.) for approaches to dynamically updating research designs.

and Plan C is the worst case scenario where the field is completely inaccessible.

These changes to the research design should be done in consultation with collaborators, mentors, or the graduate student's dissertation committee. Even if the graduate student has successfully defended the proposal, the scale of revision required may merit a Zoom meeting with some or all of the committee members. The full committee ultimately needs to sign off on the dissertation, so it is important to gain their support and input for the redesign.

All relevant Institutional Review Boards (IRBs) at the home institution and elsewhere must also be consulted and provide approvals for changes to the study scope and protocols, although IRB procedures cannot be the sole arbiter of the wide-ranging ethical dilemmas arising from research interactions in pandemic contexts (Cronin-Furman and Lake 2018; Michelson 2016; Kapiszewski, MacLean, and Read 2015; Wood et al. 2020). Already an insufficient indicator of whether research is ethical, the requirements and expectation for restarting research with human subjects vary tremendously across universities and national contexts for different IRBs. For example, on the same day, with the same information, the IRB at Indiana University did not require an amendment to proceed with survey research for an exempt project in Kenya, whereas a Canadian university prohibited any interaction with any human subjects. Researchers should also consult in an ongoing manner with local researchers and other scholars who have worked previously in the field sites to get feedback and adjudicate whether and how best to proceed.

If external or internal funding has already been obtained or is being sought, researchers should contact program officers or grant administrators early to understand how the pandemic may change the availability and requirements of the grant. Timing of future grant cycles may be changed, and extensions to current grants are often possible when requested.

Finally, changes to the research design will inevitably necessitate changes to the proposed writing. Scholars should reconsider the timeline and content for their future articles and book publications. Graduate students should revise an annotated table of contents, which summarizes the argument for each chapter and the evidence used to build the argument. Researchers can also create a project workplan document that includes the revised estimated timeline and sequencing for drafting

of each chapter. They can then revise the full project timeline to estimate the time needed to complete the new types of data collection, analysis, writing, and submission of publications.

Ethics of Digital Innovation

Redesigning a project for a COVID world requires us to rethink the ethics of doing field research. Even though there may be a good bit of data available online, and in some cases the available evidence may be so "abundant" that it is actually disorienting,⁸ we need to be continually mindful of: what is missing; what is not curated; what is not digitized; what voices are underrepresented or even absent entirely; and how we are working to address these absences and erasures.

Of course, this is not a new issue, but it may be obscured by our current remoteness. When we interact with evidence physically, we are continually reminded to think critically about our data. We need to do the same when the data is digitized. Often, the digital archive's final product may present itself to the viewer as more polished and complete when compared to a dusty box of records, loosely tied together with a tattered string, where you can see the missing or torn pages, and be viscerally reminded of the often intentional partiality of the archiving process.⁹

Just as with digital sources, scholars need to again think critically about what types of participants are even accessible for remote interviews, focus groups, or surveys. How do disparities in personal resources or in local infrastructure prevent some from participating in these kinds of virtual interactions? Who is left out completely? It is likely that those who are poor, rural, older, and/or marginalized would have less cellphone and internet connectivity, and therefore be less represented. We need to interrogate the consequences of digital engagement on the range of participants that we include in our work.

It is also vital to reflect on how digital engagement will shape the quality of the relationship we are able to construct with study participants. Even if they are able to use a video connection, much physical presence and body language is obscured from the field of view. Since everyone is seated and "ready" when they open the online platform, there is often little to no time spent on "small talk" and getting settled. Language barriers may be magnified as well when using Skype, Zoom, or the phone.

Regardless of the platform we use, we need to maintain and uphold our ethical commitments to our

8 Thanks to Diana Kim for sharing this point. See also Kim (2020).

9 Thanks to Diana Kim for sharing this point.

study participants as we navigate data collection from a distance. We must consider whether our research and methods are appropriate even if it is IRB-approved. In particular, we need to be sure that participants who agree to an interview or survey are able to do so safely. Will they have privacy from others in their household, or even from predatory hackers or the state? We may need to think about privacy and consent more broadly if household members are present and either participating or visible. Focus groups are always more challenging because the entire group of participants needs to abide by the ethical commitment to confidentiality of what is shared in the group discussion.

When contracting out to a field-based team, numerous hurdles must also be overcome for this strategy to be feasible and ethical (Mani and Barooah 2020; Copper and Sautmann 2020). To begin, researchers should seek multiple, independent sources of information about the situation on the ground and whether it is legally permissible and locally acceptable for hired enumerators or RAs to be conducting interviews or doing archival and policy research in the areas. The health and safety of the research team members as well as the study participants must be paramount.

Researchers should ensure that the research firm or RA is aware of and implementing public health protocols fully and consistently.¹⁰ As with any project, scholars should verify that any local approvals were obtained and reconfirmed given public health changes. Just like the IRB however, local responses to the pandemic vary considerably, thus necessitating careful considerations of public health practices that go above and beyond what local laws or customs may dictate. Moreover, the researcher should be aware of visa restrictions, internal mobility restrictions, evening curfews, and quarantine requirements for whatever travel is planned. In some areas, even if the data suggests that COVID cases are on the decline and the research team/RA follows strict protocols of social distance and wearing masks, residents in the field site communities may nonetheless perceive a threat and harm could result, especially if the people gathering data are considered “outsiders.”

Another issue to consider is whether participants (or RAs) would experience undue inducement to participate in a study that might pose some risk and

thus compromise the process of informed consent. In the pandemic environment of high uncertainty and economic contraction, any type of compensation or even the perception of future benefit from a connection to an outside researcher or local political elites could result in a changed calculation of risk and greater willingness to consent.¹¹ As always, the process of informed consent should take place early, be thorough, and allow for questions and discussion. The pandemic is not a time for hurriedly checking the box on voluntary and informed consent, and, yet, enumerators and participants’ shared concerns about social interactions may push them to rush the process. Ill-equipped as political scientists are to offer specific guidance, we nevertheless believe it is essential to gather up-to-date information about potential health risks in field sites and to adapt our plans as necessary to minimize risks to researchers, enumerators, and participants.

At some point, consulates and universities may begin to permit greater international travel. Or, some scholars may be already located or able to travel to their field site. These researchers still need to evaluate the ethics of continuing their project. Several scholars have argued that ethnographic and in-person, semi-structured interviews should frequently be discontinued and will be the last field methods to be restarted.¹² Researchers should think carefully about the big picture and the opportunity costs that involvement with their research project presents to participants and RAs during a traumatic time. Is this an ethical use of people’s time given the added burdens that COVID places on many people in terms of their work, caregiving, and staying healthy? If researchers do proceed with virtual, contracted, or in-person interviews, the introductions and informed consent process might need to be extended while the length and time spent on interview questions could be trimmed. Researchers also should take both social/physical distance and participant privacy into account in negotiating interview locations.

Finally, these additional ethical considerations regarding whether and how we can conduct fieldwork need to be acknowledged explicitly and discussed clearly in our writing. It is important to document the difficult tradeoffs faced and the rationale for the choices we made as we navigate this new terrain. Reflexive openness about research ethics is not simply a response necessitated by

10 Thanks to Chris Gore, Jennifer Brass, Elizabeth Baldwin, and Alesha Porisky for many of the below points as we discussed the halt and continued pause of our collaborative project.

11 Cronin-Furman and Lake (2018) discuss perverse incentives in conflict zones that also apply in the pandemic context. They usefully provide questions for scholars, reviewers, and readers to consider. See also Philipps (2011) and MacLean and Porisky (n.d.)

12 See a discussion by Pacheco-Vega (2020) on the ethical dilemmas he faced in an ethnographic project with vulnerable communities in Latin America. Wood et al. (2020) provide principles for evaluating whether and how to restart ethnographic field research.

the COVID pandemic, however. This is equally motivated by a broader discussion taking place in the discipline (see, e.g., Fujii 2012). In April 2020, APSA published new principles and guidance on research with human subjects (American Political Science Association 2020). In response, some journals are developing new policies for external review processes.¹³ For example, the editors of *American Political Science Review* are encouraging authors to be open about the ethical decisions they made in the process of conducting their research, and they are asking reviewers to consider research ethics as a dimension of their evaluation (American Political Science Review Editors 2020).

Clearly methodological innovation during COVID will not be easy. But while these new ways of working have inherent challenges that we need to articulate, acknowledge, and discuss, they should not mean that doing this type of research is impossible. Rather, COVID asks us to actively reconsider the perennial ethical questions centered on respect, justice, and beneficence—who is included, how they are treated, and what costs and benefits they experience.

Conclusion

The COVID-19 pandemic has changed our world. Clearly, some aspects of fieldwork may no longer be

possible. But new possibilities may have emerged.

We need to acknowledge the intense emotions associated with this massive disruption to our professional plans and every aspect of our lives, but we also need to adapt and be flexible and keep moving toward our goals. We may be able to think creatively about our research questions and develop alternative perspectives and approaches to the data collection and analysis. As hard as everything is, we need to dig in and keep thinking and writing.

Importantly, the same strategies and creative approaches that are offered here are also relevant for scholars who may not be constrained by COVID-19 but are limited by a lack of fieldwork funding. Other scholars may have the resources, but they are restricted in terms of their time. They may not have the same freedom to leave their professional and family obligations at home or in their home institutions.

Fieldwork remains invaluable for understanding politics. We are unable to do it the way that we have done in the past, but there may be creative alternatives that are ethically sound, methodologically rigorous, and provide some new theoretical insights into the meaning of politics in particular contexts. This may be a while, so it is worth trying.

References

- American Political Science Association (APSA). 2020. "Principles and Guidance for Human Subjects." April 4, 2020. APSA Human Subjects Research Ad Hoc Committee. <https://connect.apsanet.org/hsr/principles-and-guidance/>
- American Political Science Review (APSR) Editors. 2020. "Implementing New Norms for Research with Human Participants," August 24, 2020. <https://www.cambridge.org/core/blog/2020/08/24/implementing-new-norms-for-research-with-human-participants/>
- Asiamah, Gildfred, Mohammed Awal, and Lauren MacLean. Forthcoming "Collaboration for Designing, Conceptualizing and (Possibly) Decolonizing Research in African Politics." *PS: Political Science & Politics*. Manuscript under review.
- Bond, Kanisha D., Milli Lake, and Sarah E. Parkinson. 2020. "Lessons from Conflict Studies on Research During the Coronavirus Pandemic." *Items*, Social Science Research Council, July 2, 2020. <https://items.ssrc.org/covid-19-and-the-social-sciences/social-research-and-insecurity/lessons-from-conflict-studies-on-research-during-the-coronavirus-pandemic/>
- Boswell, John, Jack Corbett, and R. A. W. Rhodes. 2019. *The Art and Craft of Comparison*. Cambridge: Cambridge University Press.
- Copper, Sarah, and Anja Sautmann. 2020. "Best Practices for Conducting Phone Surveys." *J-PAL*, March 20, 2020. <https://www.povertyactionlab.org/blog/3-20-20/best-practices-conducting-phone-surveys>
- Cronin-Furman, Kate, and Milli Lake. 2018. "Ethics Abroad: Fieldwork in Fragile and Violent Contexts." *PS: Political Science & Politics* 51, no. 3 (July): 607-14. doi: 10.1017/S1049096518000379.
- DeHart, Monica. 2020. "Thinking Ethnographically in Pandemic Times." *Items*, Social Science Research Council, May 21, 2020. <https://items.ssrc.org/covid-19-and-the-social-sciences/social-research-and-insecurity/thinking-ethnographically-in-pandemic-times/>
- Firchow, Pamina, and Mneesha Gellman. Forthcoming. "Collaborative Methodologies: Why, For Whom and How?" *PS: Political Science & Politics*.

13 At an APSA Roundtable in September 2020, editors from eight major journals convened to discuss whether and how to revise their practices and procedures.

- Fransman, Jude, and Kate Newman. 2019. "Rethinking Research Partnerships: Evidence and the Politics of Participation in Research Partnerships for International Development." *Journal of International Development* 31, no. 7 (October): 523–44.
- Fujii, Lee Ann. 2017. *Interviewing in Social Science Research: A Relational Approach*. Abingdon: Routledge Press.
- _____. 2012. "Research Ethics 101: Dilemmas and Responsibilities." *PS: Political Science and Politics* 45, no. 4 (October): 717–23.
- Johnston, Regan M., Anwar Mohammed, and Clifton van der Linden. 2020. "Evidence of Exacerbated Gender Inequality in Child Care Obligations in Canada and Australia During the COVID-19 Pandemic." *Politics & Gender*, 1-16. doi: 10.1017/S1743923X20000574.
- Kapiszewski, Diana, Lauren M. MacLean, and Benjamin Read. 2015. *Field Research in Political Science*. New York: Cambridge University Press.
- _____. n.d. "Dynamic Research Design: Iteration in Field-Based Inquiry." Working paper, last modified July 2020.
- Kim, Diana. 2020. "Archives Interrupted: On Methodological Promises and Pitfalls of Digital Abundance (for Political Scientists)." Working paper.
- Krause, Peter, and Ora Szekely, eds. 2020. *Stories from the Field: A Guide to Navigating Fieldwork in Political Science*. New York: Columbia University Press.
- Lakhani, Nina. 2020. "Why Native Americans Took COVID-19 Seriously: 'It's Our Reality'." *The Guardian*, May 26, 2020. <https://www.theguardian.com/us-news/2020/may/26/native-americans-coronavirus-impact>.
- London, Alex John, and Jonathan Kimmelman. 2020. "Against Pandemic Research Exceptionalism." *Science* 368 (6490): 476-77. <https://science.sciencemag.org/content/368/6490/476>.
- Lupton, Deborah, ed. 2020. "Doing Fieldwork in a Pandemic." Crowd-sourced document. <https://docs.google.com/document/d/1cIGjGABB2h2qbduTgfqribHmog9B6P0NvMgVuiHZCl8/edit?ts=5e88ae0a#>.
- MacLean, Lauren M., and Alesha Porisky. n.d. "Showing Appreciation, Giving Motivation or Bribery?: The Ethics of Compensation in Field Research." Working paper, last modified July 2020.
- Males, Mike. 2014. "Who are Police Killing?" *Center on Juvenile and Criminal Justice*, August 26, 2014. <http://www.cjcj.org/mobile/news/8113>.
- Mani, Subha, and Barooah, Bidisha. 2020. "Phone Surveys in Developing Countries Need an Abundance of Caution." *International Initiative for Impact Evaluation*, April 9, 2020. <https://www.3ieimpact.org/blogs/phone-surveys-developing-countries-need-abundance-caution>.
- Michelson, Melissa R. 2016. "The Risk of Over-Reliance on the Institutional Review Board: An Approved Project Is Not Always an Ethical Project." *PS: Political Science & Politics* 49, no. 2 (April): 299-303. doi: 10.1017/S104909651600024X.
- Pacheco-Vega, Raul. 2020. "Fieldwork in the Times of COVID-19: Doing Ethnography During a Pandemic." <http://www.raulpacheco.org/2020/03/fieldwork-in-the-times-of-covid-19-doing-ethnography-during-a-pandemic/>.
- Philipps, Trisha. 2011. "Exploitation in Payments to Research Subjects." *Bioethics* 25, no. 4 (May): 209-19.
- Public Health England. 2020. "Beyond the Data: Understanding the Impact of COVID-19 on BAME Groups." https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/892376/COVID_stakeholder_engagement_synthesis_beyond_the_data.pdf.
- Schatz, Edward, ed. 2009. *Political Ethnography: What Immersion Contributes to the Study of Power*. Chicago: University of Chicago Press.
- Smith, Linda Tuhiwai. 2012. *Decolonizing Methodologies: Research and Indigenous Peoples*. 2nd ed. London: Zed Books.
- Wood, Elisabeth, Douglas Rogers, K. Sivaramakrishnan, and Rene Almeling. 2020. "Resuming Field Research in Pandemic Times." *Items*, Social Science Research Council, May 21, 2020. <https://items.ssrc.org/covid-19-and-the-social-sciences/social-research-and-insecurity/resuming-field-research-in-pandemic-times/>.

Shadow Cases in Comparative Research¹

Hillel Soifer

Temple University

Scholars conducting small-N research often deploy ancillary or peripheral cases that are intended to evaluate the more general validity of the findings of their core case studies. Yet we lack a clear set of methodological guidelines for these ancillary cases.² Drawing on scholarship in the comparative politics subfield for examples, I identify two broad approaches to ancillary cases—the *case illustration* and the *shadow case study*. The case illustration, which consists of showing that outcomes in additional cases match what we expect given theory generated from the core case and the values of independent variables in those additional cases, is widely used to evaluate the generality of findings. Part One of the paper argues that this common practice is not as analytically valuable as it could be. I show that the case illustration approach faces challenges to both external and internal validity and propose a larger-N alternative to assessing claims of generality instead.

In contrast to the case illustration's exploration of correlation, the shadow case study investigates a case more deeply, drawing on techniques of case study research. This entails the deeper investigation of cases that characterizes a well-conducted case study. It goes beyond the scoring of cases on an outcome of interest, a proposed explanation, and other key variables.³ Part Two provides a concrete and precise definition of the shadow case study that centers on its generation of inference from within-case analysis and its role as an ancillary case principally intended to shed light on external validity of

findings generated in the core cases. It also discusses the constraints under which it is investigated. The emphasis on within-case analysis differentiates it from the case illustration, and the constraints under which it is conducted distinguish it from the core case study that sits at the heart of small-N research, rendering it a diminished subtype (Collier and Levitsky 1997) of the case study.⁴ In Part Three, I show that in addition to assessing scope conditions and providing confirmatory illustrations, shadow case studies can be used to explore observable implications, learn from outliers, or assess alternative explanations. I identify conditions under which each of these tasks might be especially useful for the researcher and briefly suggest how cases for these tasks might be selected.

The "Case Illustration," its Limitations, and the Advantages of a Large-N Alternative

Case study and small-N researchers often complement their core cases with what Goertz describes as "superficial case studies used to determine generalization" (Goertz 2017, 239). These commonly entail scoring of cases on the independent and dependent variables to show that the patterns of these scores observed in a series of cases align with the expectations of theory developed and tested in the core case. For example, Ziblatt complements his study of conservative parties and democratization in nineteenth-century Europe with glances at sixteen Latin American countries, some Southeast

1 I thank Rachel Riedl, Tasha Fairfield, Adam Ziegfeld, Sean Yom, Marcus Kreuzer, Erica Simmons, and anonymous reviewers for helpful comments. I also benefitted from feedback from audiences at the IQMR Authors' Workshop at Syracuse University, the Princeton Politics Department Qualitative Research Colloquium, and the Northwestern University Comparative Historical Social Science Seminar. I am especially grateful to Cassandra Emmons for research assistance and several iterations of helpful comments.

2 Throughout, I use the terms *ancillary case* or *peripheral case* to refer to cases other than the *core cases* that are the central focus of the small-N analysis. This is, as discussed further below, the broad category within which I situate both *case illustrations* and *shadow case studies*.

3 See Goertz (2017, 1-6 and *passim*) for a similar discussion of within-case analysis. Like Goertz, and Fairfield and Charman (forthcoming), I conceptualize within-case analysis as encompassing, but not limited to, process tracing evidence about causal mechanisms.

4 I thank Daniel Encinas for suggesting this way to position the shadow case study and use the full term (shadow case study) throughout, rather than "shadow case" to highlight that what I propose is a type of case study.

Asian cases, and those embroiled in the Arab Spring, and concludes that (emphasis added) “*the point appears again and again ... ultimately we cannot understand the course of a country’s democratization trajectory without factoring in the role of a democracy’s adversaries*” (Ziblatt 2017, 361). I call these exercises that dominate existing practice *case illustrations* to highlight the fact that they do not entail any detailed within-case analysis of the ancillary cases.

Scholars select these illustrative cases in a range of ways. Many explore other cases in the region or time period covered by the core cases, building on the implicit assertion that region or time period operates as a scope condition on their theoretical claims (Levitsky 2003; Ziblatt 2006). Other scholars (Handlin 2017; Mylonas 2012; Ziblatt 2017) explicitly mention the importance of using out-of-region cases because they seek to evaluate their claims at a more general level.

In a related but distinct analytical exercise, case illustrations can also be used to explore explicitly defined scope conditions, whether via confirming that the predictions of the theory hold in cases that fall within that scope (Truex 2016, 178) or showing that predictions are disconfirmed outside of them (O’Neill 2005, 236). Though they differ in the extent to which generality is asserted and in how explicitly scope is discussed, these exercises are fundamentally common in their use of evidence about predictions and outcomes in ancillary cases to assess the external validity of claims derived and evaluated in the core cases.

Yet the case illustration is characterized by two fundamental limitations. First, the logic of case selection is rarely made fully explicit. For example, Daly (2016) argues that her account should generalize to civil wars that are non-ethnic and multi-party but does not justify how her ancillary investigations of Nicaragua, Guatemala, Peru, Uganda, and Indonesia were chosen from among the many cases that fit that characterization. The result is that inferences about external validity are, by definition, limited.

Some scholars do provide explicit logics of case selection. Most common perhaps is to select cases in terms of the overall population to which the claim is intended to apply, selecting especially salient cases as settings to assess generalizing claim. Pearlman (2011), for example, uses dramatic liberation struggles in South Africa and Northern Ireland as ancillary cases in her study of the Palestinian National Movement, and Singh (2015, 244) glances at “the two most prominent cases of subnationalism in the world” to examine whether dynamics parallel

to those she finds in India can be observed.

Scholars can also select ancillary cases to explore the outcome of interest at the full range of values of the independent variable, as seen in Barnes (2016). Here, to the extent that outcomes are scored as expected in cases that take a wide range of values of her independent variable, the case illustrations can increase our confidence in the generality of her explanation for women’s legislative collaboration. Yet other scholars provide justifications of case selection that provide weaker grounds for generalization. For example, Wilkinson justifies selecting the ancillary cases of nineteenth-century Ireland, post-independence Malaysia, and post-Communist Romania to complement his account of the political incentives underpinning religious violence in India on the grounds that they represent “one case from each of the three great waves of democratization” (Wilkinson 2004, 204). It is unclear, however, what inference we can draw from the fact that cases vary on this dimension, which is not explicitly connected to the theory developed in the core case of India.

Yet even when this sort of explicit discussion of case selection increases the external validity of these exercises, a more fundamental limitation remains. These case studies tend to do little more than score cases on the independent and dependent variable and show that the outcomes take on the values predicted by the theory given the value taken by the independent variable. These exercises, then, principally rely on evidence of correlation between an independent and dependent variable in each of a series of cases. But if we take seriously the ontological and epistemological underpinnings of a process-oriented approach to causation, we ought to discount this pattern-matching evidence very heavily for falling short of the standards for causal claims developed by qualitative methodologists (Lange 2013, 17).

Scholars are aware of this limitation. For example, Boas (2016, 186) notes that his examination of ancillary cases “is generally insufficient to uncover the mechanisms” by which causation occurred. The novelty in what I propose, then, is not in the recognition of the problem, but in the provision of suggested new ways in which ancillary cases can be used. Rather than resigning ourselves to view the use of ancillary cases as providing little inferential leverage, I suggest some ways forward ways in which insights from additional cases might inform the findings of the core case study analyses. I first propose the replacement of the case illustration with a more systematic cross-case assessment of generality—a

move to a mixed-methods approach that complements case studies with a population-wide generalizing analysis. In the remainder of the paper I identify a set of other uses to which ancillary cases might be put that draw on the tools of case study analysis to increase internal validity and thus provide stronger support for inferences about external validity.

Rather than using case illustrations, I suggest that scholars might instead assess generality or evaluate scope conditions through a larger-N exercise. This would entail coding *all* cases to which the theory is expected to apply on the independent and dependent variables and seeing whether the expected relationship holds or is stronger than alternative accounts.⁵ This approach provides a satisfactory assessment of external validity by explicitly defining a universe of cases and evaluating the performance of the theory in all of them. And though it doesn't address concerns about internal validity in any of the cases included—it relies even more baldly on correlation as evidence for claims about the performance of a theory—it makes this limitation transparent to the reader and forces the author to “bite the bullet” of resting the argument precisely on this type of inference and evidence in an explicit manner rather than relying on imprecise statements of the inference that can be drawn from case illustrations.

Some scholars, of course, already rely on a mixed-methods approach of this type. See, for example, Grzymala-Busse (2015), who explicitly states that “the argument relies chiefly on the paired comparisons” of cases (p. 14), while using large-N analysis (in the Appendix) to assess generality. Other scholars (Mylonas 2012, 170; Handlin 2017, 251) reject the large-N approach to generalization for providing no evidence of causal mechanisms and rely instead on case illustrations. Yet I have shown above that the case illustrations they choose instead also tend to fail to establish causal links precisely because they struggle with identical issues of internal validity. The difference between these approaches, then, is not in the standards of evidence for causation that they use. Instead, it is in the extent of quantification and in the breadth of cases covered. Thus, these critiques seem overstated.

Turning to a large-N approach to generalization depends on the feasibility of coding a large number of cas-

es on both the independent and dependent variables to examine the correlation between them. This may be the case either because they are prohibitively hard to identify or code, or because the population of cases to which a theory is proposed to generalize is fairly small. Where identifying and coding a sufficient number of additional cases is possible, scholars might benefit from turning to a population-wide exercise in assessing generality, rather than doing so in a more limited number of case illustrations.

The Shadow Case Study—A Definition

If assessing generality is not the best use for ancillary cases, how might scholars instead deploy them? Departing from the presumption that scholars engaged in small-N research might prefer to explore the generality of their claims in ways that highlight the inferential power of within-case analysis, the next step is to define the *shadow case study*. Despite the massive flowering of guidelines for case study research, we lack a definition in existing methodological scholarship for a shadow case. The closest approximation comes from Gerring, who suggests that the peripheral or ancillary case is part of the cross-case component of an analysis in which “the emphasis of a study shifts from the individual case to a sample of cases” (Gerring 2007, 20). These are “most often ... surveyed through a quick reading of the secondary literature or through a statistical analysis” (ibid, 22).

I suggest that the shadow case study should be seen in terms of three key characteristics. First, it is a type of case study, rather than an illustration based on a particular case—this means that it relies on “intensive” (Gerring 2007, 20) analysis of a case. Second, it is ancillary to a central case study (oriented to external rather than internal validity), and third, it is constrained in its investigation and presentation by its subsidiary place in an overall project. I therefore define the *shadow case study* as a component of small-N research that entails the examination of an ancillary or peripheral case, drawing inference from the within-case analysis of that case to shed light on the generality of claims most centrally evaluated in the core case.

Case studies and case illustrations: The threshold for what qualifies as a case study rather than a case illustration is not a clear-cut one—it is easier to articulate the standards for a maximally good case study than to iden-

5 One could in principle conduct an analogous exercise using QCA, but the goal would be identifying the set of causal configurations that produce an outcome of interest rather than, as in the approach to case illustrations in much existing scholarship, investigating the validity and generality of a single explanation. Thus, the large-N QCA exercise would produce a different kind of more general claim than the type scholars currently tend to assert via their use of case illustrations.

tify the minimal threshold for qualifying as one.⁶ I have suggested that the presentation of a case becomes more of a “study” and less of an “illustration” as it draws more on tools of within-case analysis to generate causal inference and less on correlations between key variables. This line of argument echoes the critique of King, Keohane, and Verba (KKV) articulated in the distinction between causal-process observations and dataset observations in Collier and Brady (2004). Yet as I discuss further below, one need not restrict case studies to investigations of causal process; inference can be generated from other kinds of evidence collected via intensive study of cases, and indeed this is consistent with the original definition of causal-process observations articulated by Collier and Brady. The distinction between case studies and case illustrations, then, derives from the intensity with which a case is investigated and the extent to which the researcher draws on evidence (whether of causal process or of some other aspect) that goes beyond the scoring of the case on key variables.

Ancillary and core cases: The shadow case study is distinct from the *core case study* or central analytical exercise in small-N research. This is the case study that lies at the heart of scholarship in which the shadow case study may play a complementary role. Oriented principally to internal validity, the core case study draws on tools of within-case causal analysis to leverage the strengths of case selection logics and generate causal inference. The *shadow case study* is an ancillary case, oriented to external validity. The goal of examining the shadow case is principally to show that the theory developed elsewhere holds more generally. But unlike the case illustration (and like the core case), the shadow case study seeks to generate some analytical leverage from within-case analysis.

Bringing these first two characteristics together, we can envision a two-by-two table, with the horizontal dimension scoring emphasis on internal versus external validity, and the vertical dimension scoring the extent of within-case analysis. The core case study would be in the upper-right quadrant (internal validity and within-case analysis); the case illustration in the lower-left (external validity and no within-case analysis). The lower-right cell, which refers to exercises oriented to internal validity without within-case analysis, has been shown to be logically impossible by the recent generation of qualitative

methods scholarship. The remaining, top-left cell, which uses within-case analysis to generate external validity, refers to the shadow case study. The challenge is how to do so under a set of constraints that place a ceiling on the extent of within-case analysis possible in an ancillary case and preclude it from becoming a core case study.

Constraints: Three central forms of constraints can be identified; I term these *logistical*, *presentational*, and *research cycle* constraints. Logistical constraints are principally a function of researcher time and resources. Because the examinations of ancillary cases are relatively abbreviated and are not the core focus of scholarly endeavor, the amount of effort that can reasonably be allocated to their analysis is fairly limited. Limits on researcher time, effort, and resources may limit the data available to a researcher conducting an ancillary case study. This can impact case selection, as scholars prioritize data availability over other considerations (Posner 2005, 260). It can also limit scholars to only exploring parts of their arguments in ancillary cases (Daly 2016, 220). Overall, the impact is to place an upper bound on the strength of the inferences that can be drawn from these cases.

A second type of constraint is presentational. References to space as a limitation in the discussion of ancillary cases are common. Scholars refer to them as “illustrations” (Mylonas 2012, 171), “glimpses” (Daly 2016, 230), or “brief examinations” (Pearlman 2011, 215) that “cannot... be thought of as a full-fledged engagement” (Kurtz 2013, 237 n. 15). The fact that the presentation of ancillary cases can be no more than “brief” or “schematic” places limits on their inferential leverage. In other words, even if scholars conduct a full-fledged within-case analysis of these ancillary cases, they cannot *present* that analysis. Thus, like logistical constraints, presentational constraints place an upper bound on the extent of within-case inference that these ancillary cases can provide.⁷

By research cycle constraints, I refer to the fact that these ancillary cases are investigated after the analysis of the core cases, limiting in practice the set of ways they can be of use for researchers. Their place in the research cycle constrains the analytical tasks of ancillary cases to confirmatory exercises: They cannot serve as sites for other analytical tasks like theory building. Were a scholar to find something in examining a case intended for a shadow case study that led to a fundamental rethinking

6 A similar problem exists in empirical coding of ancillary cases as well; at what point does a scholar’s investigation of an ancillary case advance from an “illustration” to a “study”? There’s a clear-cut difference between the paragraph-length illustrations and the chapter-length studies seen for example in Cammett (2014), but where should the boundary be defined for less clear-cut instances? Parallel with the discussion that follows in this paragraph, I suggest that the intensity with which within-case evidence is used should be the deciding factor.

7 Even if word limits are not a concern, the researcher must still limit the space devoted to ancillary cases in the interest of streamlining presentation and making the overall project more digestible or readable. See Fairfield and Charman 2017 for a similar point.

of the core case, that case would likely take on a different role in the overall study rather than remaining in an ancillary role.⁸

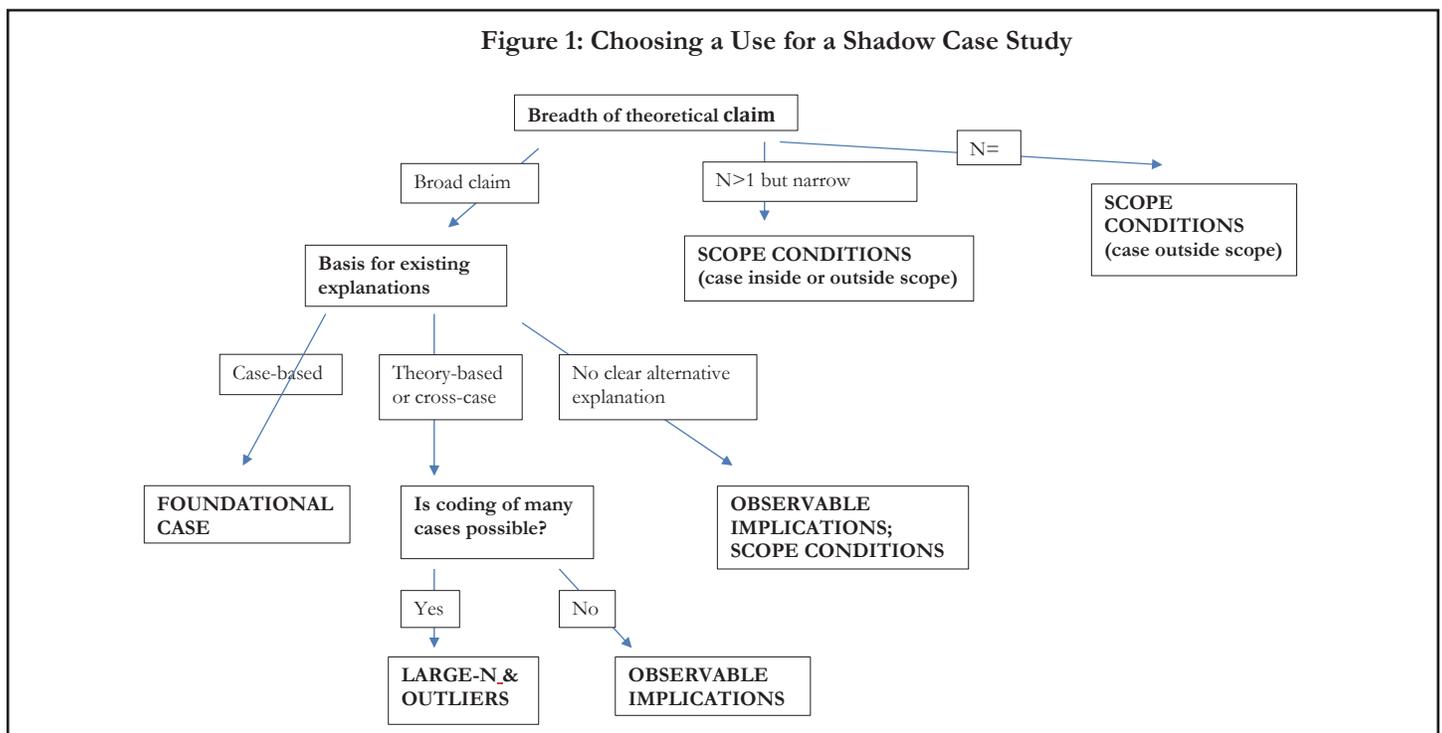
Together, these constraints make the use of full within-case analysis in shadow case studies a logical impossibility and make the shadow case study a diminished subtype of case study rather than a full member of that conceptual category. Yet shadow case studies vary in their intensity. At one end of a continuum, we have studies like Cammett (2014), which supplements an analysis of the logic of social provision by sectarian political actors in Lebanon with a chapter-length investigation of the behavior of similar actors in Iraq. To the extent that a shadow case study like this generates some internal validity, it comes close to a core case study. But the modal shadow case study, bound by the constraints discussed above, would fall closer to the boundary separating this category from the case illustration. The purpose of this paper is to explore how a scholar operating within the constraints that limit the number of core cases one can conduct might leverage shadow case studies, and the within-case inference that they allow, to best increase the validity of the inferential claims developed and advanced in the core cases.

Uses for the Shadow Case Study

As discussed above, existing practice has often deployed ancillary cases to investigate scope by exploring cases within or outside scope conditions. Where the

claims advanced are broad, I suggested that a larger-N alternative might be preferable to the investigation of a relatively small number of cases as a means for assessing generality. But there may be a valuable use of ancillary cases where the breadth of analytical claims is fairly narrow, and a large-N alternative is therefore not feasible.

Yet scholars might also consider using ancillary cases for a set of other possible analytical tasks, described below. This section shows that shadow case studies can be used as a site for a researcher to *address alternative explanations*, to *learn from outliers*, and to *explore additional observable implications*. These inferential tasks are standard fare for case-oriented scholarship, but we lack a systematic discussion of how they can be carried out within the confines of an ancillary case study. I show how each improves the inferences about theory generated and tested in the core cases, while leveraging the distinct advantages of the within-case analytical toolkit within the constraints of the ancillary case setting. I also provide, as outlined in Figure 1, some guidelines for how scholars might choose from among these inferential tasks as well as the exploration of scope conditions (seen on the right side of the figure), as they seek to generate some external validity for their studies. Before proceeding, I note that there may be other valuable uses for shadow case studies beyond those listed here, which were selected because of the obvious ways in which they provide inferential leverage for the broader study.



8 I thank Sean Yom for suggesting the research cycle constraint and Erica Simmons for helpful comments relating to this issue, though I expect she will continue to disagree with my views on it.

Assessing alternative explanations: If we think about the state of the literature on a given topic in terms of alternative explanations, we can imagine three situations, as indicated in the bottom half of Figure One. In some areas of research, no clear alternative explanation exists—this is often the case in fairly new currents of scholarship, or where arguments are especially contested. In others, there is a clear conventional wisdom or consensus around the analytical power of an existing explanation. Yet sometimes this consensus is based on theory or broad cross-case patterns—for example, the idea that the level of inequality affects regime outcomes. Finally, in other areas of research, scholarly confidence in the validity of an existing explanation is based on insight from a particularly illuminating case. I call this type of case—a case on which general claims for a particular explanation of an outcome are most centrally grounded—a *foundational case*.⁹ Argentina, for example, is a *foundational case* for the link between peripheral industrialization and democratic breakdown (O'Donnell 1973) in that the confidence scholars have in this explanation is based not on broad cross-case patterns or on theory alone, but on the within-case understanding of the Argentine case advanced by O'Donnell.¹⁰

A first possible use for shadow case studies is to reconsider the validity of these case-based alternative explanations. As Bayesian scholars have argued in particular (Fairfield and Charman 2019), the plausibility of a new argument advanced depends not only on its fit with cases but on its ability to out-perform or subsume alternative explanations. Under some conditions, a shadow case study of a *foundational case* for an alternative explanation can be an especially powerful analytical exercise.

A common existing practice in case study research is to deal with alternative explanations by controlling for them or by dismissing them based on a brief exploration of patterns in the population. Yet the vast majority of small-N scholars have an epistemological commitment to the importance of evidence about causal mechanisms as a means of evaluating hypothesized explanations for outcomes, and a central tenet of this approach is that evidence about causal process is the strongest and most valid way to evaluate proposed causal relationships. It is striking, then, that when dealing with alternative explanations, scholars tend to draw on types of evidence that

they hold to be weaker and use the evidence from cross-case analysis that is often described by advocates of process tracing as providing nothing more than evidence of correlation.

To provide convincing evidence that casts doubt on alternative explanations, we should instead evaluate them using the same within-case analysis that lies at the heart of small-N research, by examining the causal process theorized to link proposed cause to outcome and showing that it does not operate. This is not a novel suggestion: Many scholars, particularly those who take a formally Bayesian approach to process tracing, argue that one must do so in the core cases themselves (Fairfield and Charman forthcoming).¹¹ Moreover, as suggested above, case-based engagement with an alternative explanation might be especially valuable when our confidence in that alternative is itself based on a case study rather than on other kinds of evidence. It might be especially valuable to select as our core case the foundational case for an alternative explanation, and to show that our explanation out-performs the alternative on its strongest terrain.

Yet, as Gerring (2001) insists, scholars must weigh many criteria in making research design choices. This means that scholars will often choose for their core cases a case other than the foundational case for an alternative explanation. An alternative approach is to select the foundational case for a shadow case study. A scholar presenting a new explanation for a broadly occurring outcome for which a prominent alternative explanation exists can use this approach to show that her preferred theory either subsumes that explanation into a more general account or out-performs it in matching the pattern of observations in the case.

This effort relies heavily on the tools of within-case analysis. Analysis of the foundational case requires a deep engagement with process-based evidence for both the author's preferred explanation and the alternative for which a foundational case has been chosen. One example of the payoffs of this approach can be seen in Kurtz (2013), who turns to Prussia as a shadow case study for his argument that rural social structure was the key determinant of state-building outcomes, as elaborated in his core cases drawn from Latin America. He describes a robust current of scholarship that emphasizes external security threats as the determinant of state strength

⁹ Earlier versions of this paper used the term “paradigmatic case”; I have moved away from that term to highlight that the case is chosen based on the *foundation* on which rests an existing explanation.

¹⁰ Note that a foundational case is not defined in terms of an *outcome*, but in terms of an *explanation*. It is therefore distinct from existing definitions of a crucial case, and from the definition of an influential case in Seawright and Gerring (2008).

¹¹ Yet some methodological guidance diverges on this point: Goertz (2017, 78ff.) suggests that cases where both the alternative explanation and the new one being advanced might operate are characterized by over-determination and should not be chosen for process tracing.

in Prussia and describes it as the foundational case for the claim that war made the state. By showing that “external threat did not drive modernization” until a shift in agrarian social structure toward capitalist agriculture after 1850, Kurtz challenges the “bellic” school of state formation on the very grounds where arguments of that nature are given the most credence (Kurtz 2013, 247).

As this discussion suggests and Figure 1 shows, the foundational case might be an especially apposite use for a shadow case study when a scholar is advancing a broadly general claim, when the state of the literature is such that a new explanation for an existing outcome confronts a particularly prominent and widely held alternative, when confidence in that most prominent alternative derives from the dynamics of an existing case rather than from cross-case evidence or from theory, and when other research design criteria preclude the selection of this existing case as the core case for analysis. Under those circumstances, using a shadow case study to confront the alternative in its foundational case may be the best way to assess the relative explanatory power of a newly proposed explanation and a prominent alternative. In Bayesian terms, this is selecting a case in which our priors strongly favor an alternative explanation. If we can conclude after investigating that our case is favored even in such a case, our confidence in its overall explanatory power relative to that alternative is enhanced. Following the terminology of Fairfield and Charman (forthcoming, Chapter 11), then, this kind of shadow case study may be a setting for high information gain.

Learning from outliers: Rather than using shadow case studies to engage an alternative explanation, scholars might instead use them to explore the shortcomings of one’s own proposed explanation. It may be fruitful to investigate cases mis-predicted by the theory, to engage in within-case analysis of these outliers to find what accounts for the poor fit of the proposed explanation, and thus seek to increase confidence in the theory generated in the core case studies.

One possibility is that cases identified as outliers are a product of measurement error. Scholars can check their scoring of key variables in these cases based on a scholarly consensus by case experts. Doing so not only has the effect of showing that the theory fits those cases, it also increases our confidence in the explanation developed in the core cases, because a greater share of cases is shown to be consistent with theory. Another possibility,

which entails more extensive within-case investigation, is that outlier cases result from causal heterogeneity. Investigating outliers to identify why the proposed explanation does not hold can therefore result in theoretical refinement that improves the validity of causal claims.

In his study of variation in how insurgents treat civilians, Weinstein (2007, 311-326) provides examples of both of these kinds of learning from outliers. He examines four outlying cases in which the number of conflict deaths during civil war is higher than his theory would predict given the relatively low access to resources for rebels. Two (Algeria 1962-64 and Lebanon 1975-91) are found to be a product of measurement error. The other two (Colombia’s *La Violencia* 1949-63 and Algeria 1992-2000) are shown to be outliers because of causal heterogeneity, and their investigation prompts theoretical refinement to incorporate the role of state weakness and to add a scope condition on shifts in government strategy.

As the preceding discussion suggests, the exploration of outliers requires three research tasks: (1) scoring a sizable number of cases or a sizable share of the universe of cases to which the theory is said to apply on reasonable proxies for the independent and dependent variable so that outliers can be identified; (2) assessing whether the cases identified as outliers are an artifact of coding error through an examination of case-specific sources to refine scorings on these variables; and (3) drawing on case-specific secondary sources to identify the causal process producing the outcome in the cases remaining as outliers once task (2) has been completed.¹²

These tasks are feasible within the constraints of the shadow case study and potentially valuable, not because of what we learn about these cases in particular, but in terms of the broader research enterprise. As Figure 1 suggests, this might be an especially valuable use of ancillary cases where the theoretical claim being advanced in the core case is intended to have broad applicability, where ancillary cases do not serve to evaluate alternative explanations, and where the coding of many cases on key variables is feasible.

Two final points can be made about this use for the shadow case study. First, of the uses I describe, this one falls closest to the “floor” between the shadow case study and the case illustration discussed above, since it requires less intensive within-case analysis than the other ways to deploy ancillary cases. Second, and more importantly,

12 If a large number of outliers remain as task (3) is confronted, the researcher faces the problem of how to choose among them. Given the analytical payoffs from theoretical refinement that they may generate, and the feasibility of doing so in a fairly efficient manner, scholars should start with the most high-leverage outliers but explore as many as constraints allow. Thus, again, my recommendation for case selection builds on the Bayesian idea of information gain articulated in Fairfield and Charman (forthcoming, Chapter 11)

since identifying these outliers requires that the researcher code a relatively large number of cases on the independent and dependent variable, this use of the shadow case study is complementary with the mixed-methods approach to assessing generality discussed above. I suggest that these two can fruitfully be combined.

Exploring observable implications: A final use for a shadow case study entails the investigation of a theory's observable implications. The exploration of observable implications outside a core case follows logically from either one of two theoretical starting points about the nature of evidence and inference. First, we tend to hold that confidence in the explanatory power of a theory is improved to the extent that we can confirm with empirical evidence a range of observable implications it generates beyond predictions about the central outcome. This point is, I submit, too often rejected in some recent scholarship on process tracing (e.g., Beach and Pedersen 2019) that limits causal inference to the examination of causal process, reifying "mechanism" and "process" at the expense of ignoring other opportunities to use information about cases to arbitrate among alternative explanations. To be provocative, I suggest that there has been too literal an interpretation of the term "causal process observation" articulated by Collier and Brady to distinguish within-case analysis from dataset observations, and that the result has been an overly narrow approach to causal analysis in small-N research. Like Fairfield and Charman (forthcoming, Chapter 3), I argue that we should take a broad view of the kinds of evidence that generate inference about the causal claims we seek to assess, and I suggest that we can use shadow case studies to explore a range of evidence that goes beyond a narrow definition of causal process.

Second, in most traditions of the philosophy of causation (see for example Falleti and Lynch 2009), those observable manifestations of a causal process in a given case are generally taken as evidence in favor of the more general operation of that process.¹³ If we hold either or both of these two points to be true, we can easily imagine situations where a theory generates observable implications that cannot be assessed in the core cases chosen by a researcher. I present two such scenarios as examples.

First, data needed to evaluate observable implications may not exist in the core cases. This might occur, for example, if the core cases are historical but the theory has

implications that can be assessed in contemporary contexts with survey or experimental data. Singh fruitfully uses a brief discussion of ancillary case evidence in her conclusion as a way of "traveling with the core insight of this book all the way down to the individual," drawing largely on survey experimental evidence not available in her core cases (Singh 2015, 255-6).

Second, some observable implications do not operate in the core cases. For example, a theory tested in democratic contexts might have observable implications in non-democratic settings. Posner takes his theory beyond the core case of regime transition and the "relative political salience of tribal and language group identities in Zambia" and explores how "the general logic it articulates extends well beyond these particular independent and dependent variables," using examples of institutional changes different from those seen in his core case of Zambia (Posner 2005, 274).

While the analysis entailed is not a holistic case study, it is also analytically distinct from the pattern-matching exercise of the case illustration in that it focuses on observable implications that go beyond scores on independent and dependent variables. Because this within-case analysis exists principally as confirmatory evidence for the broader theoretical project, rather than as investigation of the particular case in its own right, these analytical exercises are instances of within-case analysis in the service of the broader findings and are properly placed within the shadow case study category. Scholars should be more attuned to the possibility that these analytical opportunities exist, and that shadow case studies might fruitfully be used to seize upon them. This might be an especially valuable use for ancillary cases where opportunities to weigh one's own preferred explanation against evidence outside core cases provides inferential gains.

Conclusion

Scholars whose research focuses on one or a small number of cases often use ancillary cases in an effort to strengthen their claims. Yet we lack any systematic investigation of the research design and methodological choices entailed in these or any guidelines for how to interpret the inferences they generate. This paper has sought to begin to remedy that gap and develop a framework for considering how one might select and deploy ancillary cases to complement small-N research.

13 This suggests that one might use ancillary cases to explore the generality of a causal mechanism beyond the core case. To the extent that this exercise relied on within-case analysis rather than a review of the treatments of the case by other scholars with the inferential problems that entails, and that the scholar articulated a case selection strategy oriented toward external validity, this could be another use for the shadow case study. Note that some scholarship on process tracing and causal mechanisms (e.g., Beach and Pedersen 2019) rejects the possibility of generalizing across cases about causal mechanisms.

Existing practice has tended towards what I call the case illustration, which uses ancillary cases to make claims of generality or to explore the validity of theorized scope conditions. Scholars most commonly code ancillary cases on core variables of interest and engage in a sort of pattern-matching or correlational exercise. This approach may be of value to the extent that case selection is conducted with an explicit logic in ways that maximize variation within the scope being evaluated and to the extent that the researcher is comfortable with correlational evidence as a way to assess theory. Yet one might argue that both of these criteria may be better satisfied with a large-N and universal exercise in assessing generality, especially when the researcher seeks to advance broadly generalizing claims.

I therefore suggest that scholars who seek to generate inference from internal validity might instead turn to *shadow case studies*, or more intensive investigations of ancillary cases with specific analytical tasks in mind that generate increased inferential leverage. These can serve a variety of inferential purposes, including the exploration of outliers, the search for observable implications, and the confrontation of alternative explanations in the foundational case for them. There are, of course, other uses for the shadow case study. The purpose of this paper, then, is not to provide an exhaustive menu of options. Instead, it is intended to de-center the common practice of the case illustration and to call for more thoughtful research design choices about ancillary cases, which can serve to generate inference to a much greater extent than has heretofore been realized.

References

- Barnes, Tiffany. 2016. *Gendering Legislative Behavior: Institutional Constraints and Collaboration*. New York: Cambridge University Press.
- Beach, Derek, and Rasmus Brun Pedersen. 2019. *Process Tracing Methods*. Ann Arbor: University of Michigan Press.
- Boas, Taylor. 2016. *Presidential Campaigns in Latin America: Electoral Strategies and Success Contagion*. New York: Cambridge University Press.
- Cammett, Melani. 2014. *Compassionate Communalism: Welfare and Sectarianism in Lebanon*. Ithaca: Cornell University Press.
- Collier, David, and Henry Brady. 2004. *Rethinking Social Inquiry*. Lanham: Rowman and Littlefield.
- Collier, David, and Steven Levitsky. 1997. "Democracy with Adjectives: Conceptual Innovation in Comparative Research." *World Politics* 49, no. 3 (April): 430-451.
- Daly, Sarah Zukerman. 2016. *Organized Violence after Civil War: The Geography of Recruitment in Latin America*. New York: Cambridge University Press.
- Fairfield, Tasha, and Andrew Charman. 2017. "Explicit Bayesian Analysis for Process Tracing." *Political Analysis* 25, no. 3 (July): 363-380.
- . 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.
- Falletti, Tulia, and Julia Lynch. 2009. "Context and Causal Mechanisms in Political Analysis." *Comparative Political Studies*. 42, no. 9 (September): 1143-1166.
- Gerring, John. 2001. *Social Science Methodology: A Criterial Framework*. New York: Cambridge University Press.
- . 2007. *Case Study Research: Principles and Practices*. New York: Cambridge University Press.
- Goertz, Gary. 2017. *Multimethod Research, Causal Mechanisms, and Case Studies: An Integrated Approach*. Princeton: Princeton University Press.
- Grzymala-Busse, Anna. 2015. *Nations Under God: How Churches Use Moral Authority to Influence Policy*. Princeton: Princeton University Press.
- Handlin, Samuel. 2017. *State Crisis in Fragile Democracies: Polarization and Political Regimes in South America*. New York: Cambridge University Press.
- Kurtz, Marcus. 2013. *Latin American State Building in Comparative Perspective: Social Foundations of Institutional Order*. New York: Cambridge University Press.
- Lange, Matthew. 2013. *Comparative-Historical Methods*. Thousand Oaks: Sage.

- Levitsky, Steven. 2003. *Transforming Labor-Based Parties in Latin America: Argentine Peronism in Comparative Perspective*. New York: Cambridge University Press.
- Mylonas, Harris. 2012. *The Politics of Nation-Building: Making Co-Nationals, Refugees, and Minorities*. New York: Cambridge University Press.
- O'Donnell, Guillermo. 1973. *Modernization and Bureaucratic Authoritarianism: Studies in South American Politics*. Berkeley, CA: University of California Press.
- O'Neill, Kathleen. 2005. *Decentralizing the State: Elections, Parties, and Local Power in the Andes*. New York: Cambridge University Press.
- Pearlman, Wendy. 2011. *Violence, Non-violence, and the Palestinian National Movement*. New York: Cambridge University Press.
- Posner, Daniel. 2005. *Institutions and Ethnic Politics in Africa*. New York: Cambridge University Press.
- Seawright, Jason, and John Gerring. 2008. "Case selection techniques in case study research: A menu of qualitative and quantitative options." *Political Research Quarterly* 6, no.2: 294-308.
- Singh, Prerna. 2015. *How Solidarity Works for Welfare: Subnationalism and Social Development in India*. New York: Cambridge University Press.
- Truex, Rory. 2016. *Making Autocracy Work: Representation and Responsiveness in Modern China*. New York: Cambridge University Press.
- Weinstein, Jeremy M. 2007. *Inside Rebellion: The Politics of Insurgent Violence*. New York: Cambridge University Press.
- Wilkinson, Steven. 2004. *Votes and Violence: Electoral Competition and Ethnic Riots in India*. New York: Cambridge University Press.
- Ziblatt, Daniel. 2006. *Structuring the State: The Formation of Italy and Germany and the Puzzle of Federalism*. Princeton: Princeton University Press/
- . 2017. *Conservative Political Parties and the Birth of Modern Democracy in Europe*. New York: Cambridge University Press.

Symposium:

Authors meet Critics— Veil of ignorance Process Tracing

Qualitative and Multi-Method Research

Fall 2020, Vol. 18, No. 2

[https://DOI: 10.5281/zenodo.4046580](https://doi.org/10.5281/zenodo.4046580)

The Veil of ignorance Process Tracing (VPT) Methodology

James Copestake
University of Bath

Gary Goertz
University of Notre Dame

Stephan Haggard¹
University of California-San Diego

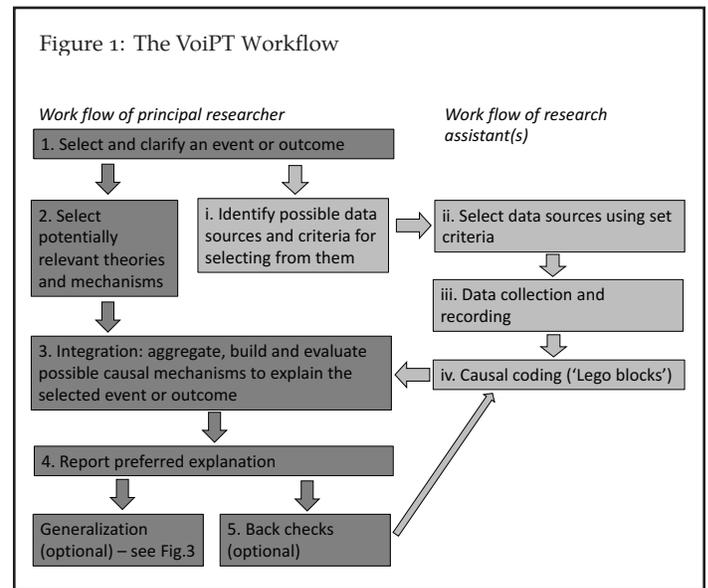
Introduction

A central problem in all qualitative and multimethod research is the risk of selection bias and resulting “cherry-picking” in process tracing. What confidence do we have that the researcher is not producing just-so stories, accounts that are biased in favor of the author’s causal story over valid theoretical competitors? Related to this is omitted variable bias, which often goes along with cherry picking. Analysis of observational data deals with this problem through the inclusion of controls. Experiments deal with selection and omitted variable biases via randomization. Process tracing research lacks similar mechanisms for guarding against biased selection of data and variables.

In this paper we introduce an approach to causal mechanism observation and process tracing methodology that addresses these issues. What we call the Veil of ignorance Process Tracing (VPT) methodology can be used to guide the collection and initial analysis of primary qualitative data for the purpose of process tracing and within-case causal inference. The approach builds on a causal attribution methodology, known as the Qualitative Impact Protocol (QuIP), developed by James Copestake and his colleagues for use in evaluating development projects (2019); in this essay, we extend it to a variety of other settings.

Our approach starts with the proposition that diverse forms of qualitative evidence—open-ended interviews,

primary sources such as archival material, as well as secondary sources—all contain a variety of causal claims. The method seeks to tap these claims for the purpose of causal inference in a systematic and transparent way, reducing possible sources of bias both in identifying sources and extracting information from them.



Core to the veil of ignorance methodology is the separation of the interpretive role of the principal researcher(s) from the task of extracting what might be called “causal claim Lego blocks.” This is a radical departure from current practice in process tracing where the two are completely fused: The same researcher de-

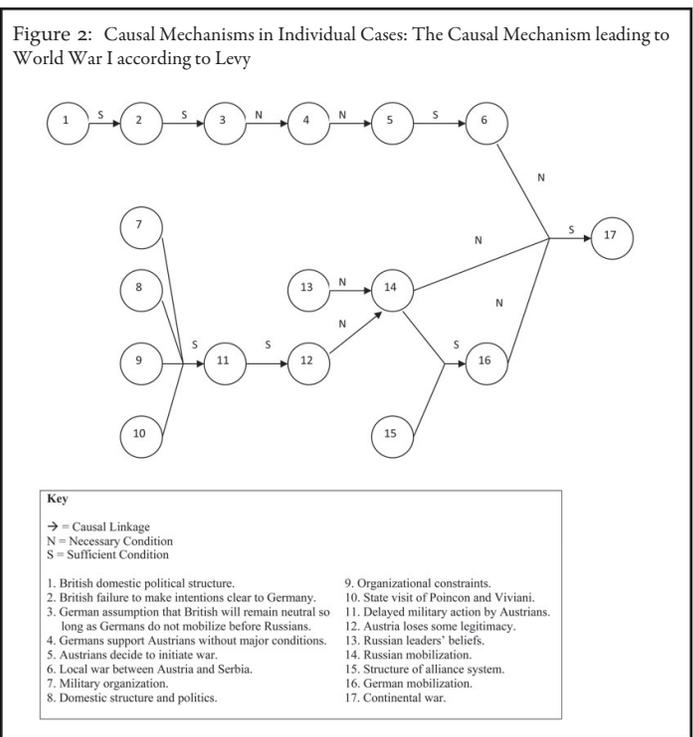
¹ The order of the authors is alphabetical. Thanks to the QuIP team—Fiona Remnant, Rebekah Avard, Marlies Morsink, and Steve Powell—for many discussions. Thanks, as well, to Jim Mahoney for comments on an earlier draft.

termines the sources, examines them for causal information, and then integrates the findings for the purpose of causal inference. 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. The result is less risk of bias arising from priors. This disciplining function is similar in some respects to procedures advanced by Bayesian approaches to process tracing (e.g., Fairfield and Charman 2017, 2019).

This division of labor is encapsulated in Figure 1, which sets out the proposed workflow sequence. As can be seen, it distinguishes clearly between research roles, with the role of the principal researcher (PR) on the left and the role of research assistant (RA) on the right. The main role of the PA on the left (Boxes 1 to 5) is to choose and theoretically frame the case study and use findings to build understanding of the causal processes revealed. In addition, the PR identifies possible sources of data and criteria for selecting from them (Box i). The RA then takes on the role of source selection, evidence collection, and coding (Boxes ii to iv). The whole process relates to a single case, but once complete it may contribute to wider generalization (discussed further below and illustrated by Figure 3).

A second core aspect of this methodology is a laser-like focus on causal mechanisms. The methodology is designed to support process tracing or causal process observation: the identification of the presence (or absence) of postulated causal mechanisms in a given case and a consideration of how they worked to produce the outcome of interest.

A causal mechanism can be represented in a variety of ways, for example, a game-theoretic model. A postulated causal mechanism is constituted in our approach by means of a causal mechanism figure (e.g., Figure 2; see Waldner 2015 and Mahoney, forthcoming for additional examples). A causal mechanism comprises a set of proposed causal links that combine to produce an outcome; these relationships can be represented via a single causal diagram that outlines characteristics of the causal relationships. Testing for a proposed causal mechanism means using empirical information, quantitative or qualitative, to assess whether the mechanism did, in fact, have causal effect. The clarity and rigor of process tracing is advanced by precision in such causal mechanism figures, as well as transparency about the evidence used to support the causal claims within.



In the context of the VPT methodology it is important that these are causal mechanisms explaining individual outcomes and events and not simply general causal mechanisms (see the conclusions for more on this point). Figure 2 shows a causal mechanism figure for the occurrence of World War I according to Jack Levy as interpreted (with Levy’s approval) by Mahoney (2007). Note that the causal mechanism graph consists of several distinct causal chains and intervening variables, some of which have multiple determinants. Note also that the graph can be used to stipulate other features of the relationships, including in this case whether they are necessary or sufficient for the next stage in the causal chain.

Most of the research conducted by the QuIP has relied on interviews with subjects of project interventions. We emphasize the potential for this approach to be used with a wider variety of sources that social scientists use in qualitative work to support both generalization and within-case causal inference. These include:

- Open-ended interviews—including elite interviews—and focus groups.
- Archival sources, including internal memoranda, diplomatic cables, committee reports, meeting minutes, letters, etc.
- Secondary sources, primarily academic books, and articles—extending to policy analysis and journalistic accounts as well.

We illustrate the data gathering and evaluation aspect of the methodology (Figure 1) with an example focusing

on Levitsky and Way's (2010) *Competitive Authoritarianism*. We use this example to discuss the use of a veil of ignorance in determining and coding secondary sources for causal claims. The example can be extended to other primary and secondary sources.

We conclude this essay with an evaluation of how the veil of ignorance methodology fits into the larger multi-method enterprise by considering how it fares against the standards for causal process observation outlined by Bennett and Checkel (2015).

Veil of Ignorance Process Tracing I: Selecting Data Sources

The critical first step in the veil of ignorance workflow—Box 1 in Figure 1—is defining the event or outcome to be explained. The approach focuses on within-case causal inference (of the causes of World War I, for example). This section and the next describe the workflow for what we call the causal claim data collection part of the workflow, which consists of lightly shaded boxes and arrows in Figure 1.

The VPT methodology involves coding causal claims in what we call the sources, whether primary or secondary. In QuIP, the sources are the intended beneficiaries of development interventions, who are sampled for interviews or to participate in focus groups to generate narrative data. These data are used to identify recurring causal claims and stories of change, and to confront the prior theory underpinning the project intervention and experimental or other quantitative evaluations. Where possible, QuIP data collection entails “blindfolding” both field interviewers and interviewees from knowledge of this prior theory and empirical work. In particular, interviewers are not briefed about the theory, hypothesis or precise causal mechanism being tested, and are informed only about the outcomes of interest.

The QuIP experience prompted us to review how blindfolding or veils of ignorance might be used to improve the credibility of process tracing. In this paper we identify at least two other opportunities for the veiling approach. First, there is the formal process of selecting sources of evidence from a wider set of possible sources (Step 3 in Figure 1). Rigor here entails full transparency about selection criteria and procedures to avoid cherry-picking those sources most favorable to a particular theory or causal mechanism. Second, bias can be reduced by delegating the task of identifying causal claims embedded in selected sources (Step 5 in Figure 1) to a research assistant without knowledge of the principal researcher's favored theories.

As an ongoing example, we discuss how the veil of ignorance procedure might be used for selecting and coding secondary sources in the construction of case studies. The coder is asked to review claims in extant sources about the causes of the outcome in question, again without bias with respect to any particular claim. In their outstanding book *Competitive Authoritarianism*, Levitsky and Way (2010) provide numerous case studies testing the causal mechanisms they propose. For example, they offer a case study of Slovakia that runs to just over six pages and draws on over two dozen sources. The purpose is to show how linkages with Western Europe served to limit the abuses of the Meciar government (1993–1998)—a competitive authoritarian regime—and ultimately led to its downfall.

The principal researcher retains responsibility for deciding what criteria should guide the assistant in selecting sources to be coded. This could include the prestige of journals or book publishers, institutional affiliation, strength of personal connections with the event, the status of relevant archives, and so on. These criteria, plus specification of the event or outcome to be explained, determine what the agent will use to look for sources.

In the Slovakia example the sources include books and articles from reputable field and regional journals. We propose a veiled design where coders are instructed to select sources on Slovakia, choosing secondary materials meeting assigned quality or credibility standards and focused broadly on the transition from Communism in Slovakia and the Meciar regime in particular.

A central challenge in selecting the best sources is to establish their potential relevance to specified outcomes in a transparent way that prevents deliberately or unwittingly cherry-picking sources to support a prior view. This does not mean randomized selection necessarily is the best strategy (Seawright and Gerring 2008), but in some settings it might make sense. In the QuIP methodology, for example, sources are most often selected through random selection of respondents within clusters that are chosen purposively to capture diversity of sources based on clearly stated criteria.

Thus, the first veil is between the researcher and the agent who is choosing the sources to be coded. The purpose is to reduce bias in source selection. The core of Veil of Ignorance I is that the agent knows little or nothing about the theory, ideas, or hypotheses of the researcher. Rather, the veiled source agent is provided with guidelines about what constitutes a credible source about the outcome being explained. To state the obvious, the quality of the sources is very important, but

this is no different than any empirical analysis, statistical or experimental.

Veil of ignorance Process Tracing II: Coding Causal Links

The second veiling procedure addresses a particular source of bias that arises in process tracing: that the researcher is the one who is extracting causal claims from the data. In the veil of ignorance methodology, the principal researcher instructs an agent, who we will call the “coder,” to undertake a series of steps designed to extract causal information from the sources. By shielding the coder from the favored causal theory or hunch, the researcher can be assured of a less-biased array of raw data that can be used to construct or test causal process accounts. This is the Veil of ignorance II.

The Veil of ignorance II rests on structuring the relationship between the researcher and coder. Typically, research assistants are cognizant of the aims of a project. For the process described here to be neutral with respect to potential sources of bias, the coder cannot be familiar with the theory-testing objectives of the project and need not even be familiar with the empirical field in question except to the extent absolutely necessary; indeed, the method benefits from the coder operating behind a veil of ignorance that assures neutrality.

With the QuIP, the main blindfolding action takes place in data collection: Field researchers ideally know nothing about the development project being evaluated. Once respondents are selected, they are asked relatively open-ended questions about important changes in their lives in specified outcome domains and why they believe they occurred. However, they do so without reference to the treatment (typically a development project). It is up to the respondents to volunteer whether the project had an impact on their lives, and in an interview context that encourages them to give equal weight to other potential causal drivers.

When it comes to coding, QuIP passes the primary data onto an analyst who performs two separate coding functions, exploratory and confirmatory. The exploratory task entails the purely inductive coding of causal claims (Lego blocks) embedded in the primary text (e.g., an interview transcript). To do this they need no knowledge of the project being evaluated; indeed, ignorance enhances the credibility of the coding as a test of the project’s influence. This is the coding activity which parallels the Veil of ignorance II method and can be equated with theory-building process tracing.

In the Slovakia example the coder would be asked to

extract the causal claims—the causal Lego blocks in our terminology—that the chosen sources make about the causal factors that contributed to the development of democracy in Slovakia and the exit of the Meciar government in particular. The coders would be ignorant of Levitsky and Way’s causal model. However, the coder could and probably would be coding variables postulated by Levitsky and Way; if the theory is robust, they will show up in the coding process. But they will be coded in ignorance of the mechanisms under evaluation.

The instructions to the coders are designed to veil the postulated causal relationships of interest to permit a review of the full array of causal claims that may emerge from the underlying data. The second veil has the interesting corollary that ideal coders can or even should come from somewhat distant substantive areas. Coders need to be familiar with the logic of social science, but ideally not with the specifics of the given subfield in which the process tracing work is being done.

Conceptualizing and Coding Causal Claims

What are the causal claim Lego blocks that constitute potential components of a postulated causal mechanism? We start with the simple causes and then proceed to more complex causal claims. This preliminary list is illustrative; other causal relationships might well be added.

What is it, exactly, that the coder is doing behind their veil of ignorance? With QuIP studies, the relevant causal claims appear in respondents’ interpretations of changes in different dimensions of their wellbeing within a specified period. Respondents might mention a string of possible causal relationships, and these constitute checks on the potential bias of focusing on the intervention only. In archival research, the claims in question are contemporaneous or ex-post assessments of key decision-makers or stakeholders about the causes of specific events, including their own motivation and that of other protagonists (for example, in a strategic game).

Much of existing process tracing centers on the identification of causal claims, derived either from primary or secondary sources. For example, Fairfield and Charman give two examples of what they call “evidence” from open-ended interviews but which are clearly causal claims. The veil of ignorance methodology is about extracting this sort of information in more systematic and transparent way. Here is one of their examples:

E1 = Governing-coalition informants told the investigator that the center-left coalition discussed including a measure to eliminate the tax subsidy in multiple prior tax

reforms, but that measure was ruled out as infeasible on every such occasion due to resistance from the right coalition (2017, 374).

A more complete list of causal attributions that could be coded include the following:

1. Whether the relationship is positive or negative.
2. Explicit claims that X is *not* a cause of Y .
3. Whether X is necessary or sufficient.
4. Interaction terms.
5. Mediator and moderator relationships (e.g., Z mediates the causal relationship between X and Y).
6. Strength of the relationship (e.g., weak, strong).
7. Certainty about the causal claim.
8. Positive versus negative feedback among component causal elements (which can also be drawn as a double arrow).

It is easy to imagine further extensions depending on the issue in question: How narrowly defined causal processes are in time and space, and whether they refer to measurable variables, general conditions, or discrete events. In QuIP studies, for example, respondents are asked to identify drivers of change in specified outcome variables within a specified time-period.

The Levy-Mahoney figure illustrates the need for a systematic set of causal attributions. Mahoney (2007) interprets Levy as proposing a variety of necessary or sufficient conditions. These central causal claim features need to be included in the figure. Often a causal factor is seen as having a positive or negative effect on the outcome. That effect needs specification in the causal mechanism figure.

It is important to realize that this approach marks a fundamental departure from traditional process tracing. Typically process tracing involves evaluating large-scale and complex explanations, as the Levy figure on World War I attests; that is, indeed, its purported advantage. *The veiling methodology helps identify elements of a causal mechanism and not the entire causal mechanism, which is typically complex and involves a variety of component parts.* That is why we call these elements “causal Lego blocks.” The principal researcher and others have the job of making sense of these elements, considering how they can be put together or aggregated, and then evaluating whether they conform with different theories and postulated causal mechanisms (Box 5 in Figure 1).

Integration: Building and Evaluating Causal Mechanisms

The coder provides the principal researcher with a set of causal statements related to the outcome or dependent variable of interest. These claims may converge on a single-favored explanation but will more typically consist of an array of causal claims that ultimately have to be sorted in some way. This final process is of course the most consequential. But it is important to recognize that the interpretation of findings is a necessary stage in quantitative designs as well, even those as simple as a pure experiment.

As our main example we use Mahoney’s causal mechanism interpretation of Levy’s influential analysis of the causes of World War I. Mahoney has extracted the core causal claims from Levy’s analysis and constructed a causal mechanism figure (i.e., Figure 2). Integration occurs here when the causal claim data confront the postulated causal mechanism.

The end product of the VPT methodology is a causal mechanism figure that synthesizes and makes sense of the causal claim data. This can be inductive, starting from the Lego blocks or starting with a given causal mechanism and confronting it with the causal Lego blocks.

At the most basic level, the coder has provided the researcher with the incidence of causal claims in the data, in the form of a set of Lego blocks. One question is therefore whether the favored causal explanation or mechanism—for example, that of Levitsky and Way (2010) for Slovakia—is supported by the evidence once a fully array of potential causal claims are assembled and evaluated. Do the causal claims tend to converge on the favored mechanism?

Most commonly, we would expect researchers to grapple with multiple causal factors and seek to weight them; this process of integration goes somewhat beyond our focus here and clearly involves a series of challenges that all multi-method and qualitative researchers have faced. But the method we propose here is designed to increase the probability of a “no results” outcome in which the favored explanation does not receive support. This could occur either because there is a confounding explanation that exerts a stronger causal effect, or simply because the postulated causal process essentially yields no results: a scattered pattern of causal statements that do not converge around the favored mechanism. Note that if this possibility is not allowed, then the process is prone to bias from the start.

Here is a core and fundamental difference between our process tracing methodology and most others: The

veiling and coding process provides the raw material for the construction of causal mechanisms and explanations as well as the evaluation of existing ones. Thus, the major work of the researcher begins with the analysis of the raw causal claim data.

We can think of our methodology using the common detective metaphor deployed by Collier (2011) in his discussion of process tracing in the context of a Sherlock Holmes story. The veil of ignorance produces causal claims—clues or what we call Lego blocks—from source material. The detective must then integrate them into a causal story that is coherent with the clues. In the integration phase the detective might then go back to the source material to look for specific evidence missing from the raw data (clues) but which might prove critical in the complete causal mechanism. The advantage to this process is that it is transparent about which causal clues originate from the researcher and which from the basic causal claim dataset.

How Does the Veil of Ignorance Methodology Fare Against the Bennett and Checkel Checklist?

We use Bennett and Checkel's list (2015, 21) of desirable features for any process tracing methodology as a framework for discussion and evaluation of the veil of ignorance methodology. We propose that the veil of ignorance methodology deals quite well with many of the items on this list (see also Bath Social and Development Research 2020, Table 3).

1. Cast the net widely for alternative explanations.

The core principle for the source agent is to locate sources with some level of a priori credibility of their causal accounts. Since the agent does not know the theory in question the net is cast widely in terms of a variety of sources. This can be handled in part through randomization with respect to sources, but usually requires more thought in purposeful selection of primary and secondary sources.

2. Be equally tough on the alternative explanations.

Because the causal claim coder does not know the alternative explanations, there is no way for her to be tough or easy on any of them.

3. Consider the potential biases of evidentiary sources.

Given that the source agent does not know the theory under investigation it is hard for her to have any significant bias. The purpose of the veiling is to choose a wide-ranging list of credible sources so as to minimize and reduce bias. The methodology does not consider biases in the sources themselves, but this is at least partly

obviated by the criteria for choosing sources at the outset and the fact that strong and credible theories will presumably get support across sources.

4. Take into account whether the case is most or least likely for alternative explanations.

The veil of ignorance prevents the coder from knowing anything about the potential explanations under consideration. The role of the researcher during the integration phase is then to assess which possible explanation is most consistent with the Lego block evidence supplied by the causal claim data.

5. Make a justifiable decision on when to start.

A core question for any causal process observation is when in time one should start the analysis. While this choice obviously will depend on the specific case being explained, the causal claims themselves can suggest a beginning point. One would probably not want to begin at a point before or after many causal claims in the data set.

6. Be relentless in gathering diverse and relevant evidence but make a justifiable decision on when to stop.

The veil of ignorance approach clearly incentivizes such a process.

7. Combine process tracing with case comparisons when useful for the research goal and feasible.

The veil of ignorance causal mechanism methodology is about within-case causal inference, but there is no reason why this method could not be extended to those favoring small-N qualitative comparative analysis or even large-N qualitative analysis (LNQA; Goertz and Haggard 2021).

8. Be open to inductive insights.

Obviously, the causal claims themselves constitute precisely inductive insights. They may be features of the causal mechanism that were not obvious to the researcher beforehand. As a result, the sources can be systematically mined for insights because they do not depend on the pre-existing positions of the researcher.

9. Use deduction to ask, "If my explanation is true, what will be the specific process leading to the outcome?"

Core to the Bayesian approach is the prior belief in the hypothesis under consideration vis-à-vis its alternatives. Absolutely central to the veiling procedure is that the priors influencing collection and initial compilation of evidence are either made fully transparent or avoided altogether. In Bayesian terms this becomes a uniform prior. One does not explicitly consider alternative explanations and mechanisms until the basic raw causal claim data are delivered to the researcher. At this point of course the various alternatives are considered in the integration phase to produce a conclusion.

10. Remember that conclusive process tracing is good, but not all good process tracing is conclusive.

As with all data, there is no guarantee that the results will be conclusive. It might be that the causal claims fit naturally into one coherent causal mechanism. But it is also quite possible that there are multiple mechanisms that are consistent with the basic causal claim data.

11. Process tracing should be transparent.

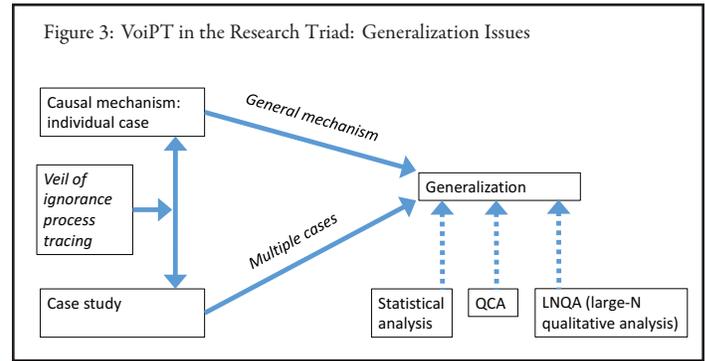
This last item is our addition to the Bennett and Checkel (2015) list, which we consider to be both important and noncontroversial. By generating a specific and comprehensive list of sources, explicitly coding causal claims, and generating causal mechanism figures, one can make the methodology very transparent.

Conclusion

This paper has proposed a methodology for within-case causal inference, the VPT methodology. Our proposal draws on experience using the QuIP to evaluate the impact of over forty development projects through analysis of open-ended interviews, field-testing the idea. Here, we extend the basic method to a variety of other primary and secondary sources.

The veil of ignorance approach provides some practical steps for making process tracing more rigorous and transparent. A core idea is to use a veil of ignorance to address potential bias in selection and coding of data for building and testing causal theories. We have explained how this entails introducing a more explicit division of roles within the research process between a lead researcher and research assistants delegated blindfolded responsibility for aspects of data selection and coding.

Looking at the eleven criteria for good process tracing discussed above, the veil of ignorance approach appears to have significant promise. It deals quite effectively with a majority of the listed criteria, and those it does not address reside in what we call the integration phase, where “clues” or Lego building blocks are assessed, weighted, and combined. But the veil of ignorance methodology assures that these bits of evidence have been extracted through a process that reduces bias. That process rests on separating these two crucial steps in the process tracing exercise: The causal claims “raw data,” and their integration into an assessment of postulated causal mechanisms.



One might object that this methodology involves a tremendous amount of effort in order to conduct a causal analysis of an individual case. This objection depends on how an individual-case causal mechanism analysis is seen as fitting into broader research efforts. Our priors (and those of this section) are that such cases are not mere add-ons or afterthoughts but are crucial components of increasing our confidence in theoretical priors. Their generation thus richly deserves closer scrutiny.

Figure 3 illustrates how we see this methodology as being connected with the larger research enterprise in the social sciences. We see multimethod research as involving three core components—a research triad. Most social scientists are less interested in causal explanation of individual cases than in generalizing from them across a wider population, bounded by appropriate scope conditions. Generalization takes two forms, as indicated in the figure. The first is moving from an individual-case causal mechanism figure and analysis to a general causal mechanism figure that works in multiple cases within some scope limits. The general causal mechanism figure must be connected to the individual case one, but probably will be somewhat different because it is meant to apply to many cases. The second form of generalization lies in the empirical analysis of multiple cases. This can be done by considering how closely selected cases embody the general causal mechanism across the wider population (the “multiple cases” arrow in Figure 3). Qualitative analysis can contribute to this goal (e.g., QCA) or other cross-case qualitative methods that multiply the number of case studies (Goertz and Haggard 2021). The final goal is a general causal mechanism figure that applies to many cases not just the one used in the VPT methodology.

References

- Bath Social and Development Research. 2020. “The Qualitative Impact Assessment Protocol.” BSDR Briefing Paper. <https://bathhdr.org/wp-content/uploads/2020/10/QUIP-briefing-paper-October-2020.pdf>.
- Bennett, Andrew and Jeffrey Checkel, eds. 2015. *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool*. New York: Cambridge University Press.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44, no. 4 (October): 823–30.

Copestake, James, Marlies Morsink, and Fiona Remnant, eds. 2019. *Attributing Development Impact: The Qualitative Impact Protocol Case Book*. Rugby: Practical Action Publishing. <https://practicalactionpublishing.com/book/105/attributing-development-impact>.

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.

Goertz, Gary, and Stephan Haggard. 2021. "Generalization, Case Studies, and Within-Case Causal Inference: Large-N Qualitative Analysis (LNQA)." In H. Kincaid and J. Van Bouwel (eds.) *The Oxford handbook on the philosophy of political science*. Oxford: Oxford University Press.

Levitsky, Steven, and Lucan Way. 2010. *Competitive Authoritarianism: Hybrid Regimes After the Cold War*. Cambridge: Cambridge University Press.

Mahoney, James. 2007. "The Elaboration Model and Necessary Causes." In *Explaining War and Peace: Case Studies and Necessary Condition Counterfactuals*, edited by Gary Goertz and Jack Levy, 281–306. London: Routledge.

_____. Forthcoming. *The Logic of Social Science*. Princeton: Princeton University Press.

Seawright, Jason, and John Gerring, 2008. "Case selection techniques in case study research: a menu of qualitative and quantitative options." *Political Research Quarterly*, no. 61: 294–308.

Waldner, David. 2015. "Process Tracing and Qualitative Causal Inference." *Security Studies* 24, no. 2 (June): 239–50.

VPT: A Bad Idea Whose Time Has Come?

Derek Beach
Aarhus University

This has been a difficult review to write. On the one hand, it is important to be constructive and open to new approaches for how to do good process tracing (PT). On the other hand, the proposed "Veil of ignorance Process Tracing" (VPT) attempts to solve one vaguely defined problem through a confused mismatch of ideas that result in a set of much bigger problems. Building on my experiences both as a practitioner of multiple PT case studies (Beach, Schäfer, and Smeets 2019; Beach and Smeets 2020), and a methodologist who has published whole chapters detailing how to guard against confirmation bias (see e.g., Beach and Pedersen 2019), my diagnosis of VPT is that there are a series of highly problematic (and puzzling) theoretical, methodological, and practical issues that make for a recipe producing poorly-evidenced empirical generalizations as to why an event took place that will not result in a real contribution to our broader understanding of how causal processes work in real-world cases.

The authors postulate—without evidence—that confirmation bias is an endemic problem in process tracing, insinuating that much existing PT consists of "just-so stories" where researchers search for evidence that confirms their pet theory. In reality, contrary to what these authors write, confirmation bias in non-experimental science is not avoided by "blinding," but instead by being one's own strongest critic when evaluating empirical evidence, achieved in PT through thorough and transparent assessment of the uniqueness of found evidence in relation to alternative explanations (see e.g. Fairfield and Charman 2017; Beach and Pedersen 2019). Further, good peer review should detect this bias—but is not mentioned by the authors.

Before I detail these problems, a background note is important to understand the context within which VPT was developed. The method comes from the field of policy evaluation, within which there is a widespread belief amongst *funders* that the only scientific method that avoids confirmation bias in evaluations is to use counter-

factual-based assessments in the form of experimental methods (termed randomized controlled trials (RCTs) in evaluation).¹ The VPT method reads as if it were originally framed to allay funder concerns about evaluation case studies by including a number of buzzwords from experimentalist research. But this begs the question: Why should political science or international relations adopt this unhappy marriage between a naïve, empiricist version of grounded theory with ideas about blinding research processes that are appropriate when using RCTs?

Confused Assertions About Types of Causal Claims Being Made

VPT asserts itself as a method for making causal attributions, but, in my opinion, it is all over the place regarding the *nature* of the causal claims being assessed. The proposal uses the terminology of causal effects throughout but also proposes eight (!) different types of causal relationships that can be explored using VPT. These include [positively/negatively/not related] as a [necessary/sufficient] cause [mediated/interacted with Z] that had a [weak/strong] impact on Y, potentially with [positive/negative] feedback. At the same time, the proposal insists that: “*The veiling methodology helps identify elements of a causal mechanism*” (this issue, italics in original).² As its name suggests, PT is all about tracing causal processes (aka mechanisms) that link a cause (or set of causes) and an outcome together (e.g., Bennett and Checkel 2014; Schmitt and Beach 2015; Beach and Pedersen 2019). This also implies that PT—as a within-case method—cannot assess net causal effects because of the problem of masking.³

At no point in the article is the *causal* nature of mechanistic explanations developed (see below), which together with the eight different types of causal claims that the VoiPT supposedly can evidence, results in complete confusion about the ontological properties of “causal claim Lego blocks.”⁴ The authors even readily admit this on the first page, writing: “Our approach starts with the proposition that diverse forms of qualitative evidence... all contain a *variety* of causal claims (this issue, italics added).” As a result, the VPT proposal forges ahead by asserting that “causal claim Lego blocks” are just lying in

wait in the empirical record, like causal gold nuggets that the research assistant (RA) can snatch up.

Process Tracing Without the Mechanisms

If we put aside the problems created by the numerous types of causal claims involved, the VPT proposes that the found causal gold nuggets are compared by the principal researcher (PR) with her “favored” causal mechanism in the “integration” stage of VPT. However, what this involves is unclear because the VPT appears to understand mechanisms as a more descriptive series of events. The article uses the example of Levy’s work on the outbreak of continental war (WWI). When we look at the series of events put forward by Levy in the figure, however, it quickly becomes obvious that causation is being *assumed* through temporal sequencing of events, but there is no causal theorization of the linkages between them. For example, there is no explanation for why the combination of “Russian beliefs” and “Austria loses some legitimacy” produces the next step, “Russian mobilization.” What is the causal linkage between a static factor (Russian beliefs) and an unexplained event (Austria losing some legitimacy)? Therefore, the explanation produced is a series of factors/events, not a mechanistic causal explanation.

But how then can *causal* gold nuggets supplied by the RA aimed at explaining *why* the outcome occurred in the case be compared with a more *descriptive* series of events (i.e. *what* happened) that the PR has developed beforehand? And even if we ignore this mismatch, the set of causal gold nuggets delivered by the RA would only tell the PR that a majority of stakeholders believed that X was causally related in one or more of the eight types of causal types (see above). How a plausible causal mechanism could be reconstructed on the basis of such a disparate set of causal gold nuggets about why an outcome or event occurred is difficult to see.

Lack of Alignment Between Type of Causal Claim and Methods to Make Inferences

Moving to the issue of inference, as we know from Hall’s now famous principle of methodological alignment (2003), our methodology has to be appropriate

1 Interestingly, while many funders hold these beliefs, many methodologists in the field assert that other methods (e.g., contribution analysis) are just as robust but importantly that they also answer different questions with different types of evidence (e.g. Scriven 2008; Mayne 2012; Cartwright 2012; Schmitt and Beach 2015).

2 One of the eight types of “causal claims” that is not even causal in nature. “Certainty about the causal claim” is an epistemological question related to the strength of the evidence backing a causal claim.

3 It is now well established in the philosophy of science that tracing a causal mechanism in a case does *not* enable assessment of net causal effects, only that there is a mechanism linking them. See Clarke et al. 2014.

4 The VPT proposal also does not get Lego terminology correct. Lego has *bricks*, not blocks. See www.lego.com.

for evidencing the *type* of causal claim we are making. Alignment means that evidencing a claim about *necessity* requires a counterfactual comparison (either using actual or hypothetical cases), whereas evidencing a claim about the *strength* of a causal relationship requires quantitative cross-case data about variation in X, Y, and confounders, ideally interval-scale data that can be assessed using multiple regression or other appropriate statistical tools to figure out the strength of the net causal effect. As regards a claim about necessity, is the VPT suggesting that when the RA asks a stakeholder about why event Y occurred, and they respond by saying that they think event X was necessary for Y to occur, that this is a valid causal inference? Can we assume that the stakeholder engaged in a proper counterfactual, cross-case assessment of the claim (ideally from a natural experiment)? Or are we just accepting the causal attribution made by the stakeholder as such, in effect accepting *hearsay* as evidence confirming the claim about necessity?

Induction and the Myth of Data Speaking for Itself

Another fundamental issue relates to the issue of inferences and moving from empirics back to theory. The VPT proposal builds on a naïve empiricist reading of grounded theory, where theories are built *inductively* from observable data. In effect, the VPT assumes that causal gold nuggets are just lying in wait in an uncontaminated state, where it is the ignorance of theory and empirics that enables the RA to extract these “causal claims from the data” without being tainted by prior theoretical bias. But despite developing elaborate coding procedures, the naïve empiricist reading of grounded theory itself never lived up to its promise, because data does not speak for itself. As is well established in the philosophy of science, observation of the world is *per definition* theoretically informed (Tavory and Timmermans 2014, 13-17; Jackson 2016). Causal claims are what give empirical data explanatory meaning, with scientific research as a sustained dialogue between theory and empirics. For example, when an analyst finds a puzzling rhetorical pattern in a speech by a politician, existing theories coupled with detailed empirical knowledge about what is typically in speeches are the tools that enable the analyst to figure out what the pattern means and thereby also what theoretical proposition it could be evidence of.

In my substantive research I have spent many hours reading and analyzing archival documents, during which I have *never* seen a causal gold nugget just lying around in a policy brief, intelligence assessment, or draft proposal.

Instead, working with archival information is a continuous back-and-forth between focused searches based on theoretical hunches, finding new puzzling things in documents, and then returning to theory (and case knowledge) to make sense of them, and back again. The idea that a novice RA could extract causal gold nuggets from the empirical record without any bias is therefore difficult to see.

Inherent Bias in Stakeholder Attributions of Causality

VPT suggests that causal attributions by stakeholders are *less* prone to bias. VPT proposes to use statements by stakeholders about why something happened, drawn from interviews or archival documents they have produced, supplemented with (ideally randomly selected!) attributions found in the secondary literature. The authors write that as the “source agent does not know the theory under investigation, it is hard for her to have any significant bias” (this issue).

But there is a very simple reason we use the term stakeholder—the person has a STAKE in the issue being discussed. These statements can act as evidence of what a given stakeholder wants to be heard saying for the sake of posterity. And even if they could be trusted, there can be a myriad of reasons detailed by cognitive psychologists for why their perceptions might not match the perceptions of an outside observer. This is why historians typically do *not* use stakeholder interviews, or trust stakeholder confessionals in archival documents. Instead, they want to gain access to archival documents that can enable them to reconstruct a historical process. Political scientists often do not have this luxury, and therefore have to subject these types of confessionals with robust source criticism—which raises the next point.

Inexperienced RAs Cannot Engage in Source Criticism

The suggestion of using *inexperienced* RAs to collect and code empirical material is problematic. The authors write that the RA “need not even be familiar with the empirical field in question except to the extent absolutely necessary” (this issue). However, this inexperienced RA would lack the case-related knowledge to be able to do a good interview (e.g., follow-up questions), and more critically, would not be able to engage in proper source criticism of interviews, archival documents, or secondary sources—all of which would contain implicit or explicit bias that requires significant theoretical and empirical knowledge to evaluate (see Beach and Pedersen 2019, 195-222; Møller and Skaaning 2018). Evaluating what

statements of stakeholders mean in a particular case, and whether we can trust them, requires significant case-specific knowledge and experience. Often evidence of parts of causal mechanisms are not simple “confessions,” but instead involve analyzing patterns of speech, sequencing of events, or even just the fact that two social actors held a meeting. All of these empirical observations would have to be evaluated in relation to a theory *before* they could be *evidence of* anything. Good theoretical explanations cannot just be extracted by a novice from the data! If this were the case, the most compelling research would be found in the first essays done by political science students in our bachelor programs.

Therefore, the blinding of the VPT creates the very real risk that the RA regurgitates the causal gold nuggets found in statements from a majority of interviewees, archival documents, or secondary sources in the report to the PR. The PR expert is several steps removed from the nitty-gritty of the sources, preventing them from engaging in real source criticism when evaluating the evidential backing for causal gold nuggets. The result is the *de facto* replacement of a good, experienced researcher with a novice.

And Would I Really Send an Ignorant RA to do Interviews?

On a final practical note, deliberately sending an inexperienced RA to collect causal gold nuggets from interviews with high-level stakeholders would run into practical difficulties. In my own substantive research on high-level crisis governance in the EU, would I really send an RA who knows nothing about how Brussels works or the dossiers I am studying to conduct an interview on my behalf? One of the first things that happens in elite interviewing is that high-level civil servants test you to

see whether you have done your homework (e.g., having read all of the proposals and public accounts of the negotiations). If they find you wanting, the interview will very quickly end because you are wasting their time. On the upside, once you have passed this “test,” they can be very useful informants, often offering further non-public documents and drafts, along with helping you understand the signals embedded in the often-arcane language they use in legislative proposals and communications. All of this empirical material that I can use to understand how things work in Brussels would be completely out of reach for an inexperienced RA.

Conclusion

This review might seem excessively harsh. It was not intended to be, but the VPT proposal has a fractal character in which the more you dig into it, the more confused and problematic the suggestions become. The article tries to validate the method by claiming that “the idea has been field tested” in many evaluations (this issue). However, just because something has been done does not make it right, and it cannot mask the fundamental flaws in the theoretical and empirical logic inherent in the naïve empiricism of VPT, coupled with the problematic idea of blinding.

Fortunately, the state-of-the-art regarding PT methods is sufficiently developed, such that there exist robust methodological alternatives that are being widely used by researchers to produce important, evidence-based findings about causal mechanisms at play in the real world (see Beach 2020 for a review of recent uses and methodological developments). VPT is a bad idea whose time does not need to come.

References:

- Beach, Derek. 2020. “Process Tracing Methods.” In *Oxford Bibliographies Online—International Relations*. DOI: 10.1093/OBO/9780199743292-0227.
- Beach, Derek, and Rasmus Brun Pedersen. 2019. *Process-Tracing Methods: Foundations and Guidelines*. 2nd Edition. Ann Arbor: University of Michigan Press.
- Beach, Derek, and Sandrino Smeets. 2020. “The Unseen Hands: Collaborative Instrumental Leadership in the British Renegotiation Case.” *European Journal of Political Research* 59, no 2 (May): 444-464.
- Beach, Derek, David Schäfer, and Sandrino Smeets. 2019. “The Past in the Present—The Role of Analogical Reasoning in Epistemic Learning About How to Tackle Complex Policy Problems.” *Policy Studies Journal*. doi.org/10.1111/psj.12372.
- Bennett, Andrew, and Jeffrey T. Checkel, eds. 2014. *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.
- Cartwright, Nancy. 2012. “Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps.” *Philosophy of Science* 79, no 5 (December): 973-989.
- Fairfield, Tasha, and Andrew E. Charman. 2017. “Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and

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.

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 testing¹—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: $E = \text{source } S \text{ made statement } X \text{ in context } C$. 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 $H_A = \text{suspect } A \text{ acting alone killed the victim with an ice pick}$, and $H_B = \text{suspect } B \text{ acting alone killed the victim with an ice pick}$, 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.

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 ar-

gument. 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.

this end, Copestake, Goertz, and Haggard offer an approach for addressing this challenge. The foundation of what they call Veil of Ignorance Process Tracing (VPT) is a division of labor. The principal researcher does the theoretical work but has little—if any—role in the data collection; in contrast, the research assistants do the data collection and the subsequent coding but in the absence of any theoretical priors; then after the collection and coding is complete, the principal researcher uses the data to test their argument. The intuition is that divorcing the theoretical priors from the data collection efforts ensures less risk of bias. Certainly, this approach is unorthodox and the efforts should be applauded. But in its current form, VPT hides behind at least three other veils of ignorance that make the method problematic and also introduce new hazards.

First, there is the veil of inequity. The division of labor is one thing: We see this increasingly in the profession with coauthored work. However, the designation of a singular person doing the theoretical work as the “principal researcher” and everyone else harvesting the data as “ignorant” is a troubling hierarchical framework. With VPT, we risk the principal researcher—without ever putting boots on the ground—taking advantage of their research assistants, who speak the language, understand the culture, and have the requisite networks to get the data collected. The data collection required for process tracing—whether it is doing interviews or conducting focus groups, going through archives, or digitalizing print media—is laborious and time-consuming. With VPT, we also risk the principal researcher—while staying safe in their home institution—overlooking the physical dangers of data collection. Fieldwork can be challenging not just because of the field site (e.g., Donbass, Ukraine) or the topic (e.g., prison gangs) but because of the research assistant’s gender, ethnicity, religion, sexual orientation, or disabilities. Even when the topic is seemingly innocuous, the risks can be immense. Consider female research assistants conducting focus groups in a place where their interactions with non-familial men are supposed to be limited. Consider ethnic Chinese (or even “Asian” more broadly) research assistants doing interviews in the aftermath of the Covid-19 pandemic. Consider research assistants with physical disabilities doing archival research in places with many steps and heavy doors but no ramps or elevators.

Second, there is the veil of impracticality. Unless there is a waiver, Institutional Review Board (IRB) protocols require informed consent—by the research subjects. It is hard to imagine an IRB approving a study involving

human subjects where the research personnel and the research subjects are both blind behind a theoretical veil. There is no way to guarantee the research assistants are collecting data within the guidelines; there is no way for anyone to ensure minimal risk and harm; and there is no way for debriefings to take place. Additionally, without any theoretical guidance, we risk one of two extremes. The first is that it puts a lot of undue pressure on the research assistants to ask the “right” questions or to find the “smoking gun” document without knowing the terrain. If the research assistant is a graduate student, there will be fears that “failure” on their part could result in funding cuts in the future or bad letters for the job market. The other extreme is that the research assistant feels unburdened in what type of data they collect: As long as the document pertains in any way to the dependent variable of interest, it is worth gathering. As a result, they end up collecting too much data, of which most—if not all—are irrelevant. In this case, the resources spent—money by the principal researcher and time by the research assistant—are wasted.

Third, there is the veil of inaccessibility. The costs associated with paying other people to do the data collection can be insurmountable. And as such, the approach is truly available only to those with the requisite resources (i.e., faculty members at elite institutions with large research budgets). It is hard to imagine a graduate student securing an APSA doctoral dissertation research improvement grant or an SSRC international dissertation research fellowship proposing such strategies: I have a theory about how X affects Y; give me the money to go to country A, where I will hire local research assistants to collect the data for me. Likewise, it is hard to imagine a cooperative project (e.g., Cooperative Congressional Election Study or Time-Sharing Experiments for Social Sciences), a survey center (e.g., any of the regional barometers), or a research institute (e.g., Minorities at Risk or Varieties of Democracy) built on the VPT principle. These aforementioned examples require substantial funds—but the generated data benefits multiple researchers. In effect, there is a public goods provision. In contrast, VPT—by virtue of focusing only on causal mechanisms—is strictly a private good that benefits just one principal researcher.

In short, VPT’s biggest contribution is calling much-needed attention to the biases with process tracing. However, by focusing on reducing these biases, VPT forfeits validity—or at a minimum, the ability to validate. The separation of the principal researcher doing the theorizing from the research assistants doing the data col-

lection means we cannot gauge the validity of the data. If there is a piece of evidence that is still missing—or turns out to be the only snapshot of a smoking gun—the principal researcher cannot differentiate whether the research assistant fabricated the data, did not look in the right place, or got the resulting data because of their identity (e.g., positively because they had inside access or negatively because they were shunned by the subjects).

One way to help with validation—and possibly address some of the inequity, impracticality, and inaccessibility concerns—is a hybrid arrangement that marries the traditional process tracing model with VPT. The principal researcher is still required to put their boots on the ground (i.e., to do their own data collection and coding). They can then employ a scaled-back version of VPT with research assistants on a random selection—not all—of the causal mechanisms. With the data collected through VPT, the principal researcher can then demonstrate where the data collection efforts were congruent, and where there were possibly omitted data sources. As a parallel, when Ted Gurr started the Polity dataset, and when Jennifer Gandhi created the dictatorship dataset, both did the first iteration of data collection and data coding. They both then subsequently expanded to have research assistants to help validate the measures. Because both had done the first iteration and there was a “baseline,” the research assistants did not have the undue fear of “failing” at or “ruining” the data collection efforts.

Alternatively, we can identify best practices from other methods and consider whether they can be adopted for process tracing—allowing us to minimize the two-headed bias concerns while maintaining validity. Two best practices show promise. The first comes from game theory. When solving for equilibria, game theorists have to consider off-path behavior. In non-formal parlance, this is the counterfactual. Given that VPT is focused on causal mechanisms, the implications for data collection are that the principal researcher should not just find evidence of X causing Y, but also why ~Y did not happen. For example, in the prisoner’s dilemma, the prisoner defecting is only half the story; the other half is why they did not cooperate. By focusing on this non-outcome, the principal researcher adds an additional observation—forcing them to think through alternative causal

outcomes. Methodologically, ruling out alternatives lends more credibility to the inferences than simply proving X causes Y. When researchers think through alternative causal outcomes, this serves as a “checklist” of all the different pieces of evidence needed to support their argument. And in doing so, the principal researcher can avoid unintentional omissions in data collection. Note that this is one of the steps in the Ricks and Liu (2018) guide for process tracing.

The other best practice is from experiments. There is a norm among experimentalists to submit pre-registered reports. These reports lay out their theoretical arguments and the research designs—inclusive of predictions, data sources, and plans for analyses. Registration happens through a third-party site (e.g., Evidence in Governance and Politics, Open Science Framework). This ensures minimal fishing by the researcher. Of course, there will be some slight deviations and possibly unexpected findings. However, with their hands tied through the pre-registered reports, the researcher can ensure transparency. Even if the researcher needs to update their argument given the data collected, there is a paper trail. Note that there are ongoing efforts to adopt such practices beyond experimental methods. The debates surrounding Data Access and Research Transparency (DA-RT) are a testament to scholars making qualitative methods, fieldwork, and process tracing transparent. And with transparency, we can identify our sources of bias.

The authors of VPT deserve credit for raising the issue of biases when doing process tracing. This is certainly a legitimate problem for future scholarship to address. However, in its current form, VPT is unequal, impractical, and inaccessible. It perpetuates a hierarchy. The research assistants who do the hard work—arguably the brunt of the work—are considered merely as assistants and coders and never as equals (i.e., researcher #2 or co-researcher). The inaccessibility of the method to those outside of elite institutions with large resources exacerbates another hierarchy. And it introduces new problems just as compelling as the selection and omission biases for which they seek to correct. All these built-in hierarchies—whether intentional or not—make the method one full of veils of ignorance.

References

Ricks, Jacob I. and Amy H. Liu. 2018. “Process-Tracing Research Designs: A Practical Guide.” *PS: Political Science and Politics* 51, no. 4 (October): 842-846.

Unpacking the Assumptions of the Veil of Ignorance to Highlight the Politics of Knowledge Production

Lauren M. MacLean

Indiana University-Bloomington

The space for qualitative research has unquestionably narrowed in political science since the 2000s, when the Perestroika movement demanded greater inclusion in the discipline, and qualitative and interpretive methods were institutionalized within APSA and CQMI (Monroe 2005).¹ Skeptics now question the basis for causal identification, frequently doubting whether qualitative and interpretive scholars have analyzed the evidence fairly and not simply “cherry-picked” what they were already inclined to believe. These critics even suggest that “bias” has led qualitative and interpretive researchers to omit the key causal factors altogether.

I share the overarching goal of Copestake, Goertz, and Haggard to enhance the credibility of qualitative methods. It is vital to establish the foundation for causal inference in qualitative and interpretive research. And, yet, I disagree with many of the basic assumptions and implications of their approach. The Veil of Ignorance Process Tracing methodology (VPT) will not solve old problems but instead create new ones. The approach is based on several key assumptions that do not hold. Moreover, the method amplifies existing power dynamics and increases inequalities involved in the production of knowledge.

To begin, VPT is predicated on several unfounded and unreasonable assumptions. First, the whole notion of a “veil of ignorance” depends critically on the Principal Researcher (PR) delegating the process of selecting data for analysis and conducting the preliminary coding to a Research Assistant (RA), who has no knowledge of the theoretical framework, hypotheses, or causal mechanisms. Still, the PR must develop some kind of criteria for the RA to select data and identify potential causal factors even in this first stage of the workflow. The authors suggest that the criteria might include: “the prestige of journals or book publishers, institutional affiliation, strength of personal connections with the event, the status of relevant archives, and so on” (this issue).

The problem is that the above types of criteria are essentially focused on the credibility of the sources rather than any substantive aspect of the source materials. Since this does not provide any systematic mechanism for filtering data sources, the search for data would likely be very labor-intensive and inefficient. The RA would be led to include a near-census rather than a selection or sample of materials from certain “more credible” sources. This effort would become very costly in terms of time and wages, which would be prohibitive for junior faculty or graduate students, who experience growing pressure to publish quickly but with limited research funds.

Furthermore, limiting the dataset to certain journals, book publishers, or institutional affiliations may introduce particular analytic biases, which could result in the exclusion of whole categories of causal factors. Certain publishing outlets or academic institutions might be more inclined to recognize some kinds of explanations as more “scientific” than others. For example, structuralist factors or constructivist explanations might not be emphasized in certain journals or by scholars working in some institutions.

Or, if the RA only consulted sources who had strong personal ties to the events, another sort of “bias” might creep into the dataset and subsequent coding. Working behind the veil of ignorance, the RA would be relying on the causal factors suggested by the sources themselves (this issue). And yet, the actors that produced these sources would have had the most to gain or lose from the telling of their story. As a result, these source materials may include strategic manipulation and obfuscation by the principals most directly involved in the set of events. In the archival documents or interviews, these actors may subconsciously marginalize or intentionally hide other causal factors that would disadvantage their own personal interests and self-representation.

Second, this approach assumes that the RA’s brain is a *tabula rasa* and that this person does not have any prior knowledge of theories, hypotheses, or potential causal

1 See also https://www.maxwell.syr.edu/moynihan/cqrm/CQMI_at_Syracuse/

mechanisms that might influence their data collection and preliminary analysis of potential causal factors. Using absolute terms, the authors suggest: “To do this they need *no knowledge* of the project being evaluated; *indeed ignorance* enhances the credibility of the coding as a test of the project’s influence,” (my emphasis; this issue).

But it is impossible to come to any data collection without any priors. To the contrary, most RAs would likely be graduate students who were immersed in digesting the canon of social science theories, prone to generating hypotheses for even everyday dilemmas. To carry the authors’ example forward, how would such an RA be able to analyze the causal factors that contributed to the development of democracy in Slovakia without any prior knowledge of theory or a range of hypotheses? Or without any prior contextual knowledge of the history of regime types in Slovakia, Eastern Europe, and the world during this particular time period?

Copstake, Goertz, and Haggard explain that the RA would probably be coding for Levitsky and Way’s (2010) variables, but it is unclear why this would be the case. Is there something inherently “logical” about a particular set of variables? This might suggest that the Veil of ignorance method would only produce causal factors that were patently obvious and might somehow be more centrist or middle of the road.

An RA might not think of and then code for hypotheses or causal mechanisms that were more radical or revolutionary, different, or unique somehow to this particular time period or set of events. These data and variables might be perfectly legitimate to consider, but they might not occur to the RA because of the admonition to stay behind the “veil of ignorance.” In contrast, I would contend that multidisciplinary knowledge of the area and historical period is critical to rigorous data collection and analysis.

Third, this approach assumes that what emerges from this process of data collection is “raw data” or what the authors term “causal claim Lego blocks” (this issue). Not being guided by any particular theoretical framework does not make these causal claims more “raw,” pure, or unbiased. As shown above, the resultant data are likely to contain a different set of biases. The conceptualization of causal claims as “Lego blocks,” which are relatively small, machine-made cubes of plastic, implies that causal factors are neatly linear and [click] fit perfectly together. This does not seem to capture the complex fluidity of the political world, with its nuance, blurred liminality, and subsequent need for contingent explanations.

Fourth, this approach assumes that the PR is then

capable of aggregating these Lego blocks in a way that is not biased or fulfilling their favored hypothesis. But, precisely because the dataset now claims the illusion of being raw and unbiased, the PR may be even less able to systematically evaluate or identify biases. This is particularly the case since the PR was kept separated from and was not involved in the collection and coding of the dataset. In both observational and experimental analysis, PRs are expected to be intensely knowledgeable about the creation and coding of the datasets they employ. In interpretive methodology, scholars do not seek the Truth through objectivity but rather conceptualize knowledge production as inherently subjective and thus think reflexively about their own and others’ positionality in that process (Yanow 2006). This arbitrary separation between RA and PR is neither required nor encouraged by other methodologies to be deemed systematic and rigorous.

In addition to being premised on several unreasonable assumptions, Copstake, Goertz, and Haggard are advocating for an approach that depends on the persistence of inequalities in the production of knowledge. John Rawls (1971) also draws on the notion of a “veil of ignorance” in the development of his theory of justice. For Rawls, individuals will make just decisions for the society as a whole, when they make their evaluation from the “original position” where they imagine themselves to enjoy the same equality as every other individual. Rawls is often critiqued that it is actually impossible to position oneself “behind the veil.” Others go even further by contending that the Rawlsian fantasy of liberal egalitarianism and social justice ignores the persistent reality of inequalities based on gender, sex, race, income, etc. Below, I highlight how several dimensions of power and inequality in the production of knowledge are magnified by VPI.

First, on a very basic level, the methodology requires the PR to have the financial resources and authority to hire at least one RA. This may not be feasible for graduate students, junior faculty, faculty at under-resourced institutions, or non-tenure track faculty without research funds. The veil of ignorance is founded on the separation of the project workflow among different roles; this may not be possible for many less senior scholars in weakly endowed research settings.

Second, the approach constructs a strict partition between the basic and concrete tasks of “extraction of data” or “data collection” by the RA and the more complex, abstract and theoretical “analysis” carried out by the PR. Importantly, not only are these tasks separated, but there is a clear hierarchy of value, with the PR con-

ducting, alone, what is deemed most difficult and important. Notably, this division of labor seems to provide little opportunity for mentoring and learning between the RA and PR. To be an RA is often considered to be an opportunity for a research apprenticeship, to learn from a more experienced PR about the art and craft of research, from beginning to end. Yet, this Veil of ignorance methodology severs the ongoing communication and collaboration between the RA and PR.

Third, the approach eliminates the power and agency of the human participants in the research process and magnifies the possibility for exploitation of these participants by the PR. The participants are not able to inform the development of the working theory, hypotheses, or causal mechanisms either. Instead, the human participants are portrayed as vessels from which the RA extracts data. The authors advocate that:

“Where possible, QuIP data collection entails “blindfolding” both field interviewers and interviewees from knowledge of this prior theory and empirical work. In particular, interviewers are not briefed about the theory, hypothesis, or precise causal mechanism being tested, and are informed only about the outcomes of interest” (this issue).

The “blindfolding” of field interviewers and human participants in these studies raises significant ethical questions about the level of deception that is acceptable within this approach. Do interviewers and study participants know enough to qualify as having “informed consent”? Recent deliberation through the QTD and the APSA Task Force on Human Subjects Research has concluded that our ethical commitments should always supersede other methodological considerations and goals.²

To conclude, our collective answer to questions about the credibility of qualitative research should not be to stick our heads in the sand and pretend that we

are ignorant of theory when we are assembling and interpreting our data. Neither quantitative nor experimental scholars take this tact. Quantitative scholars draw on extensive knowledge of theories, hypotheses, and causal mechanisms to identify which array of controls should be included in the model and which indicator best operationalizes these concepts. Likewise, experimentalists base the randomization of the treatment on deep knowledge of theory as well as the politics in a particular context.

Researchers should be systematic and open about the process used for the selection of cases and the collection of data. Reflexive openness throughout every stage of research can enhance the possibility of rigor as well as the ability to maintain our ethical commitments to our study participants and field sites (MacLean et al., forthcoming).

Researchers should not aim to hide behind a “veil of ignorance” in order to evaluate their evidence and interpret causal processes fairly. In contrast, we should read more widely—well beyond our subfields and disciplines—to gain a profound understanding of how politics, history, economics, cultures, and geographies intersect. We should also acknowledge explicitly and document openly how theory, research design, field research, and analysis inform each other in an iterative process throughout our research projects (Kapiszewski, MacLean, and Read, n.d.; 2015).

And instead of trying to “blindfold” our RAs and participants, we should be consulting with and learning from their lived experiences and knowledge (Fujii 2017). A less authoritarian and more collaborative methodology that decenters the sacred role of the PR, and instead legitimizes the knowledge of RAs and study participants, would make a greater contribution to new theoretical innovations and insights (Asiamah, Awal and MacLean, forthcoming).

References

- American Political Science Association. 2020. “Principles and Guidance for Human Subjects Research.” April 4. APSA Ad Hoc Committee on Human Subjects Research. https://www.apsanet.org/Portals/54/diversity%20and%20inclusion%20prgms/Ethics/Final_Principles%20with%20Guidance%20with%20intro.pdf?ver=2020-04-20-211740-153
- Asiamah, Gildfred, Mohamed Awal and Lauren M. MacLean. Forthcoming. “Collaboration for Conceptualizing and (Possibly) Decolonizing African Politics.” In *Symposium on Collaborative Methodologies*, edited by Mneesha Gellman and Pamina Firchow. *PS: Political Science & Politics*.
- Fujii, Lee Ann. 2017. *Interviewing in Social Science Research: A Relational Approach*. New York: Routledge.

² See Jacobs and Büthe (*forthcoming*) for a summary of the Qualitative Transparency Deliberation that highlights the primacy of ethical obligations to human participants. See also APSA’s Principles and Guidance from the APSA Ad Hoc Committee on Human Subjects Research: <https://connect.apsanet.org/hsr/principles-and-guidance/>.

- Jacobs, Alan M. and Tim Büthe. Forthcoming. "Transparency in Qualitative Research: An Overview of Key Findings and Implications." *Perspectives on Politics*.
- Kapiszewski, Diana, Lauren MacLean, and Benjamin L. Read. 2015. *Field Research in Political Science: Practices and Principles*. New York: Cambridge University Press.
- _____. n.d. "Dynamic Design: Iteration in Field-Based Research." Unpublished manuscript.
- MacLean, Lauren M. 2013. "The Power of the Interviewer." In *Interview Research in Political Science*, edited by Layna Mosley, 67-83. Ithaca: Cornell University Press.
- MacLean, Lauren, Elliot Posner, Susan Thompson, and Elisabeth Wood. Forthcoming. "Research Ethics and Human Subjects: A Reflexive Openness Approach." *Perspectives on Politics*.
- Monroe, Kristen Renwick, ed. 2005. *Perestroika!: The Rancous Rebellion in Political Science*. New Haven: Yale University Press.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.
- Yanow, Dvora. 2006. "Neither Rigorous Nor Objective?: Interrogating Criteria for Knowledge Claims in Interpretive Science." In *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*, edited by Dvora Yanow and Peregrine Schwartz-Shea, 67 – 88. Armonk: M.E. Sharpe.

Veil of ignorance Process Tracing and Contested Critical Observations

James Mahoney
Northwestern University

The problem of bias in the sources and evidence used to evaluate explanations in qualitative research is a longstanding topic (Goldthorpe 1991; Lustick 1996; Isacoff 2005; see also Thies 2002). The specific problem addressed by Copestake, Goertz, and Haggard's Veil of ignorance Process Tracing (VPT) does not concern biases built into the available evidence itself (e.g., systematic error across all sources). Rather, Copestake, Goertz, and Haggard are concerned with investigator-induced biases in the *selection of sources* and the *selection of specific pieces of evidence* from within a given source. As a solution to these selection problems, the authors propose a new method—Veil of ignorance Process Tracing—in which a "veiled" research assistant is used to help neutrally select sources and evidence as well as code and interpret that evidence.

How serious of a problem is "cherry picking" sources and evidence in case study and small-N research? I think the extent of this problem should not be exaggerated. Scholars often explicitly acknowledge that the existing literature presents competing interpretations of events. In the course of their case narratives, researchers frequently acknowledge differences among historians and historical experts in the interpretation of particular events or processes. Mahoney and Villegas (2006) argue that comparative-historical researchers routinely call

attention to major differences in the historiography if these differences are important for their arguments. The professional penalty of ignoring or downplaying obvious contrary evidence is to have one's scholarship regarded as misleading, poorly informed, and not worthy of publication in peer-reviewed outlets.

Researchers often discuss how they work to resolve differences among sources. They may follow the contemporary consensus position in the literature, drawing on more recent studies that highlight flaws in earlier interpretations. They may side with the interpretation that is grounded in the more careful and thorough research. They also often explore the implications of a particular contradiction in the evidence for the specific argument being advanced. Researchers often try to build their arguments by using "basic information"—information about well-known events that is relatively free of interpretation and not subject to a high level of contestation (Thies 2002, 353-354). Quite frequently, they discover that divisions within the literature do not have a bearing on their arguments. To provide an example from my work on colonialism in Spanish America (Mahoney 2010), historical demographers disagree quite a lot about the size of the indigenous population immediately prior to colonialism (e.g., did 5 million or 20 million people live in Central Mexico in 1492?). Given my argument, however, I was

able to note these differences without having to stake out a specific position one way or the other.

My view is that “cherry picking” is a major threat to validity for only select inferences in case study and small-N research. As such, I do not believe that Veil of ignorance Process Tracing is needed for all aspects of process tracing. However, I do see Veil of ignorance Process Tracing as offering an interesting solution to a particular and serious problem that often arises in case study and small-N research. This problem is the use of *contested critical observations* as the basis for reaching conclusions about the validity of a theory or proposition. VPT is a potentially powerful tool for dealing with contested critical observations in case study and small-N research.

Contested critical observations (CCOs) are at the intersection of critical observations and contested observations. A *critical observation* is evidence or data that is highly consequential in the evaluation of a proposition or theory. In the context of process tracing, a critical observation is a piece of evidence or a set of evidence that substantially alters prior beliefs about the validity of a proposition or theory (Barrenechea and Mahoney 2019; Fairfield and Charman 2019). A *contested observation* is a piece of evidence or a set of evidence (including perhaps an entire source) whose validity is disputed. The nature of this dispute may involve either descriptive or causal inference, and the dispute may or may not be a highly politicized and normatively charged.

The methodology developed by Copestake, Goertz, and Haggard offers a useful strategy for tackling the problem of contested critical observations. At an early stage of research, a veiled assistant could help the principal investigator (PI) determine whether specific observations are in fact CCOs that need to be carefully handled. Under this approach, the PI would direct the veiled assistant to various observations that could be important and contested. The advantage of having a veiled assistant at this stage is that she or he can remain neutral with respect to the question of whether these observations are critical and whether they are highly disputed. For qualitative researchers, it can be quite helpful to have a neutral viewpoint regarding the extent to which a critical observation is contested in the literature. The PI may be more apt to view a helpful CCO as less contested than it really is. Likewise, the PI may be more apt to view an unhelpful CCO as more contested than it really is. The use of a veiled assistant could help assure both the PI and the broader scholarly community that key observations have been correctly weighted and appropriately discussed when contested.

Figure 1 presents a two-by-two table in which observations are situated along two dimensions: criticalness and contestation. My argument is that Veil of ignorance Process Tracing is mainly useful for inferences that depend on observations from cell 2, where both criticalness and contestation are high. If critical observations are not highly contested (cell 4), VPT is not needed to ensure unbiased interpretation. If contested observations are not critical (cell 1), the payoff of VPT is probably not worth the effort, time, and resources. I am arguing that VPT is a methodology for dealing with inferences that depend on observations that are both very influential and highly contested—that is, CCOs.

Figure 1
Types of Observations

		CRITICALNESS	
		Low	High
CONTESTATION	High	<i>Contested Routine Observations</i> Cell 1	<i>Contested Critical Observations (CCOs)</i> Cell 2
	Low	<i>Routine Observations</i> Cell 3	<i>Critical Observations</i> Cell 4

Allow me to discuss the example of late-colonial Chile from my own work on colonialism and development in Spanish America (Mahoney 2010, 174-76). One part of the argument about Chile turned on the question of whether this region was a marginal colonial territory during the late Bourbon period of Spanish colonialism (from about 1770 to 1820). I worked to gather all the secondary sources I could find relevant to this issue, and I concluded that the bulk of the evidence suggested that Chile was a colonial periphery. However, a few pieces of evidence and one particular source suggested that it was actually a vibrant colonial region. I noted in the text that some evidence was consistent with an alternative view, and I briefly discussed the contradicting source in a footnote (331). However, for this specific set of observations, I believe that VPT would have been an attractive approach. It would have helped resolve what I still consider to be an important question regarding my argument: Would a neutral (though knowledgeable) reader looking at this set of evidence arrive at the conclusion that Chile was a late-colonial periphery?

Another example is Jack Levy’s (1990-91; 2007) work on the origins of the continental war in Europe that led to World War I (see the summary in Copestake, Goertz, and Haggard’s Figure 2, this issue). How might VPT

have been useful to Levy in developing and assessing explanations of the continental war? It seems doubtful that a veiled assistant would have been helpful with the basic theoretical framing of Levy's explanation, which views the war as a product of Russian and German mobilization into the local war between Austria and Serbia. Nor would a veiled assistant have been helpful for informing Levy about the kinds of factors that are normally emphasized in explanations of wars in general and the continental war in particular. Levy knows as much as anyone about these topics.

Instead, the assistant could help in two other ways. First, the assistant could help decide whether any observations used by Levy are contested critical observations. For instance, consider the causal chain (circles 1-6 in the summary figure) leading to the local war between Austria and Serbia. A veiled assistant could examine these individual linkages and identify which ones are clearly supported by the literature and which ones seem more contestable. This information can be shared with Levy, who could then report the disputes in an appendix or even better using an active citation system (Moravcsik 2014). Second, for these contested linkages, the assistant can help identify the alternative factors and explanations that Levy may want to consider. For instance, what other factors besides those listed (circles 1-4) tipped the balance in favor of the Austrians initiating war with Serbia (circle 5)? Are any of these additional factors more important than those listed by Levy? At the end of the day, Levy's original explanation may not change. Nevertheless, Levy could report that a veiled assistant reviewed the individual linkages in the chain, looking for any obvious discrepancies in the sources and evidence.

What about selecting the sources to use when conducting research on a given topic? Is VPT useful for avoiding biases with that task? With case study and small-N research, my view is that a veiled assistant would generally offer only modest value-added in the selection of sources. The main reason why is that excellent case study and small-N research requires that the investigator her or himself become an expert in the literature for every one of the cases under analysis. Developing this expertise often requires years of reading and studying the literature and sources related to the cases under study. I think it would signal serious problems if a novice research assistant could come along and point out important additional sources of which the principal investigator was unaware.

The place where the veiled assistant could be helpful is identifying sources for secondary cases in small-N research. These cases receive only brief treatment because

they are used for strategic purposes in comparative research, such as introducing a control case when evaluating a particular causal factor. Again I see the role of the assistant as targeted: The assistant would gather sources relevant to the strategic way in which the researcher is using the case. This kind of research assistance is already often used by scholars. Professors often hire RAs to help with gathering literature and evidence for secondary cases. The difference with VPT is that the assistant must not know in advance what kinds of evidence will and will not support the PI's hypotheses. Hence, the assistant has every incentive to be as neutral as possible when identifying sources.

Perhaps the main context in which VPT can be useful for selecting sources is medium-N studies (e.g., 20 to 100 cases). With these studies, the principal investigator usually cannot be an expert on every single case under analysis. Moreover, the goal of the PI is not to develop a rich sequential narrative for each case. Instead, the PI is interested in specific propositions (e.g., a specific causal linkage) and the observations that do and not support their validity. For instance, Haggard and Kaufman (2012) test theories proposing that inequality shapes democratization via the mechanism of distributional conflict. They code a medium number of cases across a small number of variables, including especially the variable of distributional conflict. Would an unbiased researcher arrive at the same codes for all cases? Certainly, asking and answering this question is good social science. I think the veiled coding of particular variables often makes sense when conducting medium-N research. With high-quality qualitative research, such as Haggard and Kaufman (2012), the investigators have worked hard to put the veil over themselves throughout the research process. Well-trained qualitative researchers repeatedly ask themselves how an objective critic might find and use evidence to show that their argument is wrong. That said, good scientific practice suggests that actually putting the veil on from the start is far better than trying to imagine wearing a veil after you have already seen much evidence.

Practical issues of implementation would doubtless stand as obstacles to using even the focused kind of VPT that I have discussed. Perhaps most important, the new central actor in this methodology—the veiled research assistant—would need to have appropriate characteristics to help carry out high-quality VPT. For instance, let us pretend that we seek to reanalyze Haggard and Kaufman (2012) using a veiled assistant. What kind of assistant would be most appropriate? To begin, we would need an assistant who either does not know this Haggard and

Kaufman article or who does not know we are reconsidering its findings. The assistant probably would need to have a background in comparative politics and excellent overall research skills. The assistant would also need a great deal of time and energy to go through every single one of Haggard and Kaufman's cases! It seems unrealistic to imagine that a single research assistant could do a good job carefully coding every single case using an unbiased sample of sources.

Instead, what I have argued in this paper is that one might use VPT to deal with the contested critical observations (CCOs) in Haggard and Kaufman's (2012) article. In this particular article, the CCOs would be particular codes of distributional conflict that are influential but potentially wrong, given all of the evidence. The challenge would be identifying which specific codes fall into the CCO category. Haggard and Kaufman themselves explicitly note some borderline cases, indicating that they coded them conservatively to work in favor of the theories they critically assess. Nevertheless, one might still want to direct a neutral veiled assistant to explore these cases more carefully.

Methodologists could also develop general rules and heuristics for veiled assistants to use when locating potential CCOs. When deciding if an observation is a CCO, one can start by asking about the extent to which sources agree and how many sources agree or disagree concerning the truth of the observation. What is the nature

of the disagreement? One could then follow guidelines for weighting different sources and their interpretations. Some of the factors considered might include the year of publication of the source; the quality of the press, journal, or outlet that published the source; the extent to which the authors of the source have potential biases; other author characteristics, such as reputation, job rank, and employer; the extent to which the source's evidence is directly on topic; and the kind and quality of the primary sources that are used to derive conclusions. In developing rules, researchers could certainly draw on insights from historians, who have thought a lot about how to deal with conflicting sources and interpretations. Through the development of more objective rules for identifying contested critical observations, scholars could make the most of a veiled assistant's time and skills.

I want to end on a note of thanks to the authors for helping us think through the ways in which qualitative research might be enhanced by introducing neutral assistants who do not have a stake in the results. Whether or not Veil of Ignorance Process Tracing becomes a significant new instrument in the qualitative toolkit, Copestake, Goertz, and Haggard have done a service by stimulating this discussion about the role of researcher biases in process tracing and qualitative analysis. Their work encourages us to repeatedly ask and seek answers to the following question: What would an objective investigator find?

References

- Barrenechea, Rodrigo, and James Mahoney. 2019. "A Set-Theoretic Approach to Bayesian Process Tracing." *Sociological Methods and Research* 48, no. 3 (August): 451-484.
- 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, no. 1 (March): 154-167.
- Goldthorpe, John H. 1991. "The Uses of History in Sociology: Reflections on Some Recent Tendencies." *British Journal of Sociology* 42, no. 2 (June): 211-30.
- Haggard, Stephan, and Robert R. Kaufman. 2012. "Inequality and Regime Change: Democratic Transitions and the Stability of Democratic Rule." *American Political Science Review* 106, no. 3 (August): 495-516.
- Isacoff, Jonathan B. 2005. "Writing the Arab-Israeli Conflict: Historical Bias and the Use of History in Political Science." *Perspectives on Politics* 3, no. 1 (March): 71-88.
- Levy, Jack S. 1990/91. "Preferences, Constraints, and Choices in July 1914." *International Security* 15, no. 3 (Winter): 151-186.
- Levy, Jack S. 2007. "The Role of Necessary Conditions in the Outbreak of World War I." In *Explaining War and Peace: Case Studies and Necessary Condition Counterfactuals*, edited by Gary Goertz and Jack S. Levy, 47-84. London: Routledge.
- Lustick, Ian. 1996. "History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias." *American Political Science Review* 90, no. 3 (September): 605-618.
- Mahoney, James. 2010. *Colonialism and Postcolonial Development: Spanish America in Comparative Perspective*. New York: Cambridge University Press.
- Mahoney, James, and Celso M. Villegas. 2007. "Historical Enquiry and Comparative Politics." In *Oxford Handbook of Comparative Politics*, edited by Carlos Boix and Susan C. Stokes, 73-89. Oxford: Oxford University Press.

The Veil of Ignorance Process Tracing (VPT) Methodology: Some Practical Considerations and Limitations (to Decoupling)

Sandrino Smeets
Radboud University

With the introduction of their new approach to process tracing (PT), Copestake, Goertz, and Haggard seek to tackle what is arguably the most prevalent and most serious shortcoming in qualitative case study research: the "cherry-picking" of cases, data sources, or evidence to fit the claim or theory that the researcher is trying to substantiate. The authors should also be complimented for their attempt to make process tracing analyses "more rigorous and transparent" (this issue). By now, we have a variety of handbooks that offer high quality, hands-on advice on how to set up and conduct a process tracing analysis (see e.g., Beach and Pedersen 2019; Bennett and Checkel 2014). However, the term "process tracing" is still being stretched by scholars, who essentially provide us with a narrative reconstruction of a (causal) process. The question "how valid are my causal inferences compared to the inferences of others?" remains central to process tracing analyses. And we should welcome attempts to further develop the methodology of *causal attribution*, specifically for within-case analyses, in which we deal with one or a few instances of a cause-and-effect relationship.

The proof of the pudding, however, is in the eating. So, the relevant question is: Would I be able to apply the VPT methodology in my own research and, more importantly, would it lead to better causal process analyses? One central causal claim in my research concerns the important role and influence of institutional actors—within the context of the European Union (EU)—vis-à-vis national actors in determining the course and outcome of international negotiations (see e.g., Smeets and Beach 2020). Within the field of EU studies, the most influential actors are generally considered to be Germany and, to a lesser extent, France (see e.g., Moravcsik 1998). If I were to go against the dominant interpretations that it is "all about Germany," and seek to show that institution-

al actors, like the European Commission and European Council President, play a prominent role, would VPT help me do a better job?

The solution that Copestake, Goertz, and Haggard offer to the problem of cherry-picking is straightforward: to *decouple* the process of collecting and recording/coding the raw data (exploratory phase) from the process of interpreting/evaluating that data in relation to theorized causal mechanism (confirmatory phase). The exploratory work is to be done by a competent, but not too informed or involved, research assistant (RA), while the confirmatory work is left for the principal researcher (PR). To be able to conduct the fieldwork, the RA should be somewhat familiar with the field but not too informed about the PR's theoretical predispositions. This would allow the RA to collect "pure," "untainted" empirical material regarding the role and influence of particular actors—what Copestake, Goertz, and Haggard refer to as "causal claim Lego blocks" (this issue). The PR would subsequently receive this complete set of blocks of causal process observations (CPO), and therefore not be able—or less inclined—to focus only on the blocks that support her theoretical point of view.

There is some merit to the idea of "blindfolding" the data gatherer. Confirmation bias is a persistent problem, specifically (but not only) in the social sciences, in which definitive proof in favor or against a particular theorized hypothesis is often hard to come by. In the field of EU studies, any scholar will be able to find supporting evidence for the claim that either Germany or the European Commission was very influential in the negotiations, particularly when interviewing stakeholders from one of these two sides. In fact, even a view that Malta has been very influential in EU negotiations can be grounded in *some* empirical observations. The question is whether this evidence, and the interpretation of this evidence, holds

up in the eyes of well-informed peers.

I am not sure that VPT is on the right track by trying to “objectify” the process of data gathering and primary analyses. The core of the problem is *not* that researchers generally seek supporting evidence for their theoretical claims and larger research programs. This is very much the “normal” or “positivist” approach to scientific research, along the lines of Karl Popper, Thomas Kuhn, and Imre Lakatos. The core of the problem is that the way in which case study researchers *evaluate* or *weigh* different pieces of evidence is often obscured.

We need not concern ourselves with such meta-theoretical discussions about what constitutes facts and evidence but rather focus on a more practical limitation: Once the RA has collected and recorded all the raw CPOs, it is still unclear how the PR should go about the job of evaluating this extensive collection of raw data. Provided that she cannot simply present or include the entire “set of blocks” in the textual analysis, which pieces of evidence make it into the theoretical synthesis, and on what grounds? The core problem with “cherry-picking,” in my view, is *not* that researchers pick and choose amongst pieces of evidence. One of the major advantages of working with CPOs instead of data set observations (DSOs) is that different weight can, and should, be attached to individual pieces of evidence. To go back to my example, if out of the ten people I interviewed, eight indicated that the German Chancellor played a prominent role in the negotiations, that number as such would not constitute meaningful evidence in a process tracing analysis. Process tracing is not about stockpiling pieces of evidence but about identifying the weakest link in a causal argument.

The bigger problem with “cherry-picking” is that the researcher attaches far too much weight to some of these observations—for example, if the researcher finds a few circumstantial pieces of evidence (e.g., through interviews, archival research, or secondary data such as press reports) that indicate that the European Commission played a notable role and then uses these as leverage to support a comprehensive theory of Commission leadership. This weighing/evaluating problem is *not* overcome by using VPT. In fact, the use of “blindfolded RAs” for data gathering and primary analysis (recording/coding) might aggravate the problem. This is because the process of weighing and evaluating pieces of evidence starts *during* the data gathering. In elite interviewing, the goal is not to harvest as many individual pieces of evidence as possible. The purpose is to get at real, in-depth, and

meaningful evidence in light of the claims that the researcher is trying to investigate.

This means that the interviewer needs to be *embedded* in the field and topic, so that she is able to judge and evaluate the statements of the interviewee and ask follow-up questions on the evidence to back up a specific (causal) claim. For instance, a claim by the interviewee that the Commission was very influential in a particular phase of the negotiations is substantiated by a comparison of the final Treaty text with the initial (“untainted”) draft text presented by the Commission. The interviewer needs to be able to pick up subtle signals that are being sent and respond to them accordingly.

This is why such interviews can rarely be done through standardized or even semi-structured interviews. It also means that interviews should be conducted by deeply informed and involved researchers. The interviewees, whether they are government representatives or high-level civil servants, already have an informational advantage over the interviewer. This means that the latter can either be “taken for a ride” or that she is sent home with very general data (e.g., that the German Chancellor, Angela Merkel, was very influential at an EU Summit). Even more problematic, interviews done by “outsider-interviewers” tend to replicate the conventional wisdom and prevalent views on a matter, instead of allowing the researcher to look beyond these. The PR would thus very likely be presented with a lot of transcripts that are either biased or of little added value.

Finally, how to proceed with VPT (and other types of PT for that matter)? VPT highlights an important problem/limitation with many existing PT analyses, which are still characterized by confirmation biases and limited source criticism. The distinction that VPT makes between an exploratory and a confirmatory phase in data processing is very interesting from an *analytical* point of view. In qualitative, case study analysis, many scholars still tend to seek and therefore find the evidence that they were looking for to substantiate pre-held claims. I support the call for more openness, perhaps even a “whiff of inductivism” when gathering and analyzing data. However, I am not sure whether the two phases—exploratory and confirmatory—can be neatly separated *on an empirical level*. A significant amount of (often implicit) selecting already takes place in the very first, exploratory stage of data gathering, irrespective of whether an RA or PR is doing the fieldwork.

The solution to this problem lies not in *ex-ante* “neutral” data gathering. The solution lies more in the *ex-post* critical evaluation of the pieces of evidence that the PR

decides to put forward in support of her theoretical claim(s). “Cherry-picking” by individual researchers is not something we can weed out entirely by means of a rigorous, transparent methodology. We will always need critical evaluation, by means of informed peer review, of the causal weight/leverage that is attached to certain

pieces of evidence. In the terminology of Copestake, Goertz, and Haggard, it is not that all the “Lego blocks” should converge on the favored causal mechanism, but rather about whether the structure that is built with these blocks can carry the theoretical weight that is put on it.

References

- Beach, Derek, and Rasmus Brun Pedersen. 2019. *Process-Tracing Methods. Foundations and Guidelines*, 2nd ed. Ann Arbor: University of Michigan Press.
- Bennett, Andrew, and Jeffrey T. Checkel, eds. 2014. *Process Tracing. From Metaphor to Analytical Tool*. Cambridge: Cambridge University Press.
- Moravcsik, Andrew. 1998. *The Choice for Europe. Social Purpose and State Power from Messina to Maastricht*. Ithaca: Cornell University Press.
- Smeets, Sandrino, and Derek Beach. 2020. “Intergovernmentalism and its Implications—New Institutional Leadership in Major EU Reforms.” *Journal of European Public Policy* 27(8) 1137-1156.

Lost in Translation: A Reprise on the Veil of Ignorance Process Tracing Method

James Copestake
University of Bath

Gary Goertz
University of Notre Dame

Stephan Haggard¹
University of California-San Diego

Introduction

To say that our modest proposal for Veil of Ignorance Process Tracing (VPT) did not get much support amongst this prominent group is an understatement. Nonetheless, the symposium raises important questions about the current state of process-tracing methodology that still need to be addressed. Moreover, it suggests there are competing proposals for handling the issues we have raised, including the Bayesian vision outlined with great clarity by Fairfield. We focus on the four questions we consider most important: (1) the role of causal mechanisms and the causal claims that constitute mechanisms in process tracing; (2) the extent to which our critics think there is a problem to be addressed in the first place; (3) practical issues; and (4) normative questions. We believe these normative issues arise not simply with respect to our proposed method but in the organization of research—and particularly large-scale research projects—more generally.

We start, however, with a brief review of the QuIP methodology that inspired our proposal, because the few comments it elicited suggest a limited understanding of

it. This recap also serves as an introduction to our four overarching questions regarding the VPT methodology.

Understanding the QuIP

Beach notes that the QuIP methodology “reads as if it were originally framed to allay funder concerns about evaluation case studies by including a number of buzzwords from experimentalist research.” In fact, the QuIP is a product of more than six years of action research in the context of wider efforts to strengthen the status and use of qualitative methods in the massive field of evaluation studies (e.g., Stern et al. 2012). Many commissioners of evaluations have been concerned that causal claims derived from experimental designs are both atheoretical and contain a quite obvious cherry-picking bias by typically testing the hypothesis that a favored intervention (i.e., the treatment) worked. The QuIP is one response to these concerns and to the demand for better qualitative complements to experimental approaches. It also drew explicitly on Copestake’s reading of the process tracing literature (e.g., Copestake, Morsink, and Remnant 2019, 37–38).

¹ The order of the authors is alphabetical. Thanks to the QuIP team—Fiona Remnant, Rebekah Avard, Marlies Morsink, and Steve Powell—for many discussions. Thanks, as well, to Jim Mahoney for comments on an earlier draft.

The QuIP set out primarily to do one thing as clearly and transparently as possible: to find out what selected respondents perceive to be the most important causal drivers of change in selected domains of their lives. As in the new generation of experimental work that also seeks to tap more open-ended interviews—most notably the work of Paluck (2010)—these interviews were seen as tapping into the vast storehouse of knowledge held by participants on social relationships, feedback mechanisms, and individual motives for action. If qualitative process tracing methods are going to add any value to choice- or game-theoretic models, it can only come by probing these connections.

A second aim of the QuIP research has been to improve on the way respondents' views can be aggregated in order to generalize to wider populations. To this end it has developed methods for systematically coding discrete causal claims. It uses diagrams to visualize the corresponding causal mechanisms, while recognizing that these diagrams typically reflect only the bare skeleton of the narratives that underpin them.

A number of contributors to the symposium fault the QuIP for uncritically accepting that what subjects say is causally—or even factually—correct. The QuIP addresses this problem in some detail. First, it goes to considerable lengths to reduce potential bias in subject responses. Second, QuIP studies are always integrated with other sources of evidence. Third, it is often useful to know what respondents perceive to be true, even if it is rejected. Lastly, the problem is hardly unique to the QuIP, indeed it is almost universal to all research. For example, it comes up in the use of archival material ultimately crafted by self-interested actors, in secondary sources in which authors have preferences for particular explanations, and in structured as well as open-ended interviews.

Is confirmation bias a major problem? A surprising outcome of this symposium for us was the extent to which respondents tended to minimize the problem the VPT was seeking to address. This contrasts sharply with the attention it attracts in other fields, including psychology (Chambers 2017). Evaluators and policy makers often cite confirmation bias, along with other response biases, as a reason for doubting what people are reported to have said rather than what they have revealed through observable actions.

Is blindfolding a practical way of responding to this problem? QuIP experience of partly blindfolding researchers as a response to this problem is now extensive. The approach does limit the scope for more pointed probing about a particular intervention, and for this reason is

generally combined in different ways with unblindfolded data collection. The pay-off is that the QuIP empowers researchers and respondents to explore drivers of change in specified outcome domains through open-ended conversations that avoid steering respondents toward favored causal stories. Researchers and respondents can thereby also talk on a more equal footing, because neither is privileged with detailed prior knowledge of what the commissioner of the research is most interested in. The approach also increases the importance of working collaboratively with skilled and well-briefed interviewers—in sharp contrast to studies that limit the scope for field researchers to exercise their skills and discretion.

A second practical objection is that data collection should not be separated from data analysis and interpretation. This criticism overestimates the nature of the firewalls that the QuIP seeks to create. Falling somewhere between the total immersion of the lone researcher and the rigid task demarcation common with quantitative surveys, the QuIP research process hinges on team work and iteration: Not a strict formalization of roles but a structured process of deliberation based on the shared goal of articulating as accurately as possible the views of those interviewed (Copestake et al. 2018; Copestake, Davies, and Remnant, 2019).

Is blindfolding ethical? QuIP research is motivated by the goal of giving voice to the often-ignored experience of intended beneficiaries of international development activities in a more credible way. The approach reduces power imbalances between interviewers and respondents arising from an asymmetry in prior knowledge about what the study is seeking to establish. Transparently limiting what interviewers and respondents know in advance also enhances the credibility that commissioners attach to the findings (Copestake, Morsink, and Remnant 2019). Of course, the reasons for blindfolding should be fully explained to all involved, as is the case for blinding in randomized control trials. Respondents should also consent, be able to withdraw at any time, and be informed about who they can consult if they have further queries about the research. These principles apply equally to interviewers; their full understanding and assent to being temporarily blindfolded is critical to the whole approach, as is their systematic debriefing. QuIP studies have been routinely subjected to ethical approval and found consistent with extant research norms.

Having briefly clarified the QuIP approach, we return to the central question of this symposium: Is there something about this experience that could be extended to the evidence that makes up much of what qualitative process

tracing does more widely; namely, extracting evidence—including causal claims made by the principles—from primary and secondary sources?

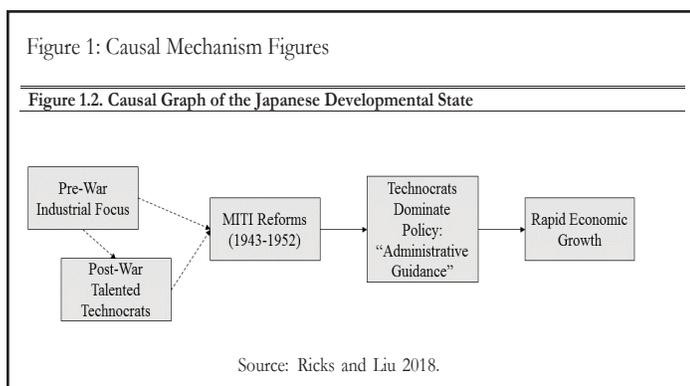
Causal Mechanisms and Causal Claims

It is useful to start from a central goal of our methodology—and we think of all process tracing—which is the generation of causal mechanism figures. With the exception of the commentary by Derek Beach (this issue), none of the pieces makes significant mention of causal mechanisms at all. The symposium has made it clear to us that there remain a variety of unresolved issues about the role of causal mechanism figures in process tracing, including the semantics as well as ontology surrounding them. Given that they are a central goal of the proposed approach, it is worth considering these issues in some detail.

As a starting point, consider Robert Axelrod's 1976 *Structure of Decision* (a book cited nearly 4500 times). Axelrod uses a methodology that has many similarities with ours. He defines a cognitive map exactly the way we define a causal mechanism:

"A cognitive map is designed to capture the structure of the causal assertions of a person with respect to a particular policy domain... A cognitive map has only two basic types of elements: concepts and causal beliefs. The concepts are treated as variables [boxes], and the causal beliefs are treated as relationships [arrows] between the variables" (Axelrod 1976, 58).

As his main example, he and his graduate students at Berkeley approached an archival source and coded recently declassified transcripts of elite committee meetings regarding the deliberations of the British Eastern Committee in 1918.



One can find ample examples of causal mechanism figures in the qualitative comparative politics space, in-

cluding exemplary ones generated by participants in this roundtable. As an example, we reproduce a causal mechanism figure, as Figure 1, from Ricks and Liu's (2018) article on process tracing.

This is quite typical of causal mechanisms that include exactly what Axelrod (1976) described above: entities and causal relations between them. If one breaks up these causal mechanism figures they decompose into a discrete number of what we call causal claims or causal Lego blocks. So, for example, in the one reproduced here explaining the Japanese developmental state, there are two causal claims that correspond to the two solid arrows in the figure.²

This example from Ricks and Liu (2018) is very similar in spirit to how we suggest that the VPT might be used to evaluate extant qualitative research, for example, the Levitsky and Way (2010) claims made in their outstanding case studies. If we were to use this technique to assess their causal claims, we would first need to draw the causal mechanism figure corresponding to their theory and then investigate the validity and veracity of the causal Lego blocks used in it. In effect, these are the steps in the theorized and observed causal process that figures such as those in Ricks and Liu (2018) and Beach and Pedersen (2013) use in outlining the approach.

Beach (this issue) expresses concern that these causal mechanism figures are in fact descriptive ("the VPT appears to understand mechanisms as a descriptive series of events"). This is a serious misreading, although the fault may be ours. A causal mechanism—and a related causal mechanism figure—consists of a set of postulated causal relationships. Beach argues that Jack Levy's work in this vein is descriptive, but this is not correct: His claims are causal.³ We see the core of process tracing to reside in the articulation of causal mechanism figures.

Such causal mechanism figures are generated through both deductive and inductive research processes as Beach and Pedersen (2013) point out. At no point do we say that the generation of either theory or causal claims can be achieved through induction; this is a red herring. For example, game theory provides a large class of deductive models that can be mapped as game trees and thus have figures embodying causal mechanisms. Contrary to our critics, we in fact believe—as we are sure they do—that work in the qualitative process tracing tradition is an iterative process between theory, an extant literature, and

2 The Ricks and Liu (2018) example also illustrates the need for a much more complex and sophisticated set of semantics for these figures. They use dotted lines to indicate relevant background information that is not causally relevant. In the online appendix they write: "Alternative information about the points prior to this time period, although interesting, is not essential to this causal story, and thus we link it with dotted arrows," (Ricks and Liu 2018, 5).

3 Goertz co-edited the volume with Levy (2007) containing his World War I analysis.

within-case causal inference.

Causal Claims, Evidence, and Sources

A great deal of confusion among the participants was caused by some problematic features of our own workflow figure. And those problems go directly to the discussion in the symposium about the relationship between causal claims and evidence but also to our focus on sources. Few, if any, of the participants picked up on a core feature of our approach, which is to force thinking about the dependence of qualitative research on sources who are not only potentially biased but the very selection of whom might also be biased.

In the column dealing with processing and collecting causal claims there was some quite ambiguous language—particularly “data collection and coding”—that was not useful in expressing what the methodology is about. What exactly is this “data collection and coding”? As we will show, the lines between causal claims and evidence are not as sharp as the participants think.

A concrete example dealing with secondary sources might help. Goertz was attracted to the QuIP methodology because it was similar to what he did in preparation for a 2007 anthology with Jack Levy on the causes of World War I. In writing the parts of the book dealing with the war, he downloaded all of the articles published in the journal *International Security* dealing with the conflict. Goertz was in no way an expert on World War I; this is important to underscore given the skepticism about research assistants doing this work. But he was interested in systematically gathering important causal claims about the war. He was in no position to evaluate the substance of the claims but was very concerned about correctly cataloging and describing those claims. That is exactly what QuIP does and what the veil of ignorance asks research assistants to do.

But what exactly is being collected here? Perhaps the core theoretical critique of our approach is that it confuses causal claims or hypotheses with the evidence that is used to support them. But these lines are far less clear than the participants think. Moreover, the responses do not address some of the defining features of qualitative work, which include reliance on interviews, primary and archival materials, and secondary sources. These sources do not contain “raw evidence,” but causal claims. In their exposition of Bayesian methods, for example, Fairfield and Charman (2017) cite as a piece of “evidence” what is in fact a causal claim: An informant tells Fairfield that particular measure was “ruled out as infeasible on every such occasion due to resistance from the right coalition”

(374). The causal claim of the participants is in effect taken as a piece of evidence in support of the hypothesis, aka causal mechanism, of the analysts.

Fairfield chides us for muddying the relationship between causal claims (hypotheses) and the evidence collected for them. However, we clearly did not fully recognize this problem at the time and did not adequately think through the important distinction between cataloguing causal claims, collecting causal claims from sources that support causal claims, and collecting other types of evidence that might support a causal claim. This method is addressed to the first of these two processes, and not to the third.

Fairfield and Charman (forthcoming) provide one ideal principle that we completely support regarding information and process tracing: There should be informational completeness and all relevant information should be taken into account. This is precisely the goal of the veil of ignorance approach, to provide a set of causal claims regarding the target event that are as complete as possible. This can only be done based on an unbiased, systematic search of the relevant sources.

Is There a Problem?

Scholars should not get credit for solving problems that either don't exist or are of their own making. One of the things we have learned from the symposium is that there is a range of opinion about the severity of the cherry-picking or researcher bias problem, which the veil of ignorance is designed to address. We were honestly surprised by this, because arguably the entire evolution of the field of qualitative methods has been aimed at increasing the reliability of—and reducing possible biases in—qualitative analysis.

We reiterate that the first step in this process—selecting data sources—is nothing more than structured due diligence. The process we have identified assures the researcher that she is accessing a full and as unbiased set of sources as possible, what one reviewer dismissively calls a “census” or catalogue. But particularly in fields where issues are contested—including through bias arising from power or discrimination—such assurances should not be dismissed. We need to not only be aware of the universe of causal claims, but to satisfy ourselves and others that we are drawing them from an appropriately diverse set of sources in order to minimize bias.

Mahoney (this issue) is the one reviewer who sees a role for the methodology in what he calls “contested observations.” We take this to mean some causal links in the causal mechanism are noncontroversial while others gen-

erate much debate. If one looks at the causal mechanisms surrounding virtually any social science outcome of serious substantive interest, there are almost certainly going to be aspects of those causal claims that are controversial. Minimally, such an assessment requires a reasonably systematic survey of competing causal claims and sources, precisely the activity that the veil of ignorance approach is designed to strengthen.

Mahoney is the most straightforward about his priors when he says that the extent of the source bias problem should not be exaggerated. Beach (this issue) says the same thing: “The authors postulate—without evidence—that confirmation bias is an endemic problem in process tracing, insinuating that much existing PT consists of “just-so stories” where researchers search for evidence that confirms their pet theory. In reality, contrary to what these authors write, confirmation bias in non-experimental science is not avoided by “blinding,” but instead by being one’s own strongest critic when evaluating empirical evidence, achieved in PT through thorough and transparent assessment of the uniqueness of found evidence...” These and other contributions effectively appeal—as does Fairfield—to virtue. In addition to the role of peer review raised by Beach, the assessment of any given piece of research would thus rely at least in part on personal knowledge of the ethical standards of the researcher or on a good-will assumption.

Again, a concrete example underlines that the problem is an ongoing one. Haggard edits a social science journal focused on East and Southeast Asia and recently published a roundtable on four new books dealing with power transition theories and their applicability to China’s rise (Shiffrinson and Haggard 2020). All were based on high-quality qualitative historical scholarship on other power transitions, mostly in Europe. All wrote closing sections in which they speculated on the Chinese case. Yet one of the critics in the roundtable, David Kang, argued that whatever the merits of their primary work, the authors had not spent adequate time surveying either competing sources or contrary evidence in seeking to extend their inferences to the Chinese case. The “thinner” the cases used for process tracing purposes, the more risk that such biases might in fact arise if only as a practical matter. How much effort can a researcher devote to a two- or three-page case study?

We should note that our approach is in fact compatible with Fairfield’s Bayesian approach and in fact addresses a particular problem with it. The Bayesian approach stresses the need for *mutually exclusive and exhaustive* hypotheses. How are we to know that the hypotheses are

really exhaustive? One way is to broaden the search by involving other researchers who are unaware of the leading hypothesis—precisely what a systematic veil of ignorance search of sources does.

One thing we were not very clear about was the engagement and other activities of the principal investigator. Normally this person will be highly engaged in the topic under consideration and in field settings will certainly have extensive field research experience. Smeets and MacLean in the symposium both stress the value of being deeply embedded in the interview process. For example, Smeets (this issue) writes: “This means that the interviewer needs to be *embedded* in the field and topic, so that she is able to judge and evaluate the statements of the interviewee and ask follow-up questions on the evidence to back up a specific (causal) claim” (emphasis in the original). This is similar to Mahoney’s comment that the principal investigator will have, ideally, extensive knowledge about relevant debates.

If we go back to back to the QuIP setting, the commissioners of these studies have started out with a deep prior interest in the development activities being evaluated. The role of the QuIP is to provide an independent perspective with sufficient credibility to prompt them to amend their prior views. We see the role of the veil of ignorance in process tracing in similar terms, with the principal investigator in the position of the commissioner. The veiling method generates an independent set of causal claim data that can be used to re-evaluate what the PI previously thought and is proposing and then can potentially be used by others to replicate the process tracing.

We note that the field of psychology has been going through a major crisis over the last decade. A great deal of this has to do with issues of replication and transparency. If one reads a critique such as Chambers (2017), which traces the crisis, one cannot help but be struck by its direct relevance to qualitative methods and process tracing. For example, can one replicate a given process tracing analysis? Currently we think that this is almost impossible. Does the author give all the sources actually consulted or just those used?

Clearly, we think that researcher bias is a significant problem in qualitative research, not least because uncertainty over its magnitude undermines perceptions of reliability. In all the sciences, including the natural sciences, researchers are psychologically attached to their favored theories and ideas and quite naturally seek to defend them. We all know of stories about how prominent authors seek to suppress publication of highly critical work. It is extremely difficult to be objective and evaluate and

choose evidence without being subject to bias about your own preferred theory. The physicist Niels Bohr went so far as to say that physics progresses funeral by funeral. There may well be other ways to address the problem of bias, but we stand by our claim that it is an issue and we should explore systematic ways to address it beyond the standard ones advanced by the respondents.

Practical Issues

As the Axelrod (1976) example suggests, we think “coding” causal claims is doable by research assistants who are not experts in the area. The reason is exactly that they are not expected to evaluate the quality or validity of the causal claims but to describe them accurately. This is exactly what we expect, for example, of Ricks and Liu (2018) in their description of someone else’s causal mechanism explanation of the Japanese developmental state—that it is accurately described.

The evaluation of causal claims is definitely not the job of the research assistant, but of the principal investigator who is an expert in the area. Going back to Goertz’s example, he was not trying to evaluate the substance of various claims about World War I, but to understand how they were expressed and explore and evaluate them on methodological grounds.

A number of the contributors raise important practical questions. Is it doable? Is it affordable? When considering research design, it often makes sense to start with an optimal design and *then* make necessary adjustments on pragmatic grounds. Our proposal is precisely a call for process tracing researchers to explore the practical possibilities of applying a veil of ignorance approach. We are also emboldened in doing so by the experience of “blind-folding” data collection—and of separating it from the task of coding causal claims in text thereby obtained—in more than three dozen impact evaluations using the QuIP, as already discussed. The QuIP experience demonstrates that there is scope for enhancing procedural transparency and credibility of research by thinking more carefully both about who does what within process tracing and the choreography of who knows what, and when (Copestake et al. 2018).

A separate question is whether it is affordable. For impact evaluation, dominated by experimental and quasi-experimental designs often costing hundreds of thousands of dollars, the QuIP approach is relatively inexpensive. What we propose can also be meaningfully compared with the effort and expense to develop a quantitative data set. It has become commonplace to recruit graduate students to help build datasets that typically involve a signif-

icant expenditure not only of time but of resources as well.

We contend that this particular objection about practicality reflects a broader bias. While it is not considered out of the ordinary to spend six or seven figures on large scale, data-oriented quantitative projects, it is assumed that high-quality qualitative work can, and even should, be turned out by lone researchers on shoe-string budgets (and self-exploitation). Why? The question of expense should be considered with reference to designs that most effectively and transparently get the job done. The bar with respect to what constitutes an appropriate expense should be the same for both quantitative and qualitative work and not subject to implicit or explicit double standards.

Normative Issues

We were particularly surprised, but ultimately engaged by, the normative objections to the approach. Those criticisms revolve around questions of hierarchy, power, and fairness that may indeed arise when using the veil of ignorance approach. But we would argue these problems are not particular to the VPT but arise equally if not more prominently in large swathes of how social science research is currently structured.

Liu (this issue) offers a succinct statement: “It perpetuates a hierarchy. The research assistants who do the hard work—arguably the brunt of the work—are considered merely as assistants and coders and never as equals (i.e., researcher #2 or co-researcher). The inaccessibility of the method to those outside of elite institutions with large resources exacerbates another hierarchy.” MacLean echoes similar concerns.

It is useful to reiterate briefly the well-known logic of Rawls’s (1971) use of the veil of ignorance and why we appropriated the term and corresponding thought experiment. The point of the veil of ignorance was to remove differences of status, income, and power from decisions about the rules that would guide a justly organized society. Similarly, those assessing sources or causal claims would not be biased by a favored explanation—in effect by self-interest—in seeking our sources or opening the door to contending causal claims. MacLean (this issue) is critical of Rawls, and cites the possibility that his liberal egalitarianism “ignores the persistent reality of inequalities based on gender, sex, race, income, etc.” This is a perfectly legitimate debate. But we are not invoking Rawls to defend liberalism; we are invoking the method of putting potentially self-interested decisions at arm’s length.

Amy Liu and Lauren MacLean both raise important

practical and ethical issues. In particular, they home in on the complex collaborative relationships that are becoming more and more prevalent as the discipline moves toward large-scale data-intensive research. Think, for example, of the scale of the Varieties of Democracy (V-Dem) effort. At stake are issues of power and hierarchy, how credit for research is apportioned, whether research meets IRB guidelines, and even—as Liu argues—the distribution of risk between those directing projects and those engaged in fieldwork. We found the discussion of these ethical issues quite trenchant. However, we reject the claims that they are distinctive to the proposed veil of ignorance methodology.

Research can be conducted *in toto* by lone researchers and by teams taking fully equal credit. The VPT methodology would clearly not make sense in these contexts (although the problems of bias do not simply go away). Yet social science research increasingly takes place in the context of lead researchers contracting for the services of others who operate under the direction of PIs (and who may or may not receive appropriate credit for their work). Additional complexities arise in comparative politics research in which lead researchers are employing teams in developing countries where income gaps to the PIs are huge.

Yet it is important to recall that the purpose of the VPT is precisely to tie the hands of lead researchers—to check any tendencies he or she may have to direct research assistants in biased directions. To the extent that such checks are not in place, it would appear to compound—rather than alleviate—the ethical issues raised. Contra MacLean, we see the RA as effectively empowered by this process. They are potentially the carriers of bad news: the researcher who comes back to the principle with the information that other causal possibilities need attention and that favored causal stories are not, in fact, true. Rather than isolating the contracted researcher from the PI, the method forces engagement and even confrontation.

MacLean (this issue) raises a still deeper critique of Rawls that individuals are—as a general proposition—unable to place themselves behind a veil of ignorance. However, our method does not rely on individuals placing *themselves* behind a veil of ignorance. The method is designed to *place* a researcher behind a veil of ignorance with respect to one relatively narrow question: the nature of the favored causal claim held by the researcher.

Whether Liu and MacLean accept our claims, on one point we need to be particularly insistent. Liu and MacLean both suggest that the method we outline raises risks

of the violation of norms of informed consent. Yet we are puzzled why they believe that would be the case. Issues of human subjects do not arise in the use of archival or secondary material and thus only come up using this methodology in the context of open-ended interviews of research subjects. Such studies would need to go through IRB protocols and subjects would of course need to provide informed consent. It is completely unclear to us why such a method would violate current standards.

It is possible that the brevity of our explanation of the veil of ignorance idea was misleading in its portrayal of the role of the PI as directing a research assistant in ways that appeared hierarchal, non-collaborative, and even non-consensual. Yet it is not clear how this particular method—which is in fact designed to check bias on the part of the principal investigator—would generate a relationship that is any more—or less—hierarchical than the relationship any PI would have with a research assistant.

Conclusion

In closing, it is useful to contrast the causal mechanism approach with the Bayesian one, because they sit together less comfortably than Fairfield suggests. A Bayesian approach requires a comparison between competing causal mechanisms. If we take the Liu-Ricks (2018) causal mechanism figure that we use as an example, one needs to contrast that with some alternative(s). As stressed by Fairfield—and foundational to the Bayesian approach—these alternatives must be *mutually exclusive and exhaustive*. Some of the same factors can appear in these alternatives as long as they produce mutually exclusive alternatives to compare.

It is difficult to come up with a causal mechanism figure for an individual case or a small-N phenomenon typical of much qualitative work. The Bayesian approach requires the researcher to construct not only alternative hypotheses, but a complete set of alternatives that covers the probability space. Much work has been spent on how to simplify this potentially challenging injunction, but we just point out here that the task becomes more difficult when causal mechanism figures begin to get anything more than slightly complex. From our perspective, Bayesians overemphasize the need to compare mutually exclusive and exhaustive set of causal mechanisms. For example, it is useful to show that a particular drug does not work as a vaccine for COVID; why do I need a contending hypothesis? In a social science setting it is useful for Trachtenberg (2012) to say that the Fearon audience cost model does not work for the Fashoda crisis without

providing an alternative, or for Haggard and Kaufman (2016) to note that more than half of all democratic transitions during the Third Wave do not comport with the class conflict theory espoused by Acemoglu and Robinson (2006).

We appreciate the care with which our critics raked over this proposal. We have taken away several lessons. One is that significant differences remain, revealed in part by conflicting terminology about what process tracing in its different forms (singular, theory generating, and theory testing) is supposed to produce. In terms of the veil of ignorance, this debate has revealed to us a variety of opinions about how severe “cherry-picking” and confirmation

bias problems in process tracing might be. Our focus is in the production of a causal mechanism figure for the case. We believe this effort benefits from a full assessment of causal claims contained in an appropriately diverse array of sources and a review of evidence for competing causal claims. If nothing else, we hope this symposium will stimulate more investigation of these issues.

Our Veil of ignorance Process Tracing methodology rests on two core principles for good process tracing:

No process tracing without causal mechanisms (expressed in figures).

Process tracing must involve veiled and systematic coverage of sources and their causal claims.

References

- Acemoglu, Daron, and James A. Robinson. 2006. *Economic Origins of Dictatorship and Democracy*. Cambridge: Cambridge University Press.
- Axelrod, Robert, ed. 1976. *Structure of Decision: The Cognitive Maps of Political Elites*. Princeton: Princeton University Press.
- Beach, Derek, and Rasmus Brun Pedersen. 2013. *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.
- Chambers, Chris. 2017. *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton: Princeton University Press.
- Copestake, James, Claire Allan, Wilm van Bekkum, Moges Belay, Tefere Goshu, Peter Mvula, Fiona Remnant, Erin Thomas, and Zenawi Zerahun. 2018. “Managing Relationships in Qualitative Impact Evaluation of International Development: QuIP Choreography as a Case Study.” *Evaluation* 24, no. 2 (April):169–84.
- Copestake, James, Gabby Davies, and Fiona Remnant. 2019. “Generating Credible Evidence of Social Impact Using the Qualitative Impact Protocol (QuIP): The Challenge of positionality in Data Coding and Analysis.” In *Myths, Methods and Messiness: Insights for Qualitative Research Analysis*, edited by Bryan Clift, Julie Gore, Sheree Bekker, Ioannis Costas Battle, Katharina Chudzikowski, and Jennifer Hatchards, 17-29. Bath: University of Bath.
- Copestake, James, Marlies Morsink, and Fiona Remnant, eds. 2019. *Attributing Development Impact: The Qualitative Impact Protocol Case Book*. Rugby: Practical Action Publishing.
- Fairfield, Tasha, and Andrew Charman. 2017. “Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats.” *Political Analysis* 25, no. 3 (July): 363–80.
- Fairfield, Tasha, and Andrew Charman. Forthcoming. *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*. Cambridge: Cambridge University Press.
- Goertz, Gary, and Jack Levy, eds. 2007. *Explaining War and Peace: Case Studies and Necessary Condition Counterfactuals*. London: Routledge.
- Haggard, Stephan, and Robert R. Kaufman. 2016. *Dictators and Democrats: Masses, Elites, and Regime Change*. Princeton: Princeton University Press.
- Paluck, Elizabeth Levy. 2010. “The Promising Integration of Qualitative Methods and Field Experiments.” *The Annals of the American Academy* 28, no. 1 (March): 59–71.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.
- Ricks, Jacob I., and Amy H. Liu. 2018. “Process-Tracing Research Designs: A Practical Guide.” *PS: Political Science & Politics* 15, no. 4 (October): 842–46.
- Shiffrinson, Joshua, and Stephan Haggard. 2020. “Power Shifts: Connecting IR Theory to the Chinese Case.” *Journal of East Asian Studies* 20, no. 2 (August): 131–86.
- Stern, Elliot, Nicoletta Stame, John Maybe, Kim Forss, Rick Davies, and Barbara Befani. 2012. “Broadening the Range of Designs and Methods for Impact Evaluations.” DFID Working Paper 38, Department for International Development, London, UK. <https://www.oecd.org/derec/50399683.pdf>.

Introducing a New Web Resource for Digital Fieldwork

Qualitative and Multi-Method Research

Fall 2020, Vol. 18, No. 2

[https://DOI: 10.5281/zenodo.4428970](https://doi.org/10.5281/zenodo.4428970)

Built by researchers for researchers, the Digital Fieldwork website (www.digitalfieldwork.org) (www.digitalfieldwork.org), which will formally launch in mid-February 2021, offers a collaborative forum where scholars with limited ability to conduct traditional fieldwork can help each other to identify and capitalize on data-gathering opportunities, and to explore and address the data-gathering challenges involved in conducting digital fieldwork.

We invite others to join us in this endeavor. Please visit the site and consider contributing material. The site offers three types of content:

- *Resources*: information about events and webinars, organizations and websites, trainings and workshops, as well as collections of digital archival and news media resources related to digital fieldwork.
- *References*: citations for books, articles, and guides on digital fieldwork.
- *Reflections*: original, personal, and experiential thought pieces developed in a variety of media (photo essays, poems, artistic representation, audio, videos, prose, poems, etc.) that discuss how to carry out, manage, or address challenges in conducting digital fieldwork or offer reactions to the experience of doing so.

Traditional field research requires being physically present. Invaluable insights are gained and complex dynamics are understood through the face-to-face interviews, everyday interactions, serendipitous meetings, and discovery of new information sources that fieldwork makes possible.

The COVID-19 pandemic profoundly disrupted how we do fieldwork. Yet it is also the case that fieldwork has been, and will continue to be, out of reach for many scholars around the world for reasons that have little to do with the global health crisis that emerged in 2020. Scholars with limited research budgets often struggle to cover the expenses associated with fieldwork. Additional responsibilities and commitments such as teaching, administrative duties, children, and elder care also restrict scholars' ability to spend time in the field.

In the context of these new and long-standing challenges, technological advances and increased global engagement with technology encourage and facilitate new ways of thinking about and conducting field research. Broadening our conceptualization of fieldwork to include digital techniques can democratize field research, allowing more scholars to imagine conducting fieldwork. Further, accessing digital historical archives, interacting with research interlocutors using digital platforms, and drawing on digital media can catalyze the development of new techniques for data generation and research interaction.

Of course, as we develop new techniques for gathering data, new intellectual, ethical, and practical challenges emerge. How can we protect research subjects and researchers when new tools are used in the contexts and digital spaces in which we work? Whose voices are silenced—and whose are amplified—when fieldwork is conducted digitally? How does conducting human participant research digitally affect the process of gaining approval from Institutional Review Boards (IRBs)? Which technologies are safest? How can our on-the-ground experience with data generation inform our use of digital tools and techniques and help us overcome barriers to employing them? How can researchers determine when it is safe, ethical, and effective to resume on-the-ground fieldwork?

To date, little training has emerged to help scholars wrestle with these questions. We hope this new website will encourage dialogue about the nature of research in a digital age, and about how to accomplish our intellectual goals as ways of thinking about and accessing research sites, participants, and materials evolve.

We invite you to join us. Please visit the website (www.digitalfieldwork.org) – formally launching in mid-February 2021 – and become a contributor. Please also feel free to email Diana Kapiszewski (dk784@georgetown.edu), Lauren MacLean (maclelanl@indiana.edu), or Lahra Smith (Lahra.Smith@georgetown.edu) with comments, queries, or content.

Longform APSA Awards (2020)

Qualitative and Multi-Method Research

Fall 2020, Vol. 18, No. 2

[https://DOI: 10.5281/zenodo.4046696](https://doi.org/10.5281/zenodo.4046696)

Giovanni Sartori Award for Best Book on or Using Qualitative Methods

This award recognizes the best book, published in the calendar year prior to the year in which the award is presented, which makes an original contribution to qualitative or multi-method methodology per se, synthesizes or integrates methodological ideas in a way that is itself a methodological contribution, or provides an exemplary application of qualitative methods to a substantive issue. The selection committee consisted of Abigail Williamson (Trinity College), chair; Simeon Nichter (University of California San Diego); and Jim Goldgeier (American University).

Winner of the 2020 Award: Jennifer Bussell. 2019. *Clients and Constituents: Political Responsiveness in Patronage Democracies*. Oxford: Oxford University Press.

Prize Citation: This beautifully written book reveals the oft-overlooked significance of constituency services in patronage democracies. By shadowing a range of public officials, coupled with interviews, surveys, field experiments, and secondary data analysis, Bussell reveals that constituency service is crucial for redistribution and representation in India. Further, she convincingly substantiates a surprising argument: Often constituent services are not contingent upon party affiliation, although they are related to broader systems of clientelism. Higher-level politicians provide constituency services to secure electoral support from citizens blocked from accessing benefits due to local-level clientelism. The book's stellar use of shadowing provides deep insights into the logic of its argument, revealing how officials cater directly to everyday citizens without regards to partisanship. This meticulous qualitative work provides tremendous inductive leverage, which is then tested using both experimental and observational data. Throughout, the book marshals impressive quantitative and qualitative analyses to elaborate each part of its nuanced, compelling argument. The book's closing comparative analyses suggest that its findings have implications for multiple continents.

Honorable Mention: Gwyneth H. McClendon and Rachel Beatty Riedl. 2019. *From Pews to Politics: Religious*

Sermons and Political Participation in Africa. Cambridge: Cambridge University Press.

This methodological tour-de-force examines a rarely-addressed question: Do religious sermons affect political behavior? As the authors remind us, religious sermons are among the most ubiquitous forms of elite communication across the globe. Yet it is challenging to sort out whether the content of messages, as opposed to other elements of religious participation, shape civic engagement. By pairing observation of churches in four African cities with focus groups, lab experiments, analyses of nationally representative surveys, and historical case studies, the authors convincingly argue that differing sermons in sub-Saharan Africa shape political behavior. Specifically, Pentecostal and Mainline churches' distinct messages about the individual or structural source of the world's problems and the possibility of making change were reflected in the form and intensity of congregants' political participation. The authors' mixed-methods approach serves as a model to all researchers. Following meticulous hand-coding of sermons and congregational observation, lab experiments masterfully tease apart causal effects. The underlying mechanisms, as well as the duration of effects, are then insightfully explored using focus groups, interviews, and surveys that rely on the random incidents of the day of data collection to illustrate how the effects of Sunday sermons decay over time. The closing historical case studies persuasively argue that these innovative findings travel more broadly beyond the African context.

Alexander George Award for Best Article or Book Chapter on or Using Qualitative Methods

This award recognizes the journal article or book chapter, published in the calendar year prior to the year in which the award is presented, which—on its own—makes the greatest methodological contribution to qualitative research and/or provides the most exemplary application of qualitative research methods. The selection committee consisted of Janet Lewis (George Washington University), chair; Jessica Rich (Marquette University);

and Sean Yom (Temple University).

Winners of the 2020 Award: Ezequiel González-Ocantos, and Jody LaPorte. 2019. “Process Tracing and the Problem of Missing Data” *Sociological Methods & Research*: 1-29. DOI: <https://doi.org/10.1177/0049124119826153>

Prize Citation: This article makes explicit a problem that is present, if often not explicit, in the work of process tracers: How to handle missing data—that is, steps in a causal process for which evidence is simply not available? As González-Ocantos and LaPorte point out, this is an especially big problem in the study of non-democratic regimes and violence, among other contexts. The answer that González-Ocantos and LaPorte propose is clear yet subtle—and it is simple, yet not simplistic. They propose three approaches. First, they propose that scholars consider whether the missingness is consistent with the theory’s microfoundations, especially the strategic incentives of key actors. The missingness can, therefore, *help* to confirm or disconfirm the theory. Second, they propose scholars consider out-of-sample evidence for the mechanisms; to assess the mechanisms in another case for which data on the proposed mechanism *is* available. Their third suggested approach is to specify the analytic status of each link in the causal chain, and to consider carefully whether missing data on one link calls into question the broader causal process.

In the committee’s conversations about this article, words that came up repeatedly to describe it were “clear,” “useful,” and “helpful.” We agreed it would go straight to our graduate methods syllabi and would have an enduring impact on how scholars approach missing data in process tracing. Congratulations to the authors on winning the 2020 George Award!

Honorable Mention: Rana B. Khoury. 2020. “Hard-to-Survey Populations and Respondent-Driven Sampling: Expanding the Political Science Toolbox.” *Perspectives on Politics* 18, no. 2 (June): 509-526. (published online December, 2019) <https://doi.org/10.1017/S1537592719003864>

This ambitious article takes on a challenging and important problem: how can researchers access representative samples of hard-to-reach populations, from undocumented migrants, to extremists? These populations’ beliefs and behaviors are at the heart of core theories of conflict, clientelism, and authoritarianism, yet as Khoury notes, perfectly representative sampling of them is often not an option—not because researchers do not try hard enough, but because many people simply do not want to be found.

Khoury proposes that we consider the technique of

Respondent-Driven Sampling (RDS), which is a tool used more commonly in epidemiology to reach populations at high risk of HIV. Khoury demonstrates RDS’s promise with a study of Syrian refugee activists in Jordan. In doing so, she argues persuasively and boldly that political scientists should seriously consider adding RDS to their methodological toolbox. Congratulations!

Kendra Koivu Award for Best 2019 APSA Paper on or Using Qualitative Methods

Recipients: Megan Becker, Jonathan Markowitz, Srividya Dasaraju, Lindsay Lauder, Isabelle Nazha, and Sarah Orsborn. “Replicating the Resource Curse: Ross (2004) and Prospects for Qualitative Replication.”

Committee: (University of Arizona), chair; Benjamin Read (University of California Santa Cruz); and Wendy Pearlman (Northwestern University).

Prize Citation: For this year’s Qualitative and Mixed Methods Research Group’s Kendra Koivu Best Paper Award, the committee unanimously selected Becker et. al’s “Replicating the Resource Curse: Ross (2004) and Prospects for Qualitative Replication” as the winner. This paper was a clear standout both for its methodological and theoretical ambition. Given its methodological rigor and inventiveness, the committee also felt the paper was a fitting tribute to Kendra Koivu’s legacy.

Methodologically, the paper raises terrific questions and suggestions regarding what it means to do careful, reliable, systematic qualitative evaluation of evidence. Starting from the understanding that political science faces a replication crisis, this paper is a potentially groundbreaking effort to address these concerns in the field of qualitative methods. In replicating Michael Ross’s influential qualitative examination of the resource curse arguments, the paper demonstrates how use of an explicit codebook and multiple coders enable researchers to transparently replicate existing qualitative studies. Far from nit-picking, their efforts get to the heart of the important problem of using written source material to identify causal mechanisms and evaluate hypotheses concerning those mechanisms—that is, the very essence of answering research questions using qualitative data.

Theoretically, the reexamination of the data also potentially leads to some important correctives to the existing theoretical account. It finds that contrary to Ross, greed and grievance play some role in non-separatist civil war onset and duration. They also find that resources do not shorten conflict in any cases, which differs from Ross’s finding that resources can shorten conflict in some cases. Given the centrality of resource curse arguments to international relations and comparative politics, these

will surely be important findings that scholars in both fields will have to reckon with.

We want to congratulate the authors of this year's winning paper and strongly encourage all members of the group to read it. We wish them all the best in their efforts to publish this fine work of social science!

David Collier Mid-Career Achievement Award

This award honors the important contributions of David Collier to the discipline through his research, graduate teaching, and institution-building and, more generally, as a founder of the qualitative and multimethod research movement in contemporary political science. The award is presented annually to a mid-career political scientist to recognize distinction in methodological publications, innovative application of qualitative and multi-method approaches in substantive research, and/or institutional contributions to this area of methodology. The selection committee consisted of Jason Seawright (Northwestern University), chair; Andrew Bennett (Georgetown University); Diana Kapiszewski (Georgetown University); and Colin Elman (Syracuse University).

Winner of the 2020 Award: Jennifer Cyr, University of Arizona

Prize Citation: Jennifer Cyr has made substantial and distinctive contributions to the field of qualitative and multimethod research in political science, with two areas of special emphasis. She has been an institution-builder and a leader within the QMMR community, broadly construed, and she has developed a substantial research trajectory that has reinvigorated discussion and application of focus groups as a qualitative method in political science. These contributions have already opened new professional opportunities and research avenues for a wide range of political scientists, and they give every impression that Jennifer is just getting started.

First, let us consider Jennifer's impact with respect to focus groups in qualitative political science. Before her research in this area began, focus groups had a healthy methodological life in sociology, market research, and some other disciplines. Within political science, however, the method had taken on a seemingly stable niche role as a largely unreported and under-analyzed tool for survey and experimental pretests. Jennifer's work has changed this landscape substantially, starting with her 2016 article, "The Pitfalls and Promise of Focus Groups as a Data Collection Method," which has already become the most cited work by a political scientist on focus groups in at least a generation. The essay builds a compelling framework in which three different analytic approaches to fo-

cus groups could support divergent goals in a research trajectory, thereby opening the method up to make a broader range of contributions in political science.

Jennifer shows how these contributions play out in practice with her 2017 book, *The Fates of Political Parties: Crisis, Continuity, and Change in Latin America*. One key component of that book's fascinating comparative-historical analysis of 21st-century party trajectories in three crisis-ridden countries—a compelling contribution in its own right to the substantive debates provoked by David and Ruth Collier's work on Latin American parties and institutions—is the argument that some parties, but not others, can draw on substantial ideational resources and more robust party brands to recover from electoral disasters. This concept of ideational resources is the kind that political scientists often struggle to measure, resorting to indirect proxies of various kinds. Jennifer instead shows us how focus groups can much more directly measure ideational resources by revealing the kinds of beliefs about parties that are shared by politically dissimilar members of a given society, and also by demonstrating that, for some parties, there are few if any such beliefs beyond perhaps nostalgia for the past. Jennifer's book brings her methodological work to life by showing in a compelling way how focus groups can be introduced to the heart of applied qualitative research in political science.

Her work on focus groups continues to build, and alongside other articles, her 2019 book, *Focus Groups for the Social Science Researcher*, provides a pragmatic guide to the method for beginners that is likely to substantially amplify the importance of this technique in our discipline. It is not a commonplace event that a mid-career researcher is as powerfully associated with a method as Jennifer is with focus groups, and her intellectual creativity and entrepreneurship in this domain would justify the Collier award in itself.

Of course, in Jennifer's case, this work does not stand alone as a basis for the award. She is also an institutional innovator and leader within the QMMR universe. Jennifer is one of the founding organizers of the Southwest Workshop on Mixed Methods Research, a recurring conference that has offered a high-value intellectual forum for the development of new ideas in the QMMR space for several years. She has taught a series of short courses on focus groups at APSA, at IQMR, and in a variety of Latin American venues. Finally, she has served first as co-editor and then as sole editor of the QMMR publication in recent years, keeping that central space active and exciting through a series of hardships at every scale from the global to the personal. In short, Jennifer is the kind of scholar for which the Collier award was invented.

