QMMR

Qualitative and Multi-Method Research

Fall 2021 - Spring 2022 | Volume 19.2/20.1

Inside:

ISSN: 2153-6767

QMMR Qualitative & Multi-Method Research

Fall 2021 - Spring 2022 | Volume 19.2/20.1

Table of Contents

Symposium: On Integrating Surveys, Experiments, and Qualitative Work

Symposium: Author-Meets-Critics: James Mahoney, 2021. *The Logic of Social Science*. Princeton, NJ: Princeton University Press.

QMMR Qualitative & Multi-Method Research

Letter from the Section President

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448208

The development of the field of qualitative and multi-method research has, over the last two decades or so, been a notable success story in many respects. Intellectual progress has been nothing short of extraordinary, with striking advances in methods such as process tracing, political ethnography, qualitative comparative analysis, and comparative-historical analysis as well as in a range of new approaches to combining methods. Alongside these analytic developments, we have seen substantial innovation in and increasingly sophisticated guidance on effective and ethical strategies of fieldwork. These intellectual leaps forward have been accompanied and facilitated by remarkable institution-building over the last 20 years. Consider, for instance, the founding and scaling-up of the Institute for Qualitative and Multi-Method Research, which has to date trained over 3,000 students and junior faculty; investments in new research infrastructure, such as the Qualitative Data Repository; and, not least, the establishment and growth of this vibrant section, one of the largest in the American Political Science Association (APSA), in addition to the strong commitment of several other APSA sections and groups (e.g., Politics and History, International History and Politics, Interpretive Methodologies and Methods) to the advancement of qualitative scholarship.

In the remainder of this letter, I would like to draw attention to one significant exception to this broad pattern of forward movement: the representation of qualitative research in leading disciplinary journals. In fact, a vast new dataset created by Georgetown's Tranae Hardy, Diana Kapiszewski, and Daniel Solomon (HKS) suggests that qualitative methods have *lost* ground in political science's most visible outlets. As part of their ambitious "Mapping Methods in Contemporary Political Science Research" project, HKS have hand-coded the methods used in a randomly selected quarter of all articles published in 10 top disciplinary and subfield journals[1] over a 20-year period. Though the project is still ongoing, HKS kindly shared some of their data with me; the story those data tell is, to my mind, rather sobering.

According to HKS's data, from 2000 to 2009, 16.8% of single-method articles in the sample primarily employed qualitative methods.[2] That is to say, in the first decade of this century, qualitative methods already represented only a narrow slice of the work published in the discipline's leading journals. From 2010 to 2018, however, things got worse: the percentage of single-method articles using qualitative methods dropped to 9.4%, and the annual absolute number of such articles fell as well.

Even in articles that combine quantitative and qualitative approaches, quantitative methods almost always dominate the mix: only 3.9% of multi-method articles relied largely on qualitative or interpretive methods in the first period, rising just slightly to

[1] *American Journal of Political Science, American Political Science Review, British Journal of Political Science, Journal of Politics, Perspectives on Politics, International Organization, International Studies Quarterly, World Politics, Comparative Politics,* and *Comparative Political Studies.*

[2] For simplicity, I combine here HKS's categories of "qualitative" (falling from 15.3% to 9.1% of single-method articles between the two periods) and "interpretive" (1.5% to 0.3%). HKS use "qualitative" to refer to the "analysis of a small-N number of cases (e.g., countries, parties, laws) in order to draw conclusions about causal relationships" and "interpretive" to refer to methods focused on "disclosing the meaning-making practices and interpretations of human actors located within particular linguistic, historical, and values standpoints, and revealing how those practices configure to generate observable outcomes; allows concepts to emerge from encounters with people and text (i.e., from subjective meanings with no objective truths assumed); sees human action as historically contingent."

5.5% in the second period.

One might qualify the interpretation of these figures in various ways. For one thing, the HKS data do not take into account books, where a good deal of qualitative research is still published. Similarly, there are many other journals beyond HKS's "top ten" that publish far more qualitative political science scholarship. I think it is, nonetheless, striking that, over a period of tremendous intellectual advances and institutional investments in qualitative methodology and training, qualitative research as a *practice* become less prominent in the pages of the most visible and prestigious outlets in our discipline.

I point to this development in the hope of starting—or, really, restarting—a conversation within our community about what is going on, what might be causing it, and what we can do about it. I do not have answers to any of these questions, but I'll point to a couple of additional features of the situation with which I think such a conversation would need to grapple.

For one thing, despite some variation across journals and some progress over time, qualitative work still seems to represent a small proportion of manuscripts *submitted* to top journals. In 2010, the lead editor of the *American Political Science Review* wrote in these pages about "getting qualitative research back into the *APSR*" (Rogowski 2010), appealing to members of our section to submit more work. At the time, Rogowski reported that only 2% of submissions received by his team were designated by authors as qualitative. In response, this section established an annual award to encourage qualitative submissions to the *APSR*. And there *has* been significant forward movement on this front, with qualitative and interpretive submissions rising in the intervening decade. However, their proportion remains modest, hovering between 13% and 14% for the last few years, according to the 2021 *APSR* editors' report (Hayward, Kadera, and Novkov 2021). At the leading subfield journal *Comparative Political Studies,* qualitative work represented only a slightly higher share of submissions—about one-fifth—in 2020 (Ansell and Samuels 2021). At *International Studies Quarterly,* authors of only 15% of submissions in 2019-2020 indicated that their manuscripts made use of case studies (Wiegand and Prins 2021).

There is obviously considerable room for judgment in defining what counts as a "low" share of submissions. As one benchmark, however, consider that in the latest Teaching, Research, and International Policy (TRIP) survey of IR scholars, 60% of faculty respondents reported that they used "qualitative analysis" as their primary methodology (Maliniak 2017).

Alongside modest submission shares, it is notable that those qualitative papers that *are* submitted to top journals appear (from the scant data available from just a few journals) to perform less well, on average, in journal review processes than do papers using other methods. At the *APSR,* qualitative manuscripts have been about half to two-thirds as prevalent among acceptances as they are among submissions.[3] *CPS,* meanwhile, accepted a mere 5 of the 176 qualitative papers submitted in 2020.

One could imagine many possible reasons for low submission rates and lower-than-average acceptance rates, and I greatly look forward to HKS's own in-depth analysis as the "Mapping Methods" project moves forward. I also hope that we as a community can have a broader conversation about the implications of, causes of, and possible responses to these developments.

I would also like to emphasize that I raise these issues *not* out of a sense of crisis in our field. To the contrary, I do so out of a keen sense of the exceptional *strengths* that contemporary qualitative approaches bring to the production of social knowledge: to the measurement and description of social phenomena, to accounts of how actors make sense of political life, to causal explanation and inference, and to theory-development. And when policymakers and the public seek to comprehend and respond to social upheavals and dilemmas around the world—as I write, my mind turns naturally to the Russian invasion of Ukraine—there is no substitute for the depth of contextual, case-level understanding that qualitative research brings to the table. In short, I believe that there would be large intellectual and social payoffs to getting more qualitative scholarship into the pages of our discipline's preeminent outlets.

I invite section members to get in touch with me (alan.jacobs@ubc.ca) over the next few months to let me know your thoughts about how we might work toward this goal. We will then devote time at the section's business meeting in Montreal to discuss possible steps that the section might take to encourage or facilitate the publication of more qualitative research in leading journals.

Before closing, I would like to thank the many colleagues who have committed their time and talents to the important work of this section, over the last year and going forward. Thank you to Jason Seawright for his excellent leadership as section president over the last two years. It is an honor to be following on from Jay in this role. I am also absolutely delighted to be working with Veronica Herrera as section Vice President and Tasha Fairfield as Secretary-

[3] Qualitative articles represented 7% of all acceptances in the Mannheim editorial team's last two years and 10.4% during the first 10 months of the current team's tenure.

Treasurer, and am grateful to Chloe Thurston, our 2022 Division Chair, for all of the work that she has put into crafting a fantastic program for the Annual Meeting in Montreal. Heartfelt thanks as well to the at-large members of the Executive Committee; the nominating committee; and the members of our book, article, paper, and mid-career award committees for the investments they are making in the section's institutional infrastructure and intellectual life.

I would like to extend special thanks to Jennifer Cyr, the amazing editor of *QMMR*, who with this double issue is closing out her five-year editorial tenure. Under Jen's editorship—initially a co-editorship with Kendra Koivu, who passed away in 2019—*QMMR* has flourished as a site of vibrant methodological debate and as an indispensable outlet for the advancement of new methodological arguments and ideas. This issue is at the same time the *first* by our new editors, Ezequiel Gonzalez Ocantos and Juan Masullo. Many thanks to Ezequiel and Juan for signing on to this major undertaking. I am very much looking forward to seeing where they take the thriving publication that Jen is handing over to them.

Finally, I would like to express deep gratitude to Colin Elman. As most readers know, Colin was a key force at the section's inception and was, until this past fall, the section's Secretary-Treasurer for all but three years of its existence. During this time, he has been this community's anchor in ways that have gone well beyond the ordinary duties of his section role. Colin has been a constant source of sage advice to the other section officers and the repository of this section's institutional memory. And he has done—and continues to do—so much of the hard work of building and sustaining this community. It is no exaggeration to say that, without Colin's skill, dedication, and tireless efforts, much of the institutional infrastructure that I mentioned earlier in this letter would simply not exist. More than that, Colin's bedrock belief in the importance of qualitative research and of the section's mission has inspired so many others, myself included, to contribute to this community and to continue pushing the field forward. On behalf of all of us: thank you, Colin!

Alan M. Jacobs
University of British Columbia

References

Ansell, Ben and David Samuels, eds. 2021. "*Comparative Political Studies:* 2020 Annual Report." Last updated September 2021. https://drive.google.com/file/d/12cPWffIr-KwpV44nFczR2s6Mu862Le0k/view.

Hayward, Clarissa, Kelly Kadera, and Julie Novkov. 2021. "American Political Science Review Editorial Report: Executive Summary (Spring 2021)." *Political Science Today* 1, no. 3 (August): 46–53. https://doi.org/10.1017/psj.2021.67.

Maliniak, Daniel, Susan Peterson, Ryan Powers, and Michael J. Tierney. 2017. TRIP 2017 Faculty Survey. Teaching, Research, and International Policy Project, Williamsburg, VA: Global Research Institute. https://trip.wm.edu/.

Rogowski, Ronald. 2010. "Getting Qualitative Research Back into the *APSR*." *Qualitative & Multi-Method Research* 8, no. 2 (Fall): 2–3. https://doi.org/10.5281/zenodo.936247.

Wiegand, Krista and Brandon Prins, eds. 2021. "*International Studies Quarterly:* 2021 Annual Editorial Report." Last updated March 15. https://www.isanet.org/Portals/0/Documents/ISQ/ISQ%202021%20Journal%20Report.pdf?ver=2021-04-19-091122-273.

QMMR Qualitative & Multi-Method Research

Letter from the Editor

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448202

This double issue packs a pretty big punch. The content is engrossing. The contributors are diverse. They come from the global South and North. They represent new voices in qualitative and mixed methods as well as the giants in our field.

First, we have a symposium on the integration of qualitative data collection and survey and experimental research. Each contribution emerged from the scholars' personal experience in designing and carrying out a survey, experiment, or survey-experiment. In-depth fieldwork was instrumental for strengthening that design. Here, they tell us how and why this was so. As the symposium's introduction observes, the four articles allow us to "zoom in" on ways that fieldwork and case studies can complement and ultimately strengthen other data analysis methods in precise and innovative ways.

The second article considers novel approaches to storing and securing the sensitive, personal information that scholars acquire while collecting qualitative data. Most researchers who spend time in the field face the inevitable task of safely depositing the personal information of the people with whom they speak. How can researchers be sure that these personal data remain anonymized? The article provides a set of practical tools for managing the potential threat of re-identification.

Finally, we have an author-meets-critics roundtable in written form. This second symposium was inspired by a conversation held at the 2021 Congress of the American Political Science Association about James Mahoney's new book, *The Logic of Social Sciences*—a text that builds off of Mahoney's extensive publications on methodology to promote a new scientific constructivist approach to producing knowledge in the social sciences. The book is groundbreaking, not just for the methods content but for how we do social sciences in general. We invite you to learn about the book from the different perspectives that each contributor brings to the symposium. We think it could be a useful companion piece to reading the book itself—which, of course, you should also absolutely do.

On a more personal note, this is my last issue as editor of QMMR. It has been a long, winding, and thrilling (yes, thrilling) road. I started way back in 2017, when I was pre-tenure and about to give birth to my second child. Since then, I've edited articles and essays on qualitative and mixed methods during: a battle with breast cancer; the death of my friend and co-editor, Kendra Koivu; an international move; a change in institutions; a global pandemic. Throughout each trial and challenge, QMMR has been a refuge. To me, the community of scholars dedicated to qualitative and mixed methods has no parallel. I have learned so much over the past five years. I am grateful to have had the opportunity to contribute to our shared knowledge.

While I am reluctant to move on, I am so pleased to welcome the two new editors of QMMR, Ezequiel González Ocantos and Juan Masullo Jiménez. Ezequiel and Juan are accomplished scholars who have already proven they are up to the editorial task at hand. They've been active at QMMR behind the scenes for over a year. As co-editors of this double issue, their attention to detail and sharp methodological eye were vital for creating a truly excellent final product. I look forward to seeing how they continue to push the publication forward and thank them, heartily, for the work they have done so far.

As I say goodbye, I want to continue to encourage our readers to write about the methods they use. No new idea about methods, I think, is too small or obvious. If you are doing something innovative or different, then tell us about it. In this way our community of scholars grows and evolves for the better.

Jennifer Cyr
Universidad Torcuato di Tella

QMMR Qualitative & Multi-Method Research

Symposium: Varieties of Transparency in Qualitative Research

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448059

On the Potential Complementarities between Qualitative Work, Surveys, and Experiments

Ezequiel González-Ocantos
University of Oxford

Juan Masullo Jiménez
Leiden University

This symposium features four articles that discuss the complementarities between in-depth fieldwork, survey research, and experimental techniques, in particular, survey experiments. They offer fruitful and innovative ways of integrating the different data collection methods, such that the overall research design is strengthened.

For example, the first two articles, one by Rebecca Bell-Martin and the other by Virginia Oliveros, explain how they integrated qualitative and survey methods in two specific research projects, one on civilian responses to violence in Mexico (Bell-Martin) and the other on patronage in Argentina (Oliveros). The authors explain that deep contextual knowledge helped them improve construct and ecological validity, design better survey questionnaires that pay due attention to sensitive issues, and come up with better sampling strategies.

In the third contribution to the symposium, Lucía Tiscornia and Verónica Pérez Betancur draw from their research on policing in Uruguay to make the case that the kind of integration detailed in the first two contributions ought to be made explicit in a Pre-Analysis Plan. They argue this has the potential to enhance both research transparency and replicability, and offer practical advice for scholars interested in following their lead.

Finally, Daniel Encinas writes about a different way in which qualitative methods and survey-experimental research can complement each other. Instead of paying attention to the advantages of integration for, say, the design of more valid and useful questionnaires or treatment vignettes which can enhance a study's internal validity, Encinas focuses on how qualitative "casing" techniques can be used to make claims about the external validity of survey-experimental results.

The symposium builds on influential contributions to mixed-methods research that focus on the integration of qualitative data collection and experimental research (see, e.g., Dunning 2008, Paluck 2010, Thachil 2018, and Seawright 2016). By zooming in on specific areas, such as identifying sensitive issues, external validity, and research transparency, these essays move the conversation forward in useful and instructive ways.

References

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61, no. 2: 282-293. https://doi.org/10.1177/1065912907306470

Paluck, Elizabeth Levy. 2010. "The Promising Integration of Qualitative Methods and Field Experiments." *Annals of the American Academy of Political and Social Science* 628, no. 1: 59-71. https://doi.org/10.1177/0002716209351510

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

Thachil, Tariq. 2018. "Improving Surveys through Ethnography: Insights from India's Urban Periphery." *Studies in Comparative International Development* 53, no. 3 (July): 281–99. https://doi.org/10.1007/s12116-018-9272-3.

Experimentally Testing Ethnographic Insights: Benefits, Challenges, and Considerations

Rebecca Bell-Martin[1]
El Instituto Tecnológico y de Estudios Superiores de Monterrey

Political scientists increasingly pair experimental techniques with some form of qualitative research. The rich contextual knowledge afforded by qualitative approaches offers a number of benefits, including enhancing our ability to identify and validate as-if-random assumptions and interpret the substantive meaning of experimental results (Dunning 2012, 313-37). For this symposium, I consider how pairing survey experiments with one qualitative method in particular, ethnographic field research, can advance research aims. I pay special attention to the benefits and challenges of incorporating ethnographic insights into experimental tests via survey experiments. What analytical leverage do we gain from doing so? What theoretical, empirical, and practical questions should researchers contemplate? Consonant with the mixed methods tradition, my reflection is guided by a consideration of how the strengths of ethnographic work can address the weaknesses of survey experiments, and vice versa.

I first consider how ethnography can inform survey experiments and argue that ethnographic evidence facilitates greater construct and ecological validity of our instruments. I then reflect on the ways survey experiments can advance ethnographic research. I argue that survey experiments help address two limitations of ethnography—generalizability and effects measurement—by enabling out-of-sample theory testing with a systematized measurement instrument. I then highlight an important dilemma researchers face when using survey experiments to test ethnographically-generated theory and discuss how researchers can strike a balance between the pursuit of external validity and ethnographic approaches' inherent embeddedness in local contexts.

To develop my main points, I draw on examples from my research on the political consequences of organized crime violence in Mexico. There, I sought to interrogate an empirical puzzle: In contexts of violent conflict, why are some citizens mobilized to pursue civic engagement as a response to violence while others retreat from civic life out of fear? Existing explanations emphasize the role of direct victimization, but my research reveals an additional empirical relationship not yet fully explored and which I find particularly puzzling: a great deal of civic engagement in violent contexts is carried out by individuals who are not, in fact, victims. If victimization does not explain their choice to undertake civic engagement precisely when violence makes it most risky, what does? To answer this question, I designed a subnational, mixed methods project with a theory-building phase and a theory-testing phase. To inductively theorize the possible mechanisms linking violence and civic engagement, I conducted eight months of participant observation in violent and non-violent neighborhoods in Monterrey, Mexico, and over 150 hours of in-depth interviews with victims and non-victims. To test the theory that emerged, I then designed an original, nationally representative survey experiment within Mexico. My qualitative research directly informed the experimental design.[2]

What Do We Mean by Ethnographic Field Research?

There are myriad definitions of ethnography, including an interpretivist and non-interpretivist school.[3] Irrespective of this variation, long-term embeddedness in a research site through extended fieldwork is the foundation of nearly all ethnographic research. That said, ethnography is not merely a product of time spent in "the field."[4] Indeed, researchers who develop and execute surveys, carry out in-depth interviews, or conduct lab-in-the-field experiments spend a great deal of time in their research site.[5]

As leveraged by political scientists, ethnography has two distinguishing features, its method and the nature of its data. While ethnographic methods include a

[1] I thank Rob Blair, Alejandro Díaz Domínguez, Ezequiel González-Ocantos, Jerome F. Marston, Jr., and Rachel H. Meade for commenting on earlier versions of this essay. All errors are my own.

[2] Additional tests of the theory were carried out using secondary survey data from cases inside and outside Mexico.

[3] See Wedeen (2010) for a discussion of this distinction.

[4] By "the field," I mean the natural environment in which the phenomenon under study takes place. If the research phenomenon takes place in cyberspace, "the field" may mean chatrooms, message boards, social networking sites and other locales of virtual community (Wilkinson 2013, 129).

[5] As Oliveros (this symposium) suggests, field research in general advances experimental work in a number of ways.

range of data collection techniques, including (but not limited to) in-depth interviews, oral histories, and map-making workshops (see Wood 2003 and Brigden 2018 for examples), it is typically grounded in participant observation. During participant observation, the researcher lives and works among the research population for an extended period of time, acting as a member of the group or participating in the group's activities alongside research participants.

The nature of the data collected through this work is distinct. Ethnography is participant-centered, interrogating the phenomenon under study from the perspective of those who experience it. It seeks to uncover new or poorly understood aspects of the research phenomenon by examining how informants make sense of their world and how they construct political meaning around the experiences, power structures, and people around them. These insights are not merely complementary to other "hard" data points. These lived experiences and worldviews directly influence the researcher's understanding of the social and political factors that give life to the phenomenon of interest.

This latter feature lends ethnography a number of analytical strengths. Among them, ethnographic field research grants the researcher unique access and insight into the research phenomenon as it is lived and experienced. This facilitates the discovery of new or poorly understood aspects and processes, like theoretically relevant variables that previous studies overlooked, an unacknowledged mechanism explaining an established association, or a hitherto undocumented empirical relationship. By emphasizing how our interlocutors understand, experience, and process political phenomena, ethnographic research is particularly valuable for exploring political identity formation, the development of political attitudes, and political choices and behaviors—all of which are frequently examined via survey experiments as well. Using survey experiments to further examine ethnographically generated theory and concepts could thus be a logical next step for many political scientists, though doing so has received relatively little academic attention until recently (see especially Thachil 2018).[6] The purpose of the present essay is to advance debate about this particular pairing. I draw on examples of a vignette experiment, but ethnographic insights can inform various elements of experimental design, from the treatment, to question wording, and dependent variable measurement, among others.

Experimentally Testing Ethnographic Insights: The Benefits

Ethnographic field research can help us develop survey instruments and experimental interventions with greater ecological and construct validity (Thachil 2018). Ecological validity refers to how well the conditions of the experimental intervention reflect the real-world environment in which the research population would normally experience the phenomenon. Ethnographic approaches are particularly powerful in this regard for at least two reasons. First, ethnography is explicitly interested in describing research phenomena based on lived experiences from within the social, political, or economic setting under analysis. This means that experimental interventions constructed out of ethnographic evidence are likely to reflect relevant contextual information to which a researcher not engaged in ethnographic work would not be privy. Second and more practically, ethnographic fieldwork physically situates the researcher within the natural environment that is to be mimicked in the experimental design, lending her the contextual knowledge to do so with greater accuracy. This may include insights into question wording, treatment designs, or cultural references that would make an intervention or treatment particularly "real" for respondents.

To that end, researchers may consider drawing on real-world examples observed during their field research that could be duplicated in the design of the survey experiment. To illustrate this, I will first briefly describe the structure of the survey experiment that I designed as part of my research on organized crime violence and civic engagement in Mexico. Its purpose was to test the primary finding from my qualitative research, that exposure to a case of violence provoking one's sense of empathy motivates civic engagement among non-victims, all else equal. The survey featured a vignette describing an incident of violent crime. All respondents read the same vignette, but a first treatment group received pre-vignette instructions meant to prime empathy for the victims. A second treatment group received instructions meant to depress empathy.[7] For purposes of ecological validity (and ethics), it was important that the vignette mimicked the way a majority of Mexican citizens learned about organized crime violence. It was also important that the vignette described an act of violence that, while fictitious, reflected an incident the average citizen might plausibly hear about in their day-to-day life. To achieve this, I drew heavily on my ethnographic observations, incorporating characteristics of real cases I heard about during my in-depth interviews and participant observation. I then scripted the vignette in the style of

[6] Regarding the pairing of ethnographic research and other experimental approaches, see Sherman and Strang (2004) and Paluck (2010).

[7] A control group received instructions that stressed objectivity.

a brief news-style report that mimicked the structure, tone, and vocabulary of real reports from local news outlets. This strategy is analogous to Thachil's (2018) "ethnographic vignette-experiments."

Integrating ethnographic insights into our experimental analyses can also support construct validity (Thachil 2018). Construct validity refers to the degree of equivalence between the theoretical concept under study and our measurement of it. Long-term embeddedness in the research site, along with the participant-generated data that ethnographic methods produce, lend us special insight into (i) how and under what conditions the theoretical construct manifests in the context where the experiment will be carried out; and, (ii) how those who will participate in the experiment understand and interpret the concept. Both improve our ability to operationalize the concept in a way that is in agreement with the theoretical construct to be tested and the research environment. This is not merely a byproduct of the ethnographic method. Many ethnographers take concept-building as a fundamental task. As Fu and Simmons (2021, 1696) note in their reflection on ethnographic accounts of contentious politics, ethnographic work can challenge "taken-for-granted assumptions about the categories that make up the world by trying to see them through the eyes of their interlocutors."

My qualitative findings challenged pre-existing notions of "victimization." Predominant political science accounts of victimhood in Mexico understand victims as those who directly experienced physical violence to their person or a family member. Distinguishing individuals with such experiences from those without them was integral to my research. Yet, as I have recounted elsewhere, these categories did not overlap well with how my interlocutors understood their own victimization status (Bell-Martin and Marston 2021, 171 - 173). For a number of reasons, individuals who directly experience violence may not identify as a victim while others who do not experience violence do identify as such. This insight led me to conceptualize a "spectrum of victimization," in which these different experiences and perspectives can be systematically categorized (Bell-Martin 2019). It also influenced how I operationalized "victimization" in the survey experiment post-treatment questions. My aim was to measure whether or not the respondent had experienced violence directly, not whether they self-identified as a victim. Thus, rather than asking, "Have you been a victim of a crime in the last 12 months?" (a common victimization measurement),[8] the survey asked, "Have you or a family member experienced a violent crime?" The revision was simple but important because it aligned more closely with what was actually of theoretical interest: direct encounters with violent crime, rather than identification with the victimhood concept.[9]

The above points consider how ethnography can inform survey experiments. Reciprocally, survey experiments can bolster ethnographic research, particularly when leveraging survey experiments to test ethnographically-derived theory. Recall that one of ethnography's key strengths is to reveal new processes, variables, relationships, or conditions that are not adequately captured by existing accounts. What is more, the intimate knowledge the researcher gains from immersion in the research context and with the population increases the likelihood she will recognize important nuance or steps in theoretical processes that other approaches would obscure. This lends ethnography great strength in terms of building new theory, particularly about complex social processes. Yet, the generalizability of such findings is typically limited due to the small sample of research participants, the degree of embeddedness in local context, and the nature of data collection. Since ethnography requires long-term, deep engagement with a research population, it tends to be carried out in a highly localized context and with a narrow sample of the population. The issue is also epistemological and methodological. Can ethnographic data and findings ever truly be extracted from the context in which they were generated? Some might further critique ethnographic data collection as unsystematic and intersubjective since it typically engages the researcher in informal conversation, unstructured interviews, and spontaneous social activities that are nearly impossible to replicate and involve multiple, varied interactions between the research participant and the researcher. Given this, how can we demonstrate that our theory travels beyond the narrow research context, sample, and data collection method through which it was generated?

Leveraging a survey experiment to test an ethnographically derived theory can do much to assuage these concerns. Whereas ethnographic field research may limit our sample size and research context, a survey offers a vehicle through which to reach a broader and larger sample without additional, time-intensive participant observation. While ethnographic methods may appear unsystematic, a survey systematizes data collection in a single questionnaire that is administered in a uniform and replicable manner to all research participants and in which variables are measured precisely and consistently. To curtail the effect of intersubjectivity, interaction

[8] See for example the Latin American Public Opinion Project (LAPOP).

[9] For additional examples of this same principle, see Thachil (2018).

between the survey enumerator and the respondent is highly controlled and variation is minimized.[10] Testing the theory via a survey experiment can thus be a powerful way to empirically demonstrate that the new or revised theory is not limited to the specific context nor sample in which it was generated, nor is it merely a product of "intersubjective 'noise'" (Wedeen 2010, 258). On the contrary, demonstrating that one derives similar findings when drawing on an independent and external research sample, and using a second methodological approach provides compelling evidence of the theory's robustness. What is more, whereas ethnographic research is difficult to replicate, a survey is designed for replication, supporting future additional tests of the theory and its generalizability. I return to this topic in the "Challenges" section.

Measuring causal effects is an additional advantage of using survey experiments to test ethnographically derived theory. As Fu and Simmons (2021, 1711-12) note, it is somewhat of a misconception that ethnographic research does not address causal relationships. In fact, much ethnographic work is deeply interested in understanding why certain outcomes occur and their causal mechanisms, an integral part of any causal claim (see, for example, Katz 2001, 2002; Tavory and Timmermans 2013). What is more, the profound knowledge afforded by ethnographic field research prepares scholars to identify and address plausible counterfactuals, an additional and necessary component of causal arguments. Nonetheless, ethnographic approaches cannot approximate the causal identification afforded by experimental techniques. In particular, measuring causal effects is fundamentally outside the scope and capacity of ethnographic research. For this reason, scholars who aim to advance a causal claim about a variable's impact on an outcome based on ethnographic evidence may find it valuable to test that claim via a survey experiment. Whilst the ethnographic evidence can demonstrate the importance of a given variable for an outcome, a survey experiment provides a clear and precise way to demonstrate (i) that said variable has the theorized effect (itself important confirmatory evidence) and (ii) the size of that effect, or *how much* that variable matters. Indeed, McDermott (2002, 341) argues that experiments can bring clarity about a causal relationship obscured by other methods "that allow for less clear causal inference." Ethnography is one such method.

The Challenges

Above, I argued that testing ethnographic insights via survey experiments can alleviate some concerns about the generalizability of ethnographic findings. I return to this point because ethnography and experiments share weaknesses in generalizability (Shadish 1995). Experiments' external validity is a common point of critique and debate among political scientists (see, for example, McDermott 2002, 334–36; Munck and Verkuilen 2005, 389–90; Barabas and Jerit 2010; Krupnikov and Levine 2014; Findley, Kikuta, and Denly 2021). Similarly, ethnographers "are likely not claiming that the relationships identified in one field site will travel unchanged to another. Rather, they may claim that the theoretical and conceptual discoveries of one context can have relevance to another site" (Fu and Simmons 2021, 1711). Indeed, ethnographic researchers typically resist the notion that a theory could be applied to another case or population without first considering how that theory would operate in the specific context. This creates a dilemma for researchers who intend to test a theory generated through ethnographic methods because any tests of that theory should be carried out among a sample of the population external to and independent of the ethnographic sample. How can researchers reconcile ethnography's obligation to context with the need for out-of-sample tests of the theory? What role do survey experiments, themselves limited in terms of external validity, play in that effort?[11]

First, researchers may consider assuming a subnational perspective, looking for appropriate external cases or samples within the same country or city that share similar characteristics to the ethnographic research sample but are independent of it. A subnational approach is valuable in this regard because it controls for macro-level factors that could otherwise shape the research outcome, such as regime type, political institutions or economic system (Snyder 2001; Giraudy, Moncada, and Snyder 2019). Given the highly localized nature of ethnographic research, however, a submunicipal approach that generates the theory among one sample of the city's population but tests the theory via a survey experiment within the same city (or a similar one) allows the researcher to control for relevant factors that vary within a nation, such as local political competition, religious traditions, levels of insecurity, and state capacity. These highly localized subnational approaches are more likely to reflect the most important context-level factors that generated the theory, though they will still be an imperfect approximation. Additionally, the ability to generalize findings to cases further afield, such as other nations, is still restricted.

If the researcher chooses to take the survey experiment beyond the local, they should ask specific

[10] Interviewer effects pose a related, yet different problem. See, for example, Adida et al. (2016).

[11] For a discussion of how qualitative approaches can strengthen the external validity of experiments, see Encinas (this issue).

questions about how the most important variables manifest in the particular research context in which the experiment will be executed in order to maintain the instrument's ecological and construct validity. Based on field research among urban migrants in India, for example, Thachil (2018, 293-94) found that standard measurements of cooperation—contributing money—were not suitable in his research context. Sharing housing with fellow migrants better reflected how cooperation as a theoretical concept operated in his research site, a fact that was integrated into his survey experiment. This information could be gleaned through additional field research. However, if this is not feasible, speaking with local experts and other scholars who conduct research in the area can provide valuable information to inform those choices.

Finally, it is important to remember the role of replication in generating external validity. As noted previously, ethnographic field research is not easily replicable.[12] Survey experiments are replicable, even if the researcher opts to revise the instrument for application in a different context. Multiple replications in different environments and among different populations contributes to a body of evidence (dis)confirming the generalizability of the ethnographically generated theory in a way that is impossible for researchers to achieve through participant observation. McDermott (2002, 335) aptly writes, "external validity is established over time, across a series of experiments that demonstrate similar phenomena using different populations, manipulations, and measures. External validity occurs through replication." As such, researchers testing ethnographically generated theory via survey experiments may wish to incorporate replication studies into their research designs and funding proposals.

Concluding Remarks

A mixed method approach draws its analytical power from leveraging the strengths of one methodological tradition against the weaknesses of another (Seawright 2016). This is not mere triangulation of different types of data, but instead the strategic coupling of analytical techniques so that, together, the evidence they provide offers a more complete explanation of the research phenomenon than either could independently. In this short intervention I engaged existing debates about the profitability of mixed methods research designs that strategically match the strengths of ethnographic field research with the complementary strengths of survey experiments. Building on existing conversations, I described how ethnographic field research can improve the construct and ecological validity of our survey instruments. I then argued that using survey experiments to test ethnographically generated theory can bolster the persuasiveness of our arguments by addressing two weaknesses of ethnographic research, generalizability and effects measurement. The advantages of this pairing are not, however, without disadvantages. In particular, I highlighted dilemmas related to external validity, a weakness both ethnographic field research and survey experiments share. There are certainly more benefits (and challenges) than I highlight here. Rather than a comprehensive account, this short reflection aims to build on existing dialogue and stimulate future debate about how the comparative strengths and shortcomings of ethnographic field research and survey experiments can be strategically matched in the service of analytical purchase. Given the field's continued interest in experimental techniques for causal identification, the question of how ethnographic research and survey experiments can mutually contribute to the field's advancement merits such scholarly attention.

References

Adida, Claire L., Karen Ferree, Daniel N. Posner, and Amanda Lea Robinson. 2016. "Who's Asking? Interviewer Coethnicity Effects in African Survey Data." *Comparative Political Studies* 49, no. 12 (March): 1630-60. https://doi.org/10.1177/0010414016633487.

Barabas, Jason, and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science* Review 104, no. 2 (May): 226–42. https://doi.org/10.1017/S0003055410000092.

Bell-Martin, Rebecca. 2019. "'It Could Have Been Me': Empathy, Civic Engagement, and Violence in Mexico." PhD diss., Brown University.

Bell-Martin, Rebecca, and Jerome F. Marston, Jr. 2021. "Confronting Selection Bias: The Normative and Empirical Risks of Data Collection in Violent Contexts." *Geopolitics* 26, no.1 (September): 159-92. https://doi.org/10.1080/14650045.2019.1659780.

Brigden, Noelle K. 2018. *The Migrant Passage*. Ithaca: Cornell University Press.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. New York: Cambridge University Press.

[12] However, see the argument in Pérez Betancur and Tiscornia (this issue) for including qualitative elements of mixed method experimental studies in pre-analysis plans for replicability purposes.

Findley, Michael G., Kyosuke Kikuta, and Michael Denly. 2021. "External Validity." *Annual Review of Political Science* 24 (May): 365– 93. https://doi.org/10.1146/annurev-polisci-041719-102556.

Fu, Diana, and Erica S. Simmons. 2021. "Ethnographic Approaches to Contentious Politics: The What, How and Why." *Comparative Political Studies* 54, no. 10 (September): 1695–721. https://doi.org/10.1177/00104140211025544.

Giraudy, Agustina, Eduardo Moncada, and Richard Snyder, eds.. 2019. *Inside Countries: Subnational Research in Comparative Politics*. New York: Cambridge University Press.

Katz, Jack. 2002. "From How to Why: On Luminous Description and Causal Inference in Ethnography (Part 2)." *Ethnography* 3, no. 1 (March): 63-90. https://doi.org/10.1177/1466138102003001003.

———. 2001. "From How to Why: On Luminous Description and Causal Inference in Ethnography (Part I)." *Ethnography* 2, no. 4 (December): 443-73. https://doi.org/10.1177/146613801002004001.

Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1, no. 1 (August): 59–80. https://doi.org/10.1017/xps.2014.7.

McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10, no. 4 (Autumn): 325–42. https://doi.org/10.1093/pan/10.4.325.

Munck, Gerardo L., and Jay Verkuilen. 2005. "Research Designs." In *Encyclopedia of Social Measurement,* Vol. 3., edited by Kimberly Kempf-Leonard. 385–95. New York: Elsevier.

Paluck, Elizabeth Levy. 2010. "The Promising Integration of Qualitative Methods and Field Experiments." *The Annals of the American Academy of Political and Social Science* 628, no. 1 (March): 59–71. https://doi.org/10.1177/0002716209351510.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

Shadish, William R. 1995. "The Logic of Generalization: Five Principles Common to Experiments and Ethnographies." *American Journal of Community Psychology* 23, no. 1 (June): 419–28. https://doi.org/10.1007/BF02506951.

Sherman, Lawrence W., and Heather Strang. 2004. "Experimental Ethnography: The Marriage of Qualitative and Quantitative Research." *The ANNALS of the American Academy of Political and Social Science* 595, no. 1 (September): 204–22. https://doi.org/10.1177/0002716204267481.

Snyder, Richard. 2001. "Scaling Down: The Subnational Comparative Method." Studies in Comparative International Development 36, no. 1 (March): 93–110. https://doi.org/10.1007/BF02687586.

Tavory, Iddo, and Stefan Timmermans. 2013. "A Pragmatist Approach to Causality in Ethnography." *American Journal of Sociology* 119, no. 3 (November): 682-714. https://doi.org/10.1086/675891.

Thachil, Tariq. 2018. "Improving Surveys through Ethnography: Insights from India's Urban Periphery." *Studies in Comparative International Development* 53 , no. 3 (July): 281–99. https://doi.org/10.1007/s12116-018-9272-3.

Wedeen, Lisa. 2010. "Reflections on Ethnographic Work in Political Science." *Annual Review of Political Science* 13, no. 1 (June): 255–72. https://doi.org/10.1146/annurev.polisci.11.052706.123951.

Wilkinson, Cai. 2013. "Ethnographic Methods." In *Critical Approaches to Security: An Introduction to Theories and Methods,* edited by Laura J. Shepherd, 129-45. New York: Routledge.

Wood, Elisabeth. J. 2003. Insurgent Collective Action and Civil War in El Salvador. Cambridge: Cambridge University Press.

Reflections on the Importance of Fieldwork for Survey Experiments on Sensitive Topics

Virginia Oliveros
Tulane University

Fieldwork is critical for the successful implementation of field and survey experiments. Good contextual knowledge is key for any sound empirical study, but even more so in the case of experiments because these are design-based research strategies (Dunning 2012)—most of the work and important decisions need to be done before the implementation phase. Once the experiment is conducted, there is little room to fix mistakes or bad choices. Thorough preliminary fieldwork is therefore critical. In my contribution to the symposium, I focus on one particular type of survey experiment, the list experiment—a technique developed to study sensitive topics. I begin by describing my research on patronage in Argentina, which relied heavily on a series of list experiments. I then discuss three key aspects of this research for which deep knowledge of the case, as well as extensive preliminary fieldwork and being in the field while the pilot and the survey were being conducted, were key.

Public Sector Jobs and Political Services: The Machine at Work

In my book, *Patronage at Work: Public Jobs and Political Services in Argentina* (Oliveros 2021), I study the exchange of public sector jobs for political support, or patronage. Even though patronage is a widespread phenomenon, the difficulty in collecting systematic data about it means that we know very little about how patronage works. The book provides a comprehensive description of what patronage employees in low and mid-level positions do in exchange for their jobs, as well as a novel explanation of why they do it. *Patronage at Work* thus aims to understand the specific mechanisms behind the electoral returns to patronage politics.

While patronage is often perfectly legal, it is particularly difficult to study because it constitutes a "gray area" of acceptable practice (Van de Walle 2007, 52). To measure the types and extent of the political services that employees hired through patronage contracts provide to their patrons, I take an approach that allows me to elicit accurate information from public sector employees by minimizing social desirability bias. I use an original face-to-face survey of 1,200 low and mid-level public employees in three Argentinean municipalities (Salta, Santa Fe, and Tigre) that incorporates two strategies to elicit honest responses. The first, following Scacco (2010), consists of employing a number of techniques to earn respondents' trust by guaranteeing the confidentiality of the most sensitive questions. The second is the use of list experiments, a survey technique that protects the privacy of responses by using indirect questioning. Another research tool, the vignette experiment, allows me to assess why public sector employees comply with their side of the patronage agreement. I also conducted multiple interviews. Some of them were part of my preliminary fieldwork; others were conducted later on to illustrate and provide a thicker description of the main findings.

The Sample

As in many other democracies of the Global South, information on public employment in Argentina is not publicly available, and politicians and bureaucrats are reluctant to share it. The first challenge of the research project was therefore to get access to public employment data in order to be able to draw a representative sample for the survey. Preliminary fieldwork was key to achieving this. To get access to this data, I used personal connections to reach several local political authorities and then met with high-level public officials and politicians to explain the purpose of the study, gain their trust, and eventually obtain lists of public employees and receive authorization to conduct the survey. Because I am Argentinean and went to college there, I had some contacts (both academic and political) that proved a good starting point. But even with this "home" advantage, obtaining public employment data in all three municipalities was still daunting and time consuming.

For example, my initial trip to Salta was unsuccessful. My contact in the administration avoided me for a week, stopped replying to my emails, and scheduled in person or telephone appointments at times when he knew he would not be at the office. He eventually informed me that his bosses had requested that I waited until after the upcoming local election to conduct the survey. This meant withholding data for six months. In Tigre, I similarly struggled to get access to the data and the permission to conduct the survey. My contact at the municipality warned me initially that while he could probably guarantee an interview with a gatekeeper (someone close to the mayor), he was skeptical that they would share the data because it was too sensitive. Indeed, as this contact anticipated, the director of personnel proved very reluctant to assist me with the study and waited until he had written authorization from the mayor to release the data—put differently, he refused to share the data just with a phone call from a high-level official close to the mayor. Finally, in Santa Fe, there were several failed attempts to get an appointment to discuss my project with the relevant officials. A phone call from a former federal congressperson from the mayor's party facilitated access.[1]

In the end, local authorities in all three municipalities met with me, read the survey instrument, and authorized me to access the data and conduct the survey. Knowing what was sensitive in those questionnaires was key to be able to pass this barrier. That knowledge came from my familiarity with the case and the sensitivity of the issues in the Argentine context. For example, to maximize the chances of getting official approval for the survey, I described the survey to local authorities in broad terms as concerning the relationship of public sector employees with local public life (*la relación de los empleados públicos con la vida pública local*). "Local public life" included politics but also other aspects like participating in community meetings and projects, as well as volunteering. My main interest was, of course, politics, but this broader description sounded less "threatening" to the authorities whose main fear seemed to be that I might find some irregularities in public sector appointments (nepotism or too many partisan affiliates) and share that information with journalists. I also took two other precautions. First, I excluded particularly direct, sensitive questions—especially ones related to the mayor.[2] Second, I designed the survey instrument to be as short as possible to ensure employees would not be kept away from their jobs for long periods of time.

Strategies to Ask Questions on Sensitive Issues

Preliminary fieldwork and good knowledge of the case were also important to find ways to deal with the sensitive topics that did make it to the instrument.[3] In order to conduct the survey, enumerators received a random sample of names of public employees and their work addresses, and directly approached them at their workplaces during work hours. Since the focus was on mid- and low-level positions in the administration, places of work ranged from the city hall and decentralized offices, to cemeteries, construction sites, health centers, parks, and the street. Because the survey was conducted face-to-face at this broad array of locations, getting truthful answers presented a challenge. While high-ranking public officials often have private offices, most public employees in Argentina share their workspaces. The issue was that public employees could be unwilling to reveal sensitive information in front of others. How to obtain truthful answers under these conditions? I implemented two distinct but complementary strategies to elicit honest responses and thus minimize social desirability bias.

First, I designed a series of list experiments—a technique that protects the privacy of responses by using indirect questioning (more on this below). Second, I followed Scacco's (2010) strategy (originally developed to study riot participation in Africa) and split the questionnaire into two parts. The first part included background information about the respondent, the less sensitive questions, and the list experiments. The second one included the more sensitive questions about voting behavior, ideology, and political preferences. Each part of the questionnaire was marked with a different survey identification number, which could only be matched with a document not available to the enumerators. Other than this number, the second part of the questionnaire had no information—such as age, gender, or occupation—that could be used to identify the respondent.

Enumerators administered the first part of the questionnaire, while the sensitive part was read and filled out by the respondents themselves. Other public employees who were present at the time of the survey were therefore able to hear neither the questions nor the answers. This part of the questionnaire was purposely designed to be short and easy to understand and answer, with only closed-ended questions. At the end of the interview, respondents were asked to insert this second part of the questionnaire in a sealed cardboard box similar to a ballot box. Enumerators were instructed to provide a detailed explanation of these procedures before handing the sensitive part of the questionnaire to the respondents and to make sure respondents understood that the survey fully protected the confidentiality of their

[1] Note that the information I was requesting (a complete list of public employees) is something that is public information in most advanced democracies. This information is not sensitive in itself and does not put the research subjects at risk.

[2] For instance, while the survey included a list experiment question about attending rallies, there was no question about the existence of any sort of pressure from the local authorities to attend those rallies. More generally, there was no question about the mayor's role in getting public employees to perform political services.

[3] All the details of the preliminary fieldwork and the interviews were reported in the methodological appendix of the book (Oliveros 2021, 207–22). An alternative would have been to include them in a Pre-Analysis Plan, as suggested by Pérez and Tiscornia in their contribution to this symposium.

responses. Their understanding was critical to ensure the success of the data collection strategy.[4]

The specific details about this design strategy were based on preliminary fieldwork and good contextual knowledge. First, an idea of the types of places where interviews would be taking place and crucially the fact that there would be little to no privacy in most of these settings, was key in coming up with my decision to segment the questionnaire. Since the goal was to interview employees in mid-level secretarial and administrative roles and professionals, as well as employees in low-level positions such as street sweepers, janitors, drivers, maintenance workers and security officers, thinking about the places where interviews would be conducted was important. Second, knowledge of the case was also relevant to the decision to use a cardboard box similar to a ballot box. I knew this would be familiar to respondents because paper ballots and cardboard ballot boxes are used in Argentinean elections. This familiarity made the strategy easier to understand. Third, being confident that respondents would be able to fill the sensitive part of the questionnaire by themselves—literacy rates are high in Argentina—was also vital.[5]

Above all, preliminary fieldwork made it clear that some of the questions in the survey were indeed sensitive and that strategies to protect anonymity were therefore necessary. For instance, take my questions about the political services that public sector employees perform on behalf of their patrons. Along with questions on voting behavior and political preferences, these were the toughest to ask. Employees could be unwilling to reveal that kind of information in front of others, but it was also possible that they would be unwilling to reveal the information in private or, even worse, provide inaccurate responses. For these types of questions, I opted to use list experiments. List experiments (and indirect questioning in general) are typically used to improve measurement of behavior or beliefs the respondents would prefer to hide. In the case of the political services studied here, however, it was possible that some employees would actually want to broadcast their contributions and loyalty to the incumbent. But whether an employee would prefer to broadcast or hide his or her political contributions was not random. For instance, most interviews with low-skilled workers took place in front of others, sometimes including their own bosses. If bosses or coworkers were supporters of the incumbent, one could expect the employee to have an incentive to over-report his or her contributions to political services. But bosses or co-workers could also be employees appointed by the previous administration or via meritocratic processes, in which case employees might prefer to hide their political activities. The advantage of list experiments is that they prevent both underreporting and overreporting.

Considering the provision of favors (one of the political services I studied) a sensitive issue might be counterintuitive.[6] After all, providing favors is a way to help others in the community. Preliminary interviews show that in some cases, employees show pride in being helpful. In other cases, however, the sensitivity of the issue was quite evident. A broker and public sector employee from Greater Buenos Aires that I interviewed during my preliminary fieldwork provides a good example of how someone could get slightly offended by the implication that employees provide favors. After a couple of questions about favors, he replied emphatically: "But politics is not a favor machine! (*una máquina de hacer favores*)."[7] Another Peronist broker and public employee from the province of Buenos Aires wanted to make sure not to give the impression that providing favors was a broker's main role: "Peronism is not just about helping people (*no es solamente asistencia*)…Assisting people is just a small part."[8]

The literature on clientelism tends to assume that this is always a sensitive issue (González-Ocantos et al. 2012). By contrast, in the interviews I conducted with political brokers in Argentina it was clear that this was not always a sensitive issue for them, and that some were willing to discuss openly a lot of things researchers consider sensitive. Of course, the framing of the questions matters and asking bluntly if they "buy votes" may not be a good strategy. But, for the most part, brokers are proud of the work they do. They often cite helping those in need as one of their duties, and in the cases of public employees, they do not hide that their jobs were obtained because they were political brokers who could perform that sort of work. What is more: some consider patronage jobs fair compensation for their political contributions. Brokers emphasize that *on top of* doing their job in the public administration as everyone

[4] To test the effectiveness of the strategy, I included an additional question about the upcoming presidential election in the questionnaire fielded in one of the municipalities. Half of the respondents were asked this question directly (in the first part of the questionnaire); the other half found this question at the end of the sensitive part which they completed in private. The results confirm my intuition about the importance of affording respondents higher levels of anonymity. Employees responded differently when asked under the protected scheme (see Oliveros 2021).

[5] According to the 2010 Argentinean census, only 1.96 percent of the total population older than 10 years old is illiterate.

[6] On the provision of favors, see also Oliveros (2016).

[7] Author's interview, La Matanza, August 10, 2009.

[8] Author's interview, La Plata, August 5, 2009.

else does, they also perform political work. Knowing that brokers were open to discuss their political work was key for conducting successful in-depth interviews—knowledge that could not have been drawn necessarily from the existing literature.

In sum, whether a topic is sensitive or not is often an empirical question and not something that researchers can assume beforehand.[9] Moreover, the sensitivity of the issue varies by research strategy as well—a sensitive issue in a survey may not be that sensitive in an interview setting where the researcher can establish rapport with the interviewee. One can, of course, choose to err on the side of caution, but strategies to deal with sensitive questions are not without cost. For instance, using the split questionnaire strategy described above meant that the survey took longer to complete because enumerators had to spend time explaining the procedure. In the case of list experiments there is also a well-known trade-off between accuracy and efficiency. List experiments reduce response bias by minimizing the incentives for respondents to lie, but they do so at the cost of efficiency.[10] Moreover, for successful implementation, methods of indirect questioning for sensitive questions, such as list experiments, require larger sample sizes than direct questioning (Corstange 2009; Yadav 2015). For these reasons, strategies to deal with sensitive issues should only be used when the issues are indeed sensitive—a key empirical question that the researcher needs to address during preliminary fieldwork.

The List Experiment

The list experiment technique I used to ask about the provision of political services is straightforward. The sample is randomly split into a treatment and a control group. Each group is read the same question and shown a card with a number of response options.[11] List experiments work by including the item one cares about (the sensitive item) in a list containing other items, usually non-sensitive ones. Cards for the two groups differ only in the number of response categories. Respondents are asked to report the number of items on the list that apply to them, but not which ones. Since respondents are randomly assigned to either the group with the sensitive item (treatment) or the one without it (control), the two groups are, on average, indistinguishable on observable and unobservable characteristics. Differences in the mean number of items, or in my case, activities, reported by the two groups therefore provide a point estimate of the proportion of respondents who performed the sensitive activity.[12]

List experiments are, of course, not the only method of indirect questioning to deal with social desirability bias. Two interesting alternatives are the randomized response technique (e.g., Gingerich 2013) and the crosswise model (e.g., Corbacho et al. 2016). I chose to use list experiments over these alternatives mainly for their simplicity. Instructions are easy to understand, and respondents tend to trust that the anonymity of their responses will be protected (Coutts and Jann 2011). Since respondents were low- and mid-level employees, some with low levels of education, this simplicity was an important advantage.

Although the technique is fairly easy to implement and understand, it is still more demanding than direct questioning. Careful survey implementation is crucial for obtaining accurate responses. Preliminary fieldwork and being in the field at the time of the pilot were therefore key. Two examples from my experience illustrate this point. In both cases, being in the field at the time of the pilot, in permanent contact with the enumerators, and conducting many survey interviews myself made me realize two simple issues with list experiments that, at the time of the survey, were not mentioned in the literature on best practices.[13]

First, during the pilot I uncovered two types of error responses by respondents who did not follow or did not understand the instructions. One type occurred when respondents provided a count of the frequency with which they performed each of the activities on the list, instead of counting the items or activities that applied to them. The second type of error was identifying the item or items by using their numbers on the list, causing confusion about whether they were referring to the number of activities that applied to them or to a specific activity on the list (which was not what I wanted). Because of this discovery during the pilot, I decided to switch the numbers to letters, so the cards listed the items by letter (A, B, C) instead of by number. The use of letters instead of numbers to order the list made confusion with the instructions evident to the enumerators, who were instructed to repeat the instructions if respondents showed any lack of understanding. Because the survey

[9] This resonates with Bell-Martin's claim in her article in this symposium that "ethnographic evidence facilitates greater construct and ecological validity of our instruments" (p. 2). While I didn't conduct an ethnography, in-depth interviews served a similar purpose.

[10] The standard errors for list experiment estimates are larger than they would have been for a direct question with no response bias (Blair and Imai 2012; Corstange 2009).

[11] For this project, the list of responses was not read aloud to increase privacy.

[12] For other examples of the use of list experiments to measure clientelism and patronage see, for instance, Frye, Reuter, and Szakonyi (2014), González-Ocantos et al. (2012), González-Ocantos and Oliveros (2019), and Mares and Young (2018; 2019).

[13] Note that the survey was implemented in 2010 and 2011. Since then, a lot has been written about how to conduct list experiments.

included four list experiments, enumerators had a chance to explain the procedure again if the reaction to the first experiment had alerted them to a misunderstanding. Opting for letters instead of numbers was a free and easy solution that surely increased the accuracy of the responses.

The second issue had to do with "floor effects." To protect anonymity in list experiments it is crucial to avoid lists that could result in respondents choosing none or all of the items, generating "floor" or "ceiling" effects, respectively (see Kuklinski et al. 1997). If a respondent's truthful answer were "yes" or "no" to all the items in the control list, the list experiment would fail to provide the desired deniability on the sensitive item. In other words, respondents would necessarily have to reveal their participation in the sensitive activity when answering sincerely. To minimize ceiling effects, lists usually include rare activities or activities that one cannot perform concurrently. To minimize the risk of floor effects, high-prevalence activities are often included. In my survey, the strategy to minimize ceiling effects was successful and only around one percent of respondents in the control groups for all list experiments reported all four of the control items. The inclusion of high-prevalence activities to minimize the risk of floor effects was less successful. Although I am not aware of any systematic study of this issue, anecdotal evidence from the survey interviews that I conducted suggests that at least some of those zero responses were indeed "DK/NA." List experiments do not include this response option, so when respondents were in a hurry or did not want to answer for any reason, a "zero" response seemed to be the choice. This implies that even in well-designed list experiments in which high prevalence items or activities are included, a number of zero responses may be unavoidable. Although I discovered this issue while in the field, there was little to do about it. Some public employees were indeed in a hurry and chose the zero response. Knowing this, however, was key to my understanding that there was nothing intrinsically wrong with the design of the list experiment. In the end, because the presence of either ceiling or floor effects leads to the underestimation of the sensitive activity (Blair and Imai 2012), this meant that the list experiment estimates were likely conservative.

Concluding Thoughts

The importance of fieldwork and good contextual knowledge for design-based research strategies (Dunning 2012) such as experiments cannot be overstated. Experiments (both field experiments and survey experiments) require that most of the research effort is done before the implementation phase. Once the experiment is in the field, there is little room to turn back the clock on design choices. When the issues under study are sensitive political phenomena—like clientelism or patronage—preliminary fieldwork is even more critical, for both practical and ethical reasons.

From a practical standpoint, failing to acknowledge the sensitivity, or lack thereof, of a particular issue could mean ending up with poor data. In a case in which the researcher does not realize how sensitive an issue is, this could mean that responses are biased, inaccurate, or just plain refusals. But if the researcher choses one of the available techniques to deal with social desirability bias—such as the list experiment or the segmented questionnaire describe above—in a context in which this is not necessary, estimates may end up being less efficient or data more costly to gather. Research strategies designed to deal with sensitive issues should therefore only be used when the issues are indeed sensitive. However, whether an issue is sensitive or not in a particular context is an empirical question. The way to avoid both of these potential problems is to conduct thorough preliminary fieldwork.

From an ethical standpoint, preliminary fieldwork is also vital to assess the sensitivity of the issue and the potential risks for research subjects. Obtaining inaccurate responses due to misreporting or non-response bias is not the worst outcome of a poorly designed research strategy; putting subjects at risk—even if minimal—is. Of course, this is more relevant for subject areas that are more sensitive than patronage. In the end, Argentina is a well-functioning democracy and the "risk" of others finding out about the political preferences or activities of coworkers in the public administration is not serious. Still, it could lead to uncomfortable situations that need to be avoided. Strategies like the ones described above, such as not reading the questions aloud or keeping separate the responses to sensitive questions from the ones that could lead to the identification of an employee, are good examples of effective strategies to protect respondents. And protecting respondents should always be our primary goal.

References

Blair, Graeme, and Kosuke Imai. 2012. “Statistical Analysis of List Experiments.” *Political Analysis* 20, no. 1 (Winter): 47–77. https://doi.org/10.1093/pan/mpr048

Corbacho, Ana, Daniel W. Gingerich, Virginia Oliveros, and Mauricio Ruiz-Vega. 2016. “Corruption as a Self-Fulfilling Prophecy: Evidence from a Survey Experiment in Costa Rica.” *American Journal of Political Science* 60, no. 4 (October): 1077–92. https://doi.org/10.1111/ajps.12244

Corstange, Daniel. 2009. “Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT.” *Political Analysis* 17, no. 1 (Winter): 45–63. https://doi.org/10.1093/pan/mpn013

Coutts, Elisabeth, and Ben Jann. 2011. “Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT).” *Sociological Methods & Research* 40, no. 1 (February): 169–93. https://doi.org/10.1177/0049124110390768

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. New York: Cambridge University Press.

Frye, Timothy, Ora John Reuter, and David Szakonyi. 2014. “Political Machines at Work Voter Mobilization and Electoral Subversion in the Workplace.” *World Politics* 66, no. 2 (April): 195–228. https://doi.org/10.1017/S004388711400001X

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

González-Ocantos, Ezequiel, Chad Kiewiet de Jonge, Carlos Meléndez, Javier Osorio, and David W. Nickerson. 2012. “Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua.” *American Journal of Political Science* 56, no. 1 (January): 202–17. https://doi.org/10.1111/j.1540-5907.2011.00540.x

González-Ocantos, Ezequiel, and Virginia Oliveros. 2019. “Clientelism in Latin American Politics.” In *The Encyclopedia of Latin American Politics*, edited by Gary Prevost and Harry Vanden. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1677.

Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. 1997. “Racial Prejudice and Attitudes toward Affirmative Action.” *American Journal of Political Science* 41, no. 2 (April): 402–19. https://doi.org/10.2307/2111770

Mares, Isabela, and Lauren E. Young. 2018. “The Core Voter’s Curse: Clientelistic Threats and Promises in Hungarian Elections.” *Comparative Political Studies* 51, no. 11 (September): 1441–71. https://doi.org/10.1177/0010414018758754

———. 2019. *Conditionality & Coercion: Electoral Clientelism in Eastern Europe*. Oxford: Oxford University Press.

Oliveros, Virginia. 2016. “Making It Personal: Clientelism, Favors, and the Personalization of Public Administration in Argentina.” *Comparative Politics* 48, no. 3 (April): 373–91. https://doi.org/10.5129/001041516818254437

———. 2021. *Patronage at Work: Public Jobs and Political Services in Argentina*. New York: Cambridge University Press.

Scacco, Alexandra. 2010. “Who Riots? Explaining Individual Participation in Ethnic Violence.” PhD diss., Columbia University.

Van de Walle, Nicolas. 2007. “Meet the New Boss, Same as the Old Boss? The Evolution of Political Clientelism in Africa.” In *Patrons, Clients and Policies: Patterns of Democratic Accountability and Political Competition*, edited by Herbert Kitschelt and Steven I. Wilkinson, 50–67. New York: Cambridge University Press.

Yadav, Vineeta. 2015. “Lobbying, Corruption, and Non-Responses in Small Samples.” *Comparative Politics Newsletter* 25, no. 1 (Spring): 32–37.

Transparency and Replicability in Mixed-Methods Designs using Experiments[1] [2]

Verónica Pérez Bentancur
Universidad de la República

Lucía Tiscornia
Centro de Investigación y Docencia Económicas

The use of mixed methods designs containing experiments has become more popular in the social sciences over the past decades (Harbers and Ingram 2020; Seawright 2016; Weller and Barnes 2014). In the analysis of experimental results, the qualitative component is typically used to illuminate causal mechanisms (Dunning 2015; Paluck 2010). However, when it comes to improving experimental designs, the capacity of qualitative methods to improve measurement is discussed less frequently. Prior to the analysis of data, qualitative methods can be used to design better contextualized, more realistic, experimental treatments (Dunning 2008; Dunning and Harrison 2010; Seawright 2016; 2021)[3]. Yet, the process of using qualitative methods to improve treatment design, for example, through the establishment of a sequence that can be replicated, is rarely formalized. We highlight the importance of standardizing the use of qualitative research to improve experimental treatments by pre-registering it as part of a pre-analysis plan (hereafter, PAP). In formalizing this process, researchers can contribute to the transparency and replicability of the entire research process.

Defining realistic treatments is usually a challenge for experimental designs. Experimentalists debate how best to produce realistic treatments (Blair and McClendon 2021). In other words, researchers are concerned with generating treatments that make subjects react as if they had experienced the same situation in real life. If they succeed, the study is considered to have a high level of "experimental realism" (Seawright 2021). However, experiments in general lack a sense of realism because they present hypothetical situations to participants. Contextual knowledge can contribute to ameliorating this limitation because it allows researchers to anchor the experiment in real circumstances with a level of specificity not possible in a hypothetical.

Recent debates in the experimental literature reflect the advantages of incorporating qualitative elements to enhance realism (Seawright 2021). Qualitative research encompasses a wide variety of tools with varying degrees of involvement (from ethnographic research to the review of press or historiographic accounts) (Cyr 2019; Harbers and Ingram 2020; Kapiszewski, MacLean, and Read 2015). Adopting a wide range of qualitative methods and tools may not always be possible within the timeframe of a study, or given the available resources. Yet, researchers can take advantage of this wide range of tools by incorporating some inductive elements drawn from qualitative research to strengthen their experimental design. Moreover, this can be accomplished without detracting from the overall research, but rather enabling improvement of its design.

The use of qualitative research tools in the context of experimental designs allows researchers to strengthen causal inference by further developing the analytical phase of a research design, but also by enhancing the design itself.[4] For example, Oliveros's and Bell-Martin's articles in this symposium discuss the art of combining qualitative fieldwork and experiments. Moreover, some studies specifically rely on qualitative research to build more realistic treatments, but this process is rarely formalized.[5] Formalization would enable researchers to make their designs more transparent and replicable.

We argue that pre-registration—making a research design public— can boost researchers' efforts to formalize the incorporation of qualitative components to improve experimental designs. In formalizing these steps, researchers make the entire research design—not just the forward-looking parts—transparent and replicable. Experimentalists currenrly debate what components of their work should be pre-registered. There is therefore no single, standardized criteria for

[1] Authors in alphabetical order, equal contribution is implied.

[2] Part of the reflections in this article resulted from our experiences developing a research project with Juan Albarracin and Leslie MacColman, funded through the Agencia Nacional de Investigación e Innovación (Uruguay), project number FSSC_1_2018_1_147720.

[3] By experimental treatment we mean the different kind of manipulations used in surveys and field experiments. Even though a randomized controlled trial may require less emphasis on contextual knowledge given that interventions are generally more realistic, contextual knowledge is still crucial to identify the best suited intervention and to adjust it to the specificities of the place and people.

[4] Even though in some designs the distinction between analytical and design phase may be blurry, our focus is on experimental designs. In those cases, the analytical phase is defined by the statistical analysis of data.

[5] See for example Auerbach and Thachil (2020). They use ethnographic methods to define experimental treatments resulting in a compelling research design. However, the steps taken were not specified.

what should be included in a PAP (Boudreau 2021). In this piece we discuss the process of formalization of the qualitative phase in experimental research designs and propose some sections that should be incorporated as part of a PAP to that effect. In doing so, we contribute to the literature on research transparency in political science (Jacobs 2020; Jacobs et al. 2021; Kapiszewski and Karcher 2021).

The Benefits of Incorporating a Qualitative Phase in the Design of Experimental Research

Mixed-methods designs including experiments have gained popularity in the social sciences. Experimentalists have begun to consider the benefits of including qualitative elements in designs containing experiments in order to improve causal inference. Qualitative elements consist of non-numeric, detailed information obtained through ethnographies, in-depth interviews, direct observation, focus groups, systematic press review, archival research, and historical sources, among others. Designs that incorporate qualitative components typically include them in the analytical phase (Clayton et al. 2020; Dunning 2015; Paluck 2010). They focus on techniques such as interviews, focus groups, and direct observation, as well as the advantages of these techniques in identifying causal mechanisms. More recently, other researchers have highlighted the importance of the qualitative phase as a mechanism to improve the experimental design itself by enhancing internal and external validity (Jha, Rao, and Woolcock 2007; Pérez Bentancur and Tiscornia 2022; Seawright 2021; Thachil 2017; 2018; Tiscornia et al. 2021).

Improving causal inference requires high internal validity. In turn, internal validity requires a high degree of correspondence between the experimental intervention and the context. This is a challenge for experiments, as there is always some level of "fakeness" in interventions (Blair and McClendon 2021). The goal, therefore, is to devise treatments that reproduce situations that feel realistic to experimental subjects. This requires incorporating contextual knowledge, or what Seawright (2021, 371) calls "the meaning of treatment." Better measurement also ensures construct validity—how well the researcher's measurement instrument can capture the phenomenon to be measured.

Better contextual knowledge also improves external validity. Deep knowledge of the context produces a better understanding of the scope of a theoretical argument, which allows researchers to be more precise when it comes to their ability to generalize.[6] This is a central point, as a potential weakness of experiments is low external validity (Blair and McClendon 2021; Seawright 2016).

Thachil (2017, 2018) and Dunning and Harrison (2010) are good examples of the use of qualitative research to improve experimental designs. Thachil (2017, 2018) uses ethnographic fieldwork to define his sample and improve a vignette experiment in his research on identity politics in India. To build his sample, Thachil (2017, 913) used interviews to come up with a list of markets where his population of interest, which is highly mobile, would be present. To improve the vignettes, he used conversations he maintained with poor migrants to ensure that the language was reflective of the target population (2017, 909). Focusing on ethnic cleavages in Mali, Dunning and Harrison (2010) used interviews to validate and refine an experiment prior to fielding it (Dunning 2008, 21-22). By contrast, Clayton et al. (2020) conduct focus groups and experiments to assess the extent of voters' gender bias in Malawi but when they designed the experimental vignettes they did so without including relevant information from the focus groups, which they conducted in parallel. The authors speculate that this omission may have led to null findings in the experiment. From the focus groups they learned that women in Malawi face defamation campaigns, which impacts voting behavior. As a result, the authors conclude that they could have used this information to include a "rumor mongering condition" in their experiment (2020, 622).

The examples highlight the benefits of incorporating qualitative evidence to design better treatments. With few exceptions—for example, Dunning (2008) —the majority of research projects that include qualitative components to improve experiments do not incorporate them as part of the pre-registration process (Pérez Bentancur and Tiscornia 2022).

There is a wide range of qualitative methods and tools whose application would benefit researchers that employ mixed-methods designs with experiments. These methods and tools require different levels of involvement and resources (Kapiszewski, MacLean, and Read 2015; Curini and Franzese 2020). Researchers can decide which of these tools are more adequate to their design, or feasible to implement based on considerations such as time constraints and budget. Similarly, not all experimental designs require the incorporation of qualitative components to the same degree. Including qualitative elements is relevant to the extent that it provides researchers with better contextual knowledge when designing the treatment.

[6] Encinas develops this point in further detail in his contribution to the symposium.

Researchers that combine methods might even consider pre-tests of their experimental designs as part of these qualitative elements, as long as they are not simply focused on the mechanical aspects. For example, in our research we pre-tested a survey experiment and we asked participants to provide a brief reflection on the experience of taking the survey with an open-ended question at the end, which we then used to adjust the experiment that we fielded (Perez Bentancur and Tiscornia 2022). Doing so led us to reconceptualize the design phase as a non-linear process, a back-and-forth between deductive and inductive phases. Many times, researchers' starting point is deductive, but they will adjust their theory and hypotheses by alternating between deduction—induction—deduction. Acknowledging non-linearity in research design also constitutes an exercise in transparency because it reveals all the steps the researchers took to arrive at the final design (Yom 2015).

The examples above suggest that researchers can, and do, incorporate qualitative elements to improve experimental designs without pre-registering them. We argue, however, that there are important benefits in formalizing the steps researchers take through a discussion in a PAP. In documenting how qualitative components improve experimental designs, we contribute to the literature on research transparency (Elman, Kapiszewski, and Lupia 2018). There are a variety of alternatives for documentation, such as generating methodological appendices (Kapiszewski and Karcher 2021). We propose implementing pre-registration through PAPs as one of these alternatives. Using PAPs increases transparency in at least three ways. First, it improves our understanding of the process of producing research; second, it allows a better understanding of how researchers reached their conclusions; and third, it makes research more accessible because it makes all steps in the research process explicit. Replicability thus results from transparency. Replicability implies that another researcher will be able to reconstruct the research process in the same context (or a different one) and reach similar conclusions about a phenomenon.[7] This requires information about how the data was produced and how the analysis was conducted (Jacobs et al. 2021).

Transparency and Replicability in Experimental Research Design: The role of PAPs

In recent years academics have raised concerns regarding the difficulty of replicating research designs, to the point where some characterize the current state of affairs as a "replication crisis" (Druckman and Green 2021; Malhotra 2021). The inability to replicate studies results from a lack of transparency in the research process. The absence of sufficient detail in research designs prevents other research teams from following the exact same steps with the goal of attaining the same result, which is central to the concept of replication. Consequently, some researchers have proposed pre-registration as a solution to the replication crisis (Boudreau 2021; Malhotra 2021).

Pre-registration consists of developing research questions, hypotheses and analyses before observing the data, and making this information public on an independent registry. A PAP is a document describing the process that will be used to collect and analyze data (Boudreau 2021; Chen and Grady, n.d.) including, but not limited to, hypotheses, experimental designs, a description of the population to be studied. Pre-registration through a PAP prevents certain biases and increases research transparency by detailing the necessary steps for replication (Blair et al. 2019; Boudreau 2021; Jacobs 2020; Malhotra 2021; Pérez Bentancur and Tiscornia 2022).

Although researchers tend to agree on the importance of pre-registration, there is no consensus on exactly which elements of a research design to pre-register beyond the basic structure (hypotheses, measurement, tests). In fact, there is no single, standardized definition of a PAP, as evidenced by the wide variety of templates in different repositories, and the different degrees of flexibility allowed for documentation (Boudreau 2021).

Discussions around pre-registration emphasize its importance in promoting transparency. Typically, PAPs are linked to experimental designs with an emphasis on quantitative components: researchers typically register the treatment of interest, sample and subgroup characteristics, statistical power analyses, and steps for data analysis. Many times, researchers use qualitative elements to improve their experimental designs. For example, Thachil (2017) uses observations from the field to construct experimental vignettes. In our research about the micro-foundations behind public support for punitive policing, we incorporate information from interviews to refine experimental treatments (Tiscornia et al. 2021). Yet, this qualitative phase is rarely documented as part of PAPs. Our analysis of all available PAPs that use mixed-methods pre-registered in Evidence in Governance and Politics (EGAP) repository in 2019 (a total of 338) shows that only a small fraction (14, only 4%) explicitly report

[7] Replicability refers to the process of information gathering, for example, the use of qualitative tools to inform an experiment. This does not mean that interviews or ethnographic research should be replicated, as the specific qualitative technique may have to be adapted based on considerations of applicability, time, resources, risks, or others.

using the qualitative component to refine the research design. However, those PAPs do not include enough detail to allow replication of the qualitative component. For example, Brugger and Bezzola (2020) explicitly state in their PAP that they use fieldwork to identify relevant experimental outcomes. Yet, they do not discuss how fieldwork led to the changes they made (Pérez Bentancur and Tiscornia 2022).

One might wonder why they should include qualitative components as part of the PAP and not, for example, in the final manuscript. Even though a discussion of the use of qualitative research as part of the final manuscript is relevant, when it comes to transparency and replicability it is important to incorporate formalized steps as part of the PAP. Researchers may want to compare the original objectives of the research design with what was achieved in the final manuscript (Boudreau 2021, 348). Besides, not all studies become published manuscripts (for example, publication bias against research that results in null findings might prevent others from knowing their results) which can bias what we know about a specific phenomenon (Boudreau 2021, 341; Malhotra 2021, 356). In addition, the level of detail needed for replication might not be relevant or appropriate for a manuscript that might have word limits or where too much detail might distract from substantive points.

If researchers use qualitative steps to refine an experimental design, incorporating them as part of a PAP independent of publication status of the final manuscript has several advantages: (1) it increases transparency because it illustrates all the steps used to arrive at the final design; (2) it helps minimize the possibility that a null or contradictory finding is the result of design flaws derived from absence of relevant contextual factors; (3) it aids replicability by facilitating adaptation of the experimental treatment to different contexts. PAPs are public goods; when researchers register as many relevant details of the research design as possible, they contribute to the improvement of their own research but also of research as a collective enterprise.

If they are so relevant, why are qualitative components not typically included as part of PAPs? One possibility is that researchers may choose not to incorporate qualitative components as part of experimental designs because they do not view them as necessary or because they view them as "pre-scientific," and therefore do not think it is relevant to incorporate them as part of a PAP. Another alternative could be that existing PAP templates may not allow for this option. Yet, this is not the case as repositories are typically quite flexible in terms of what can be included in a PAP.

Even when researchers do not use a mixed-methods approach, experimental designs require some level of contextual knowledge that comes from qualitative sources. The kind or the extent of qualitative tools that researchers decide to incorporate might depend on considerations such as budget, time, access, even epistemological views. Other researchers can evaluate whether the resulting research design would have required additional qualitative elements. Regardless of the kind of qualitative tools they choose, if researchers choose to incorporate qualitative elements to strengthen their experimental designs, they should pre-register them.

Existing repositories provide varying degrees of flexibility to incorporate additional descriptive information (Boudreau 2021). For example, EGAP's repository allows for the inclusion of the kind of information we propose (see for instance Blair and Weintraub 2020).8 An additional tool is the possibility to include amendments to the original PAP. In general, repositories allow for the submission of revised PAPs, if they are submitted before the intervention takes place. Researchers can pre-register a design, then conduct interviews or archival research, adjust their experimental design and submit an amendment.

If one were to incorporate qualitative components as part of a PAP, where would they add them? Recent research has provided some guidelines for relevant components in PAPs (Boudreau 2021). We believe that within these guidelines there is also space to include qualitative elements. For example, if researchers used qualitative components to improve their treatment, they could specify how in the section dedicated to describing the experimental design, or within a measurement section. The qualitative component could be tied to the sections devoted to describing the design, the definition and operationalization of the treatment, or the construction of the sample, before discussing how the researcher plans to analyze the data.

Suggestions for Incorporating Qualitative Elements in PAPs

In this essay we have highlighted the usefulness of thinking about qualitative data beyond the analytical components of research. We argue that qualitative tools play a central role in mixed-methods research designs containing experiments because they contribute to improving treatments, sampling, and other components of the design. Although in many cases researchers incorporate these elements as part of their designs, they rarely make these steps explicit. We suggest that researchers can incorporate these steps as part of a PAP.

[8] This PAP represents an example of how flexible these documents can be as it includes an extensive discussion of ethical considerations.

In doing so, researchers will improve transparency and replicability.

Implementing this process may require rethinking the phases of the research design, data collection, fieldwork, or even the kind of information that needs to be collected. It may lead researchers to reconceptualize the research design phase as a non-linear process. As part of the research design, there is an inductive phase that typically happens after the initial deductive research design.

We propose that incorporating qualitative data is crucial because experiments are contextual, even when researchers do not use mixed-methods approaches. Depending on feasibility, this can be done with varying degrees of depth. It may be the case that researchers have accumulated enough contextual knowledge that a qualitative phase may be deemed unnecessary, or they may not have sufficient resources to incorporate it. Even if researchers are using their contextual knowledge to improve their design, they should still make the steps they took to arrive at their final design, and the assumptions they made along the way, explicit as part of their PAP.

Based on this discussion, we propose a series of sections that researchers can incorporate when developing their PAP, once they have decided to include qualitative elements in their research design. If the qualitative elements are incorporated to improve the treatment, researchers should discuss how they did it in the section on measurement. For example, did specific interview questions contribute to illuminating relevant treatments the researcher had not previously considered? Researchers should register what the original experimental design was and how it was modified based on the qualitative data they used. Perhaps the list of attributes in a conjoint design was expanded or reduced based on archival research, or new items were included as part of a list experiment that resulted from direct observation, or a focus group. Or maybe the type of experiment changed—for example, from a conjoint design to a vignette (Tiscornia et al. 2021). Alternatively, researchers may use qualitative information to better contextualize a vignette, or to mimic the language used in the research setting, and thus make it more reflective of the reality on the ground (Masullo and Morisi 2022; Bell-Martin in this symposium). They might choose to report this in a section on construct validity. Scholars could also use qualitative information to construct an experimental sample (see for example, Thachil 2022; 2018; Oliveros in this symposium). In this case, they may incorporate the discussion of the steps they took in a specific section dedicated to the sample. Scholars should carefully register all these steps as part of the PAP, as if they were instructions to be followed. It is not enough to say what qualitative elements they included, but how and with what aim.

These suggestions do not only apply to experimental designs; they could also be easily incorporated into observational designs. Observational research can also be pre-registered before analysis, and it can include details on the use of qualitative components. For example, survey questionnaires can be improved by incorporating qualitative information, and codebooks can be analyzed to detect possible biases in the construction of databases used for analytical purposes.

The discussion presented here echoes recent academic concerns about the importance of pre-registration of research designs as it contributes to research transparency and replicability. We build on this point to highlight that in the context of mixed-methods research with experimental components, scholars use qualitative elements as part of the process of experimental design, but without pre-registering them as part of a PAP. We propose that in formalizing these steps in a PAP by clearly specifying the role of qualitative elements, researchers can make the process of designing their research more transparent, and more easily replicable, which benefits their own research and the research enterprise as a whole.

References

Auerbach, Adam Michael, and Tariq Thachil. 2020. "Cultivating Clients: Reputation, Responsiveness, and Ethnic Indifference in India's Slums." *American Journal of Political Science* 64 , no. 3 (July): 471–87. https://doi.org/10.1111/ajps.12468.

Brugger, Fritz, and Selina Bezzola. 2020. "Social Investments of Mining Companies and Citizen Engagement in Local Governance." OSF. February 3. doi:10.17605/OSF.IO/8V5TW.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113, no. 3 (August): 838–59. https://doi.org/10.1017/S0003055419000194.

Blair, Graeme, and Gwyneth McClendon. 2021. "Conducting Experiments in Multiple Contexts." In *Cambridge Handbook of Experimental Political Science.*, edited by Donald P. Green and James N. Druckman, Second edition, 411–30. New York: Cambridge University Press.

Blair, Robert, and Michael Weintraub. 2020. "Mano Dura: An Experimental Evaluation of the Plan Fortaleza Program in Cali, Colombia." OSF. February 3. doi:10.17605/OSF.IO/95CZ3.

Boudreau, Cheryl. 2021. "Transparency in Experimental Research." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green, Second Edition, 339–53. New York: Cambridge University Press.

Chen, Lula, and Chris Grady. n.d. "10 Things to Know About Pre-Analysis Plans." *EGAP* (blog). https://egap.org/resource/10-things-to-know-about-pre-analysis-plans/.

Clayton, Amanda, Amanda Lea Robinson, Martha C. Johnson, and Ragnhild Muriaas. 2020. "(How) Do Voters Discriminate Against Women Candidates? Experimental and Qualitative Evidence From Malawi." *Comparative Political Studies* 53, no. 3-4. (March): 601–30. https://doi.org/10.1177/0010414019858960.

Curini, Luigi, and Robert Franzese. 2020. *The SAGE Handbook of Research Methods in Political Science and International Relations*. Thousand Oaks: SAGE.

Cyr, Jennifer. 2019. *Focus Groups for the Social Science Researcher*. New York: Cambridge University Press.

Druckman, James N., and Donald P. Green. 2021. "A New Era of Experimental Political Science." In *Cambridge Handbook of Experimental Political Science.*, edited by James N. Druckman and Donald P. Green, Second Edition, 1–18. New York: Cambridge University Press.

Dunning, Thad. 2008. "Natural and Field Experiments: The Role of Qualitative Methods." *Qualitative & Multi-Method Research* (Fall): 17–23.

———. 2015. "Improving Process Tracing: The Case of Multi-Method Research." In *Process Tracing. Fom Metaphor to Analytic Tool*, edited by Andrew Bennett and Jeffrey T. Checkel, 211–36. New York: Cambridge University Press.

Dunning, Thad, and Lauren Harrison. 2010. "Cross-Cutting Cleavages and Ethnic Voting: An Experimental Study of Cousinage in Mali." *The American Political Science Review* 104, no. 1 (February): 21–39. https://doi.org/10.1017/S0003055409990311

Elman, Colin, Diana Kapiszewski, and Arthur Lupia. 2018. "Transparent Social Inquiry: Implications for Political Science." *Annual Review of Political Science* 21, no. 1 (May): 29–47. https://doi.org/10.1146/annurev-polisci-091515-025429.

Harbers, Imke, and Matthew C Ingram. 2020. "Mixed-Methods Designs." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, edited by Luigi Curini and Robert Franzese. Thousand Oaks: SAGE.

Jacobs, Alan M. 2020. "Pre-Registration and Results-Free Review in Obsevational and Qualitative Research." In *The Production of Knowledge. Enhancing Progress in Social Science*, edited by Colin Elman, John Gerring, and James Mahoney. New York: Cambridge University Press.

Jacobs, Alan M., Tim Büthe, Ana Arjona, Leonardo R. Arriola, Eva Bellin, Andrew Bennett, Lisa Björkman, et al. 2021. "The Qualitative Transparency Deliberations: Insights and Implications." *Perspectives on Politics* 19, no. 1 (March): 171–208. https://doi.org/10.1017/S1537592720001164.

Jha, Saumitra, Vijayendra Rao, and Michael Woolcock. 2007. "Governance in the Gullies: Democratic Responsiveness and Leadership in Delhi's Slums," in "Experiences of Combining Qualitative and Quantitative Approaches in Poverty Analysis," ed. Ravi Kanbur and Paul Shaffer, special issue, *World Development* 35, no. 2 (February): 230–46. https://doi.org/10.1016/j.worlddev.2005.10.018.

Kapiszewski, Diana, and Sebastian Karcher. 2021. "Transparency in Practice in Qualitative Research." *PS: Political Science & Politics* 54, no. 2 (December): 285–91. https://doi.org/10.1017/S1049096520000955.

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

Malhotra, Neil. 2021. "Threats to the Scientific Credibility of Experiments: Publications Bias and P-Haking." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green, Second Edition, 354–68. New York: Cambridge University Press.

Masullo, Juan and Davide Morisi. 2022. "The Human Costs of the War on Drugs. Attitudes Towards the Militarization of Security in Mexico," OSF Preprints. https://doi.org/10.31219/osf.io/3fy4s

Paluck, Elizabeth Levy. 2010. "The Promising Integration of Qualitative Methods and Field Experiments." *The ANNALS of the American Academy of Political and Social Science* 628, no. 1 (March): 59–71. https://doi.org/10.1177/0002716209351510.

Pérez Bentancur, Verónica, and Lucía Tiscornia. 2022. "Iteration in Mixed-Methods Research Designs Combining Experiments and Fieldwork,." *Sociological Methods & Research*, (March): online-first, https://doi.org/10.1177/00491241221082595..

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. New York: Cambridge University Press.

———. 2021. "What Can Multi-Method Research Add to Experiments?" In *Cambridge Handbook of Experimental Political Science.*, edited by James N. Druckman and Donald P Green, Second Edition, 369–84. New York: Cambridge University Press.

Thachil, Tariq. 2017. "Do Rural Migrants Divide Ethnically in the City? Evidence from an Ethnographic Experiment in India." *American Journal of Political Science* 61, no. 4 (October): 908–26. https://doi.org/10.1111/ajps.12315.

———. 2018. "Improving Surveys Through Ethnography: Insights from India's Urban Periphery." *Studies in Comparative International Development* 53, no. 3 (Italy): 281–99. https://doi.org/10.1007/s12116-018-9272-3.

Tiscornia, Lucía, Verónica Pérez Bentancur, Juan Albarracín, and Leslie MacColman. 2021. "The Social Basis of Punitive Policing: Public Opinion and Perceptions of Deservingness." Paper presented at the Annual Meeting of the International Studies Association (ISA). Held virtually.

Weller, Nicholas, and Jeb Barnes. 2014. *Finding Pathways: Mixed-Method Research for Studying Causal Mechanisms*. New York: Cambridge University Press.

Yom, Sean. 2015. "From Methodology to Practice: Inductive Iteration in Comparative Research." *Comparative Political Studies* 48, no. 5 (April): 616–44. https://doi.org/10.1177/0010414014554685.

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6478365

Experiments as Case Studies: A Qualitative Approach to External Validity[1]

Daniel Encinas
Northwestern University

Over the last few decades, experimental research has gained recognition in political science for enabling the identification of causal relationships. Conventional methodological discussions establish that experiments are generally high in internal validity (i.e., the degree of confidence in causal inferences). At the same time, all experimental studies are "context-dependent" and, thus, "the generalizability of experimental effects is always at issue" (Shadish, Cook, and Campbell 2002, 5). In this sense, scholars regularly consider experiments low in external validity (i.e., the degree of generality of the causal inferences drawn) and encourage us to "be cautious about the conclusions that can be drawn from experimental methods" (Hall 2012, 44; see also Deaton 2010 426; Fukuyama 2013, 93; Thelen and Mahoney 2015, 11).[2] Indeed, while experiments are very capable of dealing with the fundamental problem of causal inference, they are less equipped to tackle the fundamental issue of external validity.

Experimental researchers are increasingly aware of these criticisms. Since the early 2000s, experimenters noticed a "near obsession" with external validity issues "to dismiss experiments" (McDermott 2002, 334). Over time, these researchers developed different empirical strategies for dealing with these issues, to the point that concerns about external validity have become a constitutive feature of the most recent wave of experimental research (Druckman and Green 2021; see also Kam and Trussler 2017, 791). However, new advances in experimental political science for enhancing the level of external validity are overwhelmingly quantitative. Hence, the potential contributions of qualitative methods have been largely neglected in the literature.

The present essay offers an alternative, qualitative approach to enhancing experiments' level of external validity. I advance an understanding of experiments (i.e., laboratory, field, survey, or natural experiments) as case studies. Experiments are extensive, cross-case analyses when considering the units randomly assigned to the treatment and control groups. But experiments might be seen as intensive, within-case analyses from the perspective of the case that contains these randomized units. For instance, a field experiment in Peru would be a study about individuals and Peru. This country would exemplify what I call an *experimental case,* defined as the context in which the randomization takes place or, more precisely, the overarching case that binds the *randomized units* (i.e., the individuals included in the experimental study) together spatially and temporally. Any given experimental study typically exhibits multiple randomized units but a single experimental case.

[1] I especially thank Jamie Druckman, Jim Mahoney, and Jay Seawright for their commentaries on a previous, expanded version of this manuscript. I also thank Pilar Manzi, Bo Won Kim, Salih O. Noor, Rodrigo Barrenechea, participants of the Northwestern University's Comparative Politics Workshop and Latinamericanist Study Group attendees, and participants of the 2020 APSA's Mixing Experiments and Qualitative Methods panel.

[2] See Shadish, Cook, and Campbell (2002) for definitions of internal and external validity (53, 83). See also Imbens and Rubin (2015, 57) and Seawright (2016, chap. 6 and 7) for a more advanced discussion on the fundamental problem of causal inference and the Neyman-Holland-Rubin theory of causation.

In turn, understanding experiments as case studies creates an opportunity to make use of a long lineage in qualitative methods of dealing with external validity issues.[3] These are methods that qualitative scholars regularly use to address criticism about the limited external validity of their findings, similar to those that experimental researchers face.[4] In particular, I discuss three kinds of qualitative tools that can enhance the external validity of experiments: tools for constituting cases, tools for selecting cases, and tools for setting up scope conditions (also called scope statements).

The role of these tools is recognized in the qualitative literature. First, tools for constituting a case (or cases) establish the membership of cases into broader categories (or sets). These tools determine that cases are instances of a population of cases by referring to either the outcome or (potentially relevant) causal conditions. Second, tools for selecting cases support the formulation of hypotheses or the analysis of evidence about the transportability or generalization of findings from one spatial and temporal context to others. Third, tools for setting up scope statements support establishing conceptual homogeneity or causal homogeneity among a population of cases. Conceptual homogeneity refers to the consistency of measurement across cases, while causal homogeneity helps set causal inferences' theoretical or empirical limits.

My contribution to the symposium describes these tools and discusses their applicability to experiments by considering two general scenarios. On the one hand, researchers might want to perform traditional experimental designs to estimate the average treatment effect (ATE) or a particular feature of the ATE (e.g., statistical significance, sign, and size).[5] On the other hand, researchers might apply these tools when researching moderators or variables (Z) that affect the effect of the treatment (X) on the outcome (Y) (e.g., Baron and Kenny 1986; Kam and Tressler 2017; Deaton and Cartwright 2016).[6] The latter approach focuses on variation in the ATE (i.e., causal heterogeneity) across different contexts and timings.

Current Approaches to External Validity Issues

Distinguishing between different conceptualizations of external validity is essential for understanding the contributions that qualitative methods can make to experimental research. Drawing substantively on Druckman and Kam (2011), I consider five dimensions of external validity. The first dimension (*sample*) is whether conclusions based on the set of units included in the experiment represent an underlying population. Another dimension refers to the *experimental realism* of a given experiment, or "if the situation is involving to the subjects, if they are forced to take it seriously, [and] if it has an impact on them" (Aronson, Brewer, and Carlsmith 1985, 485, quoted in Druckman and Kam 2011, 44). A third dimension (*mundane realism*) refers to "the extent to which events occurring in the research setting are likely to occur in the normal course of the subject's lives, that is, in the 'real world'" (Aronson, Brewer, and Carlsmith 1985, 485, quoted in Druckman and Kam 2011, 44; see also Bell-Martin's and Perez and Tiscornia's contributions to this symposium). Finally, *timing* is a dimension that addresses whether the experimental results are replicable in a different moment, while *spatial context* is another dimension that involves whether conclusions hold when in a different institutional and social context.

Considering these five dimensions, Table 1 presents the current state of the literature on existing empirical strategies for dealing with external validity. Space constraints prevent a detailed discussion of all these strategies, but we should note some salient points. On the one hand, a simple counting of rows would be misleading in determining which dimensions have received the most attention in the literature. For example, some rows mention classic types of experiments (i.e., laboratory, field, population-based, and natural experiments) that are much "denser" than others—these rows represent a whole set of empirical strategies.

[3] In this article, I use qualitative methods to refer to case studies and small-N methods. I use the terms qualitative methods, qualitative tools, and case study methods interchangeably.

[4] Indeed, even though experiments and case studies might apply different conceptions or "gold standards" of causality (Goertz and Mahoney 2012; Beach 2019), they share similar strengths and weaknesses in terms of internal and external validity (e.g., Morton and Williams 2010; Thelen and Mahoney 2015; Deaton and Cartwright 2016; Goertz 2017).

[5] The ATE estimates the causal effect that results from taking the difference between the average outcome in the treatment group, $E(Y_{i,t})$, and the average outcome in the control group, $E(Y_{i,c})$. Mathematically, the formula would be $E(Y_{i,t}) - E(Y_{i,c})$, where $Y_{i,t}$ is the unit exposed to the treatment and $Y_{i,c}$ the unit exposed to the control.

[6] In other bodies of literature, moderator variables are also called background conditions, support factors, or interactive factors. Moderation analyses, then, enable the discovery of what is termed causal heterogeneity, heterogeneous treatment effects, or second-order causal inference.

Table 1. Current Empirical Strategies for External Validity

Dimension	Empirical strategy	Literature (examples)
Sample	Population-based and survey experiments	Mutz (2012)
	Regression-based analysis	Kam and Franzese Jr. (2007)
	Machine learning analysis	Chipman, George, and McCulloch (2010); Imai and Strauss (2011); Green and Kern (2012)
	Ensemble machine learning algorithm	Grimmer, Messing, and Westwood (2017)
	Theoretical-guidance	Druckman and Kam (2011); Kam and Trussler (2017); Coppock and McClellan (2019).
	Weighting	Franco et al. (2017)
	Blocking designs	Moore (2012)
	Replications	Berinsky, Quek, and Sances. (2012), Krupnikov and Levine (2014); Mullinix et al. (2016); Coppock (2019).
Experimental realism	Laboratory experiments	Iyengar (2011)
	Process-tracing applications	Dunning (2012); Seawright (2016)
	Quantitative comparison	Baldassarri and Grossman (2013)
	Hawthorne effect-treatment	Gerber, Green, and Larimer (2008)
	Screeners	Berinsky, Druckman, and Yamamoto (2019)
	Stylization	Dickson, Hafer, and Landa (2008)
Mundane realism	Field experiments	Green and Gerber (2012)
	Natural experiments	Dunning (2012)
	Experimental ethnography	Paluck (2010)
	Deception	Dickson (2011)
	Vignette and conjoint experiments	Louviere et al. (2000); Mutz (2012).
	External validation test	Hainmueller, Hangartner, and Yamamoto (2015)
Timing	Analysis of prior effects	Gaines, Kuklinski, and Quirk (2007); Druckman (2009)
	Analysis of duration effects	Gaines, Kuklinski, and Quirk (2007); Druckman and Nelson (2003); Mutz (2005).
	Moderation analysis	Green and Kerk (2012); Schwarz and Coppock (2020)
	Longitudinal survey experiment	Gaines, Kuklinski, and Quirk (2007)
	Replications	Lawless (2015)
Spatial Context	Multi-context experiments	Dunning et al. (2019); Blair and McClendon (2021)
	Contextual reality change	Arceneaux and Johnson (2008)
	Moderation analysis	See machine learning, ensembles, and moderation analysis above.
	Analytical approach	Martel García and Watchekon (2010).

A second conclusion from this table is that the literature focuses the most on dimensions of external validity such as sample construction and, to a lesser extent, on experimental realism and mundane realism while neglecting other dimensions such as timing and context (Druckman and Kam 2011). Given that the sample dimension is related to statistical concepts such as probability sampling, it may be unsurprising that empirical strategies based on quantitative methods would be the predominant approach for dealing with these external validity issues. Nevertheless, the preeminence of quantitative methods extends beyond this dimension to those where a quantitative approach is not necessarily the most appropriate.

Let us briefly consider the empirical strategies that are currently the most salient for dealing with the timing and context dimensions of external validity: *moderation analyses* and *multi-context experiments*. Moderation analyses focus on "second-order causal inferences problems," (Seawright 2016, 181) and try to identify what was defined above as moderators (variables Z). Moderation analysis serves the purpose of addressing issues of external validity related to timing when moderators are variables related to time (e.g., the effect of the treatment on the outcome varies from one year to another). Similarly, moderation analysis helps address the context dimension when moderators are contextual variables (e.g., the effect of the treatment on the outcome varies from one spatial context to another).

Another related but distinct empirical strategy is conducting and analyzing multi-context experiments. In multi-context experiments, researchers conduct, pool, or analyze results from the same (or an equivalent) experiment in different settings (or in a given location across time). In some cases, these experiments are conducted independently from each other; at other times they are purposely done sequentially, one after another; and more recently, scholars have been conducting similar experiments simultaneously across different countries (Blair and McClendon 2021). A prominent example from the third category is the Metaketa Initiative—Phase 1 from the Evidence in Governance and Politics research network (EGAP 2020), which coordinates researchers conducting the same field experiment across six to seven contexts.

In principle, neither moderation analyses nor multi-context experiments require quantitative methods. But, in practice, scholars apply statistical techniques for both of them.[7] Indeed, moderation analysis involves various methods, from simple regression analysis to

[7] For a prominent exception, see Henrich et al. (2004). The authors combine regression analysis with ethnographic evidence for their moderation analysis.

sophisticated ensembles of machine learning algorithms (Grimmer, Messing, and Westwood 2017).[8] Meanwhile, multi-context experiments usually are analyzed using quantitative methods, including moderation analysis. Nevertheless, these quantitative methods are not always appropriate. As Seawright argues, "the assumptions involved for causal inferences in such [moderation] analysis are rather daunting, and the number of different contexts available for the study is not always large" (2016, 182).

In the same vein, another conclusion from Table 1 would be that applying qualitative methods to external validity issues is currently very limited. Paluck (2010) made a compelling case for combining ethnographic techniques with experiments, arguably as a way of improving studies' level of mundane realism (see also Bell-Martin in this symposium). Seawright (2016) has proposed process-tracing applications for checking the experimental realism of treatment (or as-if-random stimuli in natural experiments). And Dunning (2012) has made contributions along the same lines as Seawright.

To the best of my knowledge, however, there is no clear methodological guidance on how to apply qualitative methods to the timing and context dimensions. The omission of qualitative methods in dealing with those dimensions is remarkable. Arguably, qualitative scholars have little to nothing to contribute to the sample dimension: as mentioned above, the sample dimension follows a statistical understanding of external validity. By contrast, qualitative researchers regularly answer questions concerning the broader application of their findings to different temporal and spatial contexts.

In sum, while the literature tends to neglect the context and timing dimensions of external validity in experimental research, the use of qualitative techniques to address these dimensions is promising.

Multiple Units, One Case

So far, I have made the case that integrating qualitative methods with experimental research would be potentially helpful in dealing with some external validity issues (context and timing). This section argues that applying qualitative tools to experimental studies is justified because experiments can be seen as case studies. Based on this understanding of experiments as case studies, the rest of the article presents ideas on how to integrate qualitative and experimental methods.

Charles Ragin (1992a, 2) argued that it is wrong to fall into the tempting idea of conflating case studies and qualitative methods, stating that "virtually every social scientific study is a case study or can be conceived as a case study." As a heuristic, the author considered "an analysis of individual-level survey data from a sample of adults in the United States," and concluded that it "provides a foundation for statements about individuals and about the United States." Ragin (1992a, 2) mentioned that this study "can be seen both as an extensive analysis of many cases (the sample of individuals) and as an intensive case study of the United States." Rueschemeyer (2003, 318) similarly argued that "a good deal of the extant quantitative research is confined to a single country or a single community," in as much as "good analytical historical work," even though both statistical analysis and historical analysis apply different techniques.

Following a similar logic, I argue that experimental studies are also case studies. More precisely, experiments are composed of at least two types of cases or units of analysis: *randomized units* (e.g., individuals) or the units randomly assigned to the treatment and control groups, and the *experimental case* (e.g., a country), which is not randomized. These different units of analysis exhibit a hierarchical relationship: the randomized units are the lower-level unit of analysis, while the experimental case forms the boundary of the randomized units spatially and temporally. Experimental studies typically contain multiple randomized units but only one experimental case.

In other words, arguing that experiments are case studies is not a statement about how to aggregate up from individuals (or other randomized units in the experiments) to more macro units.[9] Rather, it is a statement about how any single experimental study contains more than one unit of analysis, and how our focus on what I call the experimental case can lead us to characterize the experimental study as a case study.[10]

Let us consider an example. In her contribution to this symposium, Oliveros discusses the amount of contextual knowledge needed to successfully design the series of

[8] Moderation analyses could be considered a form of meta-analysis when scholars pool findings from different studies rather than subgrouping a dataset from a single study.

[9] A question on aggregation might be related to the sample dimension of external validity (i.e., how to generalize from the sample to the population). By contrast, a focus on the experimental cases is related to context and timing: the spatial and temporal context where the randomization occurs and whether we can generalize our findings to different spatial and temporal contexts.

[10] Arguably, the experimental literature already implies the existence of an experimental case when it refers to an underlying population or "the finite set of units for which we observe covariates, treatments, and realized outcomes," as "all conclusions are conditional on this population" (Imbens and Rubin 2015, 20). But, to infer causality or estimate the average treatment effect (ATE), "it does not matter how this population was selected, or where it came from" (Imbens and Rubin 2015, 20). By contrast, reflecting around this population or, more precisely, the experimental case is essential for making sense of the external validity of experimental findings.

list experiments in Argentina. However, her research also illustrates the different units of analysis included in experimental work. Indeed, 1200 low- and mid-level public employees of three Argentinean municipalities (Salta, Santa Fe, and Tigre) participated in the study. Hence, the randomized units are the individuals assigned to the treatment and control groups. A focus on lower-level units of analysis is commonplace in experimental research.

At the same time, in her book, Oliveros (2021, 63) justifies purposely selecting these three "urban" municipalities in a qualitative manner, stating: "while similar in population size, [they] vary greatly in their political and economic characteristics. They therefore provide a good opportunity to test how the theory of self-enforcing patronage travels across different political and economic environments." Thus, the experimental cases are Salta, Santa Fe, and Tigre. But the justification provided above—and, indeed, the book title and several of its passages—makes it possible to argue that Argentina as a whole is another, upper-level experimental case. From this viewpoint, her book is simultaneously a study of multiple public employees, a small-N study of three municipalities, and a case study of Argentina.

In general, defining experimental cases forces researchers to acknowledge that they are implicitly conducting case studies (i.e., small-N and single case studies) while explicitly conducting experimental research.

Experiments and Qualitative Methods

This section presents three sets of qualitative tools and ideas for applying them to external validity issues in experimental research: tools for constituting cases, selecting cases, and making scope conditions. These methods are not focused on data collection, such as interview methods, focus groups, archival research, and ethnography. Nor do these methods focus on intensive, within-case analysis for describing and analyzing causal relationships such as process-tracing and counterfactual analysis. However, these small-N and case study methods are part of the essential toolkit qualitative methodologists use when focused on the external validity of their research.[11] Conceptualizing experiments are case studies makes it possible to integrate these qualitative methods with experimental research.

Constituting Cases

Qualitative scholars are particularly aware that the "same case may be an example of many different things, and hence representative of many different populations" (Elman, Gerring, and Mahoney 2016, 378). Indeed, the research operation of "casing" or concocting cases in varied ways is routine in the social sciences (Ragin 1992b, 217). Qualitative scholars might start their research considering that their cases are instances of a particular population but, after collecting new data and engaging in concept formation, these scholars might turn to consider another population of relevant cases (Collier and Mahoney 1996). In other words, "qualitative research's specification of relevant cases at the start of an investigation is nothing more than a working hypothesis" (Ragin 2004, 125).

Tools for constituting cases are related to the external validity dimensions of context and timing because they help to "eliminate proper names" (Pzewroski and Teune 1970, 30), "as long as 'eliminate' means 'reduce' and not 'eradicate' altogether" (Slater and Ziblatt 2013, 11). When constituting a case as part of the relevant population of cases, researchers think in broader terms than the particular instance in which they collect and analyze evidence. Thus, tools for constituting cases are not tools for achieving internally valid studies but rather, externally valid ones.

In more practical terms, constituting cases means answering the question: "What is this case a case of?" (Ragin 2004, 131; 1992). A common answer could be to circumscribe the case to general, pre-established categories such as country, district, region, state, city, municipality, province, village, squatter, and university, or a particular month, year, or decade. However, qualitative scholars usually emphasize either the outcomes or the (potential) causal conditions (Ragin 2004). In O'Donnell's (1986) seminal book, for example, Argentina (1966-1973, 1976-1983) is *a case of bureaucratic authoritarianism* (when emphasizing the outcome) and *a case of the most modernized countries in Latin America* (when highlighting the causal condition). During these years, Argentina is an instance of both sets of relevant cases: bureaucratic authoritarianism and most modernized countries in Latin America.

We could similarly constitute experimental cases by emphasizing the outcome and the potentially relevant causal conditions. When considering the outcome in an experimental study, I propose focusing on features

[11] As a clarification, these three sets of methods do not support first-order causal relationships (causal effects), but they can help make sense of second-order causal relationships (causal heterogeneity) related to what has been defined above as moderators. I do not focus on how a researcher collects evidence on these second-order causal relationships but on how they can make sense of this evidence, regardless of its origin, based on these three sets of tools. Moreover, the relevance of these tools can also be supporting theory-building and the formulation of hypotheses rather than theory-testing.

of average treatment effect (ATE) such as statistical significance, sign, and magnitude. Following Seawright (2016, 183), scholars might argue that their case is an instance of a "positive- and negative-effect cases, if the variability is so extensive" or "large- and small-effect cases, if the heterogeneity involve the size, rather than the direction, of the causal effect." For instance, Scharwz and Coppock (2022) pooled 67 factorial survey experiments on the effect of a candidate's gender on vote choice. The substantial variability across these studies helps to illustrate how to constitute experimental cases based on the outcome. The authors find 23 cases that are statistically significant with a positive ATE, and 11 that are statistically significant with a negative ATE. In this sense, we could constitute an experimental case like India as an *instance of the set of statistically significant and positive ATE*.

When considering the causal conditions for constituting an experimental case, I propose focusing on contextual and timing moderators, using the definitions provided above. Scharwz and Coppock (2022, 8) "find that the effect of gender is slightly more positive among studies conducted post-2014 whereas it appears to be negative for samples collected before 1998," perhaps due to changes in gender norms over time. Assuming that the evidence on the effect of this periodization is strong, an experimental case like India from a study conducted in 2020 would also be *an instance of post-2014 studies*.

An important caveat here is that constituting experimental cases—either by emphasizing features of the ATE or relevant moderators—is not necessarily restricted to a particular stage in the research process (e.g., after conducting the experiment) as the examples might imply. Instead, tools for constituting cases can support researchers before or after conducting an experiment. The qualitative tools might help formulate hypotheses about features of the ATE *before* conducting an experiment by observing potentially relevant moderators. For instance, consider a researcher who is about to conduct another experiment on the effect of a candidate's gender on vote choice in 2022. Since the experimental case would be an instance of the set of post-2014 studies, she might also expect that the experimental case would be an instance of a statistically significant and positive ATE. From this viewpoint, casing or recognizing the case as part of a broader category (e.g., post-2014) is a prerequisite for the case selection strategies discussed below.

Alternatively, these qualitative tools might support answering questions about the effect of potentially relevant moderators on observed features of the ATE *after* conducting an experiment. In this scenario, researchers already know the experimental findings and focus on answering second-order causal inferences problems. For instance, consider the same researcher conducting an experiment on the effect of a candidate's gender on vote choice in 2022 but finding the counterintuitive result of a statistically significant and negative ATE. After collecting evidence on the condition (i.e., moderator) explaining these unexpected results, she might report her conclusions by constituting the experimental case as a member of a set that makes sense of these features of the ATE. Casing thus becomes a tool for making claims about scope conditions, as explained below.

In sum, constituting experimental cases as instances of broader outcomes (i.e., features of the ATE) or causal conditions (i.e., timing and contextual moderators) supports formulating hypotheses or analyzing the evidence on the external validity of experimental results. At the same time, it also supports the application of the qualitative tools discussed in what follows.[12]

Selecting Cases

Case selection is intrinsically related to how externally valid a study is. Strictly speaking, why a particular context is selected is irrelevant for drawing clear, internally valid causal inferences. By contrast, statements about case selection, especially when combined with tools for constituting cases as discussed above, are crucial for hypothesizing or evaluating the transportability or generalization of findings to different spatial and temporal contexts.

When experimental scholars decide where to conduct their experiments, they are unlikely to rely on random sampling or consider the whole universe of relevant cases. These scholars commonly ponder their case expertise (e.g., American Politics, Latinamericanists) and theoretical interests (e.g., cases to test their hypotheses) to select a single or small number of contexts, even when they coordinate multi-context experiments (Blair and Mclendon 2021). As a result, experimental scholars make decisions that have much in common with the purposive case selection strategies that qualitative scholars follow (Goertz and Mahoney 2012, 185; see also Beach and Pedersen 2013).

Drawing substantively on Koivu and Hinze (2017) and Gerring and Cojocaru (2016), Table 2 lists classic case

[12] For purposes of transparency and replicability, hypotheses derived from these qualitative tools and the ones discussed below can also be pre-registered or incorporated as part of pre-analysis plans, along the lines of Pérez Bentancur and Tiscornia's contribution to this symposium.

selection strategies adapted to experimental research.[13] The table classifies strategies into three case-selection types: (1) characteristics of the experimental case, (2) relationship of one experimental case to another, and (3) relationship to the theoretical or posited relationship between moderators (Z) and features of the ATE.

Table 2. Experimental case selection strategies

Experimental Case-selection Type	Experimental Case Selection Strategy	Criterion	Requirements
Characteristics of the Experimental Case	Extreme on the moderator	Z	1+
	Extreme on the causal effect	ATE	3+
	Substantive significance	Z	1+
Relationship to each other	Least similar	Z, ATE	2+
	Most similar	Z, ATE	2+
Relationship to Theory or Posited Z/ATE Relationship	Most likely	Z, ATE	1+
	Least likely	Z, ATE	1+
	Crucial	Z, ATE	1+
	Deviant	ATE	1+
	Typological	Z, ATE	2+
	Typical	ATE	1+

Note: (Z) refers to moderators and ATE to average treatment effect. Numbers and the (+) sign refer to the minimal (without maximum) number of experimental cases needed for a given strategy.

While there is ample literature that expands on each of these strategies, the table also summarizes requirements regarding the minimum number of experimental cases needed as well as the selection criterion (either Z or features of the ATE). These criteria, in particular, resemble the decisions that qualitative researchers make regarding case selection: (i) cases are selected because of their observed outcomes (e.g., cases exhibiting the presence of a revolution); (ii) researchers ignore outcomes, and select cases based on how they score on the independent variables (e.g., cases exhibiting different levels of state capacity where the outcome is an unknown level of law enforcement); and (iii) cases are selected because of their scores on both the dependent and independent variables (e.g., cases of party survival to national-electoral crisis due to the presence of resources to remain competitive in the subnational arena).[14]

In traditional experimental research, scholars conduct studies where they ignore the experimental findings but have a hypothesis based on potentially relevant moderators (Z). When the selection criterion is Z, experimenters resemble situation (ii) in qualitative research. For example, another finding from Schwarz and Coppock's (2022) meta-analysis is that South American cases show a positive ATE: the average effect of a candidate described as a woman results in percentage point gains in vote margin. Thus, one could reasonably select an unstudied case like Peru (i.e., *another instance in the set of South American countries*) under the hypothesis that the ATE should also be positive.

When experimental findings are known, case selection is still helpful. As Seawright (2016) argues, experimental findings (i.e., features of the ATE) could be the outcomes to be explained in multi-method research designs. Thus, case selection can support formulating or evaluating hypotheses of second-order causal relationships (i.e., moderators) that are intrinsically related to external validity issues of context and timing:

[13] Before adopting these qualitative tools for selecting cases, experimenters should remember that the vast qualitative literature on the topic includes overlaps, contradictions, and controversies in existing recommendations (Fairfield and Charman 2019).

[14] The last example on party survival is partially inspired by Cyr (2017). In a review symposium discussing Goertz (2017), however, Cyr argues that, based on her experiences from previous work (e.g., Cyr 2017), selecting cases based on both the primary cause (X) and the outcome (Y) may be difficult (Waldner et al. 2019, 162).

Why does a particular case (or set of cases) exhibit a specific feature of the ATE (and others do not)?

Once again, Scharwz and Coppock (2022) provide relevant examples. As in situation (i) above, experimental researchers could select cases because of features of the ATE (e.g., Afghanistan, Jordan, or Tunisia are cases of a negative effect of a candidate's description as a woman). Furthermore, as in situation (iii) above, experimental researchers could select cases using potentially relevant moderators and the ATE. For example, imagine we conduct the vote choice experiment in another South American case like Peru, and the results are a negative effect of a candidate's description as a woman, contradicting previous results in the region (Scharwz and Coppock 2022). In this scenario, selecting Peru would use both selection criteria (Z and ATE). Rather than understanding this case selection as a justification after the fact or a counterintuitive selection after conducting the analysis, it is essential to highlight that its purpose would be starting a different analysis focused on second-order causal relationships (i.e., the conditions explaining these unexpected results).[15]

Scope Conditions

Qualitative scholars typically ask questions about the causes-of-effect (e.g., why are women underrepresented in political institutions?) rather than the effects-of-causes (e.g., what is the effect of candidates' gender on vote choice?). Thus, they embrace a view of the social world that is causally complex and:

> ...characterized by path dependence, tipping points, interaction effects, strategic interaction, two-directional causality or feedback loops, and equifinality (many different paths to the same outcome) or multifinality (many different outcomes from the same value of an independent variable, depending on context). (Bennett and Elman 2006, 456)

As a result of this causal complexity, scholars tend to find it desirable to restrict the set of cases included in the analysis for parsimony (e.g., Skocpol's (1979) theory on social revolutions limited to non-colonial states). In more formal terms, Goertz and Mahoney (2009, 307) discuss scope statements as "intimately related to generalization" because these statements "set empirical and theoretical limits on the extent to which an inference can be generalized." Thus, scope statements help determine inferences' spatial and temporal boundaries and determine the findings' external validity. As Findley et al. (2021, 369) put it, "the identification of scope conditions is perhaps the most common approach for making external validity inferences."

Two main tools for setting scope conditions that can transfer to experimental research are *conceptual homogeneity* and *causal homogeneity* (Goertz and Mahoney 2009, 312). Conceptual homogeneity refers to the existence of *measurement stability* (i.e., "the same score means the same across all cases") or *substitutability* (i.e., different dimensions or indicators have a "functional equivalence"). Its relevance for experimental case studies is evident when we consider that multi-context experiments and moderation analyses need to assure comparability across experimental cases.[16]

For instance, Scharwz and Coppock (2022, 5) collect studies based on two criteria: "(1) candidate gender is randomized, and (2) the dependent variable is, or can be transformed into, a binary vote choice for or against the candidate." Moreover, in terms of conceptual homogeneity, the authors "did not exclude studies based on the manner in which candidate gender was signaled to the survey respondent" (e.g., woman, man, male, female, text, and pictures). Scharwz and Coppock are making a statement of measurement substitutability that permits comparison across all 67 studies instead of setting a scope restriction that incorporates some experimental cases.[17]

Meanwhile, causal homogeneity refers to imposing scope restrictions to reduce, rather than eliminate, causal heterogeneity. For example, as mentioned above, Scharwz and Coppock (2022) find that the effect of gender is slightly more positive in post-2014 studies, perhaps as a consequence of changes in gender norms. If another researcher claims that gender norms have changed the most in Western countries, a possible scope statement would be that Western countries have more positive effects of gender than non-Western countries *when considering the post-2014 era*. This scope statement could be either a hypothesis (if there is no evidence to evaluate it yet) or support a causal analysis (if there is evidence in favor of it). The critical point here is that this statement considers variability (Western versus non-Western) within the causally homogenous cases in the

[15] Let us notice that this selection for purposes of a moderation analysis, after knowing the experimental results, resembles qualitative research when cases are selected on the dependent variable, which is a common strategy because of the salience of specific outcomes (e.g., Goertz and Mahoney 2012).

[16] Conceptual homogeneity is crucial for multi-context experiments and moderation analysis that need to assure comparability across experimental cases. The same applies to meta-analyses that do not perform moderation analyses but seek to find an overall causal effect across several studies and replications evaluating the reliability of findings.

[17] The same applies to meta-analyses that do not perform moderation analyses but seek to find an overall causal effect across several studies and replications evaluating the reliability of findings.

post-2014 era, which would be a moderator timing.

Conclusion

The majority of the literature on experimental political science concentrates on providing clear, internally valid causal inferences. However, the newest generation of experimental research increasingly focuses on developing empirical strategies for dealing with external validity issues. The present essay argues that these empirical strategies can use classic qualitative tools for constituting cases, selecting cases, and setting up scope statements. The transferability of these tools to experimental research is possible because, from a specific viewpoint, experiments *are* case studies.

References

Arceneaux, Kevin, and Martin Johnson. 2008. "TV Channel Changing: Choice, Attention, and Retention in Political Communication Research." Paper presented at the Experiments in Political Science Conference of the University of California, Riverside, CA. http://dx.doi.org/10.2139/ssrn.1301769

Aronson, Elliot, Marilynn B. Brewer, and J. Merill Carlsmith. 1985. "Experimentation in Social Psychology." In *Handbook of Social Psychology: 3rd edition*, edited by Gardner Lindzey and Elliot Aronson, 441–86. New York: Random House.

Baldassarri Delia, and Guy Grossman. 2013. "The Effect of Group Attachment and Social Position on Prosocial Behavior. Evidence from Lab-in-the-Field Experiments*." PLoS ONE* 8, no. 3 (March): 1-9. https://doi.org/10.1371/journal.pone.0058750

Baron, Reuben M., and David A. Kenny. 1986. "Moderator-Mediator Variables Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51, no. 6 (December): 1173–82. https://doi.org/10.1037/0022-3514.51.6.1173

Beach, Derek. 2019. "Multi-Method Research in the Social Sciences. A Review of Different Frameworks and a Way Forward." *Government and Opposition*, 55, no.1 (February): 163-182. doi:10.1017/gov.2018.53

Bennett, Andrew, and Colin Elman. 2006. "Qualitative Research: Recent Developments in Case Study Methods." *Annual Review of Political Science* 9, no. 1 (June): 455–76. doi:10.1146/annurev.polisci.8.082103.104918.

Berinsky, Adam J., Kai Quek, and Michael Sances. 2012. "Conducting Online Experiments on Mechanical Turks," *The Experimental Political Scientist* 3, no. 1 (Spring): 2-6.

Berinsky, Adam J., James N. Druckman, and Teppei Yamamoto. 2019. "Publication Biases in Replication Studies," *Social Science Research Network* Last modified September 16 2019. http://dx.doi.org/10.2139/ssrn.3454901

Blair, Graeme, and Gwyneth McClendon. 2021."Conducting Experiments in Multiple Contexts." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Greeen, 411-28. Cambridge: Cambridge University Press.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *The Annals of Applied Statistics* 4, no.1 (March): 266-98. https://doi.org/10.1214/09-AOAS285

Collier, David, and James Mahoney. 1996. "Insights and Pitfalls: Selection Bias in Qualitative Research." *World Politics* 49, no. 1 (October): 56–91. doi:10.1353/wp.1996.0023.

Coppock, Alexander. 2019. "Generalizing From Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods* 7, no. 3 (July): 613-28. https://doi.org/10.1017/psrm.2018.10.

Coppock, Alexander, and Oliver A. McClellan. 2019. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Research & Politics* 6, no. 1 (January): 1-14. https://doi.org/10.1177/2053168018822174.

Cyr, Jennifer. 2017. *The Fates of Political Parties: Institutional Crisis, Continuity, and Change in Latin America*. Cambridge: Cambridge University Press.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48, no. 2 (June): 424–55. https://doi:10.1257/jel.48.2.424.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." National Bureau of Economic Research Working Paper Series, 22595, Cambridge, MA. http://dx.doi.org/10.3386/w22595.

Dickson, Eric S. 2011. "Economic vs. Psychology Experiments: Stylization, Incentives, and Deception." In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, 102-24. Cambridge: Cambridge University Press.

Dickson, Eric S., Catherine Hafer, and Dimitri Landa. 2008. "Cognitive and Strategy: A Deliberation Experiment." *The Journal of Politics* 70, no.4 (October): 974-89. https://doi.org/10.1017/S0022381608081000.

Druckman, James N., and Donald P. Green. 2021. *Advances in Experimental Political Science*. Cambridge: Cambridge University Press.

Druckman, James N., and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base." In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. Cambridge: Cambridge University Press.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences.* Cambridge: Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis, eds. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge: Cambridge University Press.

EGAP. 2020. "The Metaketa Initiative." *Evidence in Governance and Politics*. https://egap.org/our-work-0/the-metaketa-initiative/.

Elman, Colin, John Gerring, and James Mahoney. 2016. "Case Study Research: Putting the Quant Into the Qual." *Sociological Methods & Research* 45, no. 3 (April): 375–91.

Fairfield, Tasha, and Andrew Charman. 2019. "A Bayesian Perspective on Case Selection." (working paper, adaptation from Chapter 11 in the forthcoming *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*, Cambridge University Press) https://jsis.washington.edu/wordpress/wp-content/uploads/2019/10/Fairfield_Charman_Case_Selection_2019.pdf.

Fukuyama, Francis. 2013. "Albert O. Hirschman, 1915–2012." *The American Interest* 2 (March/April): 93–95.

Franco, Annie, Neil Malhotra, Gabor Simonovits, and L.J. Zigerell. 2017. "Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments," *Journal of Experimental Political Science* 4, no. 2 (October), 161–72. doi:10.1017/XPS.2017.2.

Gaines, Brian J., James H. Kuklinksi, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15, no. 1 (January): 1–20. doi:10.1093/pan/mpl008.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102, no.1 (February), 33-48. https://doi.org/10.1017/S000305540808009X.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: WW Norton & Company.

Gerring, John, and Lee Cojocaru. 2016. "Selecting Cases for Intensive Analysis." *Sociological Methods & Research* 45, no. 3 (February): 392–423. doi:10.1177/0049124116631692.

Goertz, Gary. 2017. *Multimethod Research, Causal Mechanisms, and Case Studies: An Integrated Approach*. Princeton: Princeton University Press.

Goertz, Gary, and James Mahoney. 2009. "Scope in Case Study Research," In *The SAGE Handbook of Case-Based Methods*, edited by David Byrne and Charles C. Ragin, 307-17. Thousand Oaks: SAGE Publications

———. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. New Jersey: Princeton University Press.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76, no. 3, 491–511.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25, no. 4 (September): 413–34, https://doi:10.1017/pan.2017.15.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against the Real World." *The Proceedings of The National Academy of Sciences* 112, no. 8 (February), 2395-400. https://doi.org/10.1073/pnas.1416587112 .

Hall, Peter, A. 2016. "Politics as a Process Structured in Space and Time," In *The Oxford Handbook of Historical Institutionalism*, edited by Orfeo Fioretos, Julia Lynch, and Adam Steinhouse, 31-50. New York: Oxford University Press.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gontis, eds. 2004. *Foundations of Human Sociality*. New York: Oxford University Press. http://dx.doi.org/10.1093/0199262055.001.0001.

Imai, Kosuke, and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign." *Political Analysis* 19, no.1 (Winter): 1–19. https://doi.org/10.1093/pan/mpq035

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Iyegar, Shanto. "Laboratory Experiments in Political Science." In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. Cambridge: Cambridge University Press.

Kam, Cindy D. and Marc J. Trussler. 2017. "At the Nexus of Observational and Experimental Research*," Political Behavior* 39, no. 4 (December): 789–815. https://doi.org/10.1007/s11109-016-9379-z.

Kam, Cindy D., and Robert Franseze Jr. 2007. *Modeling and Interpreting Interactive Hypothesis in Regression Analysis*. Michigan: University of Michigan Press.

Koivu, Kendra L., and Annika Marlen Hinze. 2017. "Cases of Convenience? The Divergence of Theory from Practice in Case Selection in Qualitative and Mixed-Methods Research." *PS: Political Science & Politics* 50, no. 04 (October): 1023–27. doi:10.1017/s1049096517001214.

Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparison and External Validity." *Journal of Experimental Political Science* 1, no. 1 (Spring): 59-80. doi:10.1017/xps.2014.7.

Lawless, Jennifer. 2015. "Female Candidates and Legislators." The Annual Review of Political Science 18, no.1 (May): 349-69. https://doi.org/10.1146/annurev-polisci-020614-094613

Martel García, Fernando, and Leonard Wantchekon. 2010. "Theory, External Validity, and Experimental Inference: Some Conjectures." *The ANNALS of the American Academy of Political and Social Science* 628, no.1 (January): 132-47. https://doi.org/10.1177/0002716209351519

McDermott, Rose. 2002. "Experimental Methodology in Political Science" *Political Analysis* 10, no. 4, (Autumn): 325-42.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge: Cambridge University Press.

Moore, Ryan T. 2012. "Multivariate Continuous Blocking to Improve Political Science." *Political Analysis* 20, no. 4 (September): 460-79. https://doi.org/10.1093/pan/mps025

Mullinix, Kevin, J., Thomas J. Leeper, James N. Druckman, and Jeremey Freese. 2016. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2, no. 2 (January): 109-38. https://doi.org/10.1017/XPS.2015.19

Mutz, Diana. 2012. *Population-Based Survey Experiments*. Princeton: Princeton University Press.

Paluck, Elizabeth Levy. 2010. "The Promising Integration of Qualitative Methods and Field Experiments." *The Annals of The American Academy of Political and Social Science* 628, no. 1 (February): 59-71. doi: https://doi.org/10.1177/0002716209351510.

Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York:Wiley-Interscience.

Ragin, Charles C. 1992a. "Introduction: Cases of 'What is a Case?'," In *What Is a Case?: Exploring the Foundations of Social Inquiry*, edited by Charles C. Ragin and Howard Saul Becker, 1-18. Cambridge: Cambridge University Press.

———. 1992b. "'Casing' and the Process of Social Inquiry," In *What Is a Case?: Exploring the Foundations of Social Inquiry*, edited by Charles C. Ragin and Howard Saul Becker, 217-26. Cambridge: Cambridge University Press.

———. 2004. "Turning the Tables: How Case-Oriented Research Challenge Variable-Oriented Research." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 1st edition, edited by Henry E. Brady and David Collier, 123-39. Lanham: Rowman & Littlefield Pub Incorporated.

Rueschemeyer, Dietrich. 2003. "Can One or a Few Cases Yield Theoretical Gains?" In *Comparative Historical Analysis in the Social Sciences*, edited by James Mahoney and Dietrich Rueschemeyer, 305-36. Cambridge: Cambridge University Press.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

Shadish, William, R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Publishing Company.

Skocpol, Theda. 1979. *States and Social Revolutions*. Cambridge: Cambridge University Press.

Slater, Dan, and Daniel Ziblatt. 2013. "The Enduring Indispensability of the Controlled Comparison." *Comparative Political Studies* 46, no. 10 (January): 1301–27. doi:10.1177/0010414012472469.

Scharwz, Susanne, and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments?" *The Journal of Politics* 84, no. 2 (April): 1-14. https://doi.org/10.1086/716290.

Thelen, Kathleen, and James Mahoney. 2015. "Comparative-Historical Analysis in Contemporary Political Science," In *Advances in Comparative-Historical Analysis*, edited by James Mahoney and Kathleen Thelen, 3-36. Cambridge: Cambridge University Press.

Waldner, David, Jennifer Cyr, Kendra Koivu, and Gary Goertz. "Review Symposium: Multimethod Research, Causal Mechanisms, and Case Studies." *European Political Science* 18, no. 1 (March): 157–69, doi:10.1057/s41304-017-0147-2.

QMMR Qualitative & Multi-Method Research

Original Article

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448133

Data Security in Human Subjects Research: New Tools for Qualitative and Mixed-Methods Scholars [1]*

Aidan Milliff [2]†
Massachusetts Institute of Technology

Political science research in both qualitative and quantitative traditions frequently uses data that contain personal information about research participants. Personal information can enter the research process in different ways; sometimes researchers collect it directly via a survey or an interview, other times they gather it from an aggregator like a government agency or private company or semi-public sources like social media. In many cases, the personal data that political scientists collect is both *personally-identifiable*[3] and *sensitive*, meaning that disclosure could expose respondents to severe repercussions like legal sanction (McMurtrie 2014) or retribution from non-state actors (Venkatesh 2008), as well as more diffuse harms like the negative impacts on personal life, employment opportunities, or reputation (Ohm 2010).

Scholars who use sensitive and personally-identifiable information (PII) in their research may struggle to balance two objectives which are in tension with one another: to keep sensitive data confidential to protect the privacy of human subjects,[4] but also conduct research that meets the method-specific standards of transparency as expected by the political science profession. Researchers often promise interviewees, study participants, or ethnography subjects that the information they share will be confidential unless they explicitly consent to being identified.[5] At the same time, professional bodies like the Qualitative Transparency Deliberations (Jacobs et al. 2021) and the APSA Ad Hoc Committee on Human Subjects Research (2020) call for researchers to provide at least parts of the underlying evidentiary record while still respecting privacy and maintaining confidentiality of sensitive, identifiable information. Some researchers may therefore perceive professional incentives to a) share data as much as possible, and b) maintain copies of *all data* indefinitely.[6]

While there is increasing clarity about the normative *standards* for privacy protection and qualitative transparency that political scientists should seek to uphold, the process of meeting those standards in practice remains largely *ad hoc*, and up to the discretion of individual researchers. To maintain data security in practice (i.e., to protect sensitive,

[1] * Thank you to Andy Halterman, Minh Trinh, Lily Tsai, Lukas Wolters, and the referees and editors at QMMR for feedback. Thank you also to Molly Roberts, Jesse Driscoll, and participants in the "Multi-Method Tools for Data Security in Political Science Research" panel at APSA 2020.

[2] † Ph.D. Candidate, Massachusetts Institute of Technology.

[3] Personally identifiable here means that the data contain sufficient information to reasonably infer the identity of the individual who the data represents, directly or indirectly (McCallister, Grance, and Scarfone 2010).

[4] This essay follows the common rule definitions of privacy and confidentiality, in which privacy refers to a research participant's desire (and right) to control what other people know about him or her, and confidentiality refers to the way that researchers (promise to) handle participants' data, typically focused on protecting their privacy.

[5] This promise is frequently part of the consent forms required by Institutional Review Board (IRB) processes (Fujii 2012; Zechmeister 2015), and is probably only omitted in specific circumstances like elite interviews. Even when using pre-existing data that contains PII (King and Persily 2019), there is a growing consensus that researchers are obligated to guard "public" data as if they had secured informed consent and collected it themselves (Gibney 2017; Shilton and Sayles 2016).

[6] The new APSA guidelines suggest that political scientists facing pressure to prioritize transparency in a way that harms research participants should contact the APSA Committee on Professional Ethics, Rights, and Freedoms.

identifiable data from misuse, disclosure, or reverse-engineering) researchers need to address a range of threats that accrue when sensitive, personally- identifiable data are collected and stored, and when de-identified data are shared. Although threats to data security (and viable solutions) vary widely depending on the research context and methods used, this article attempts to provide practical advice for designing data security protocols that meet reasonable standards for privacy protection and qualitative transparency.

I focus primarily on one common threat to data security and respondent privacy—the re-identification of participants—that can occur in both qualitative and quantitative human subjects research and is a threat across the lifespan of a research project. Re-identification can occur when adversaries are able to reverse-engineer the identity of research participants from sources that have nominally been de-identified or stripped of personal information. In the second section, I describe how the threat of re-identification arises in political science research and I describe general characteristics of good practical solutions to manage re-identification threats while respecting the importance of qualitative transparency. In the third section, I introduce a complication that is also widespread in political science research: re-identification threats increase and become harder to manage for research projects that involve partners like civil society organizations, community groups, research assistants, or translators. Finally, in the fourth section, I turn to solutions.

I propose some practical tools for managing the threat of re-identification in qualitative and multi-method data, including two novel practices that rely on open-source, easy to use tools. I conclude by situating these tools in the broader, evolving landscape of threats to data security in political science research.

Re-Identification and other Threats to Data Security

Social scientists who collect and analyze sensitive data face a wide range of threats to the confidentiality of participant data. These threats are important to consider at all stages of a research project; according to recently revised ethics guidelines from APSA, ensuring participant privacy and safety is the obligation of each individual researcher (APSA Ad Hoc Committee on Human Subjects Research, 2020). In this section, I briefly describe three of the many possible threats to data security: theft, expropriation, and re-identification. I then focus more specifically on re-identification for two reasons. First, re-identification is a threat that can be especially sensitive to the way researchers try to balance data security and transparency goals. Second, strategies to guard against re-identification are likely more generalizable than strategies to guard against theft and expropriation, which depend heavily on research context and legal jurisdiction.

One of the threats to data security is the possibility that data might be stolen. Theft can occur at any point between when data are collected and destroyed. Why should political scientists worry about theft? Theft of personal data from academic institutions is already common, but so far has targeted student records, not research data (see e.g., Identity Theft Resource Center 2017). Research data may become a target in the future, as social scientists use (and store) larger and more sensitive administrative data sets. The threat of theft might also increase in collaborative projects, where co-authors store PII on a network or frequently send it back and forth (Summers 2016).

Another threat to data security arises if researchers are forced, by law or otherwise, to give up data they have collected. This possibility, expropriation, threatens any data that researchers possess. Actors with bad intentions might also try to get data through coercion. Researchers are sometimes monitored by security services while collecting sensitive data (Wood 2009) or in rare instances, closely followed or questioned (Menoret 2014). United States citizens conducting research abroad might be able to leave without risk of extradition, but leaving generally protects a researcher's physical integrity, not the data they have collected.[7] Legal threats to data security are often overlooked, but researchers in the United States, for example, can be obliged to comply when American courts demand sensitive, identifiable data (Knerr 1982; Traynor 1996). In one extreme situation in 1993, a sociology graduate student who refused to testify against former research participants suspected of vandalism was held in contempt of court and jailed (Scarce 2005). Bringing data across international borders is hardly an ironclad solution. In 2011, tapes from an oral history of the Irish Republican Army held by researchers at Boston College were subpoenaed under a provision in a mutual legal assistance treaty between the US and the United Kingdom; these tapes were then used to implicate the research participants in a murder investigation (McMurtrie 2014; Radden Keefe 2018).

A third threat to data security—the re-identification or reverse-engineering of personal information from nominally anonymous data—is more amorphous than the

[7] Leaving also does too little to protect local colleagues.

first two.[8] Re-identification is a risk that varies depending on data sharing practices. Linking data to respondents can be surprisingly easy in both qualitative and quantitative data, even if PII are removed before sharing. Though the examples below describe re-identification in quantitative data, the same logic applies to descriptions of interview subjects or ethnographic interlocutors: providing context can sometimes positively identify an individual.

Re-identification can occur when unique combinations of attributes are matched to publicly available references, or when contextual knowledge allows an adversary to recognize an individual in the data. Sparse data structures are less anonymous than researchers expect. As of 2000, 87% of US residents are uniquely identifiable by three attributes which would be easy to match with public records: ZIP code, gender, and birth date (Sweeney 2000).

Re-identification doesn't just rely on demographic variables. In a study of Netflix user data, computer scientists found that small amounts of "background knowledge" about a respondent's movie tastes was sufficient to identify their anonymized account (Narayanan and Shmatikov 2008, 2). IMDB accounts (social media accounts) with as few as 5-10 movie ratings could be reliably linked to Netflix accounts because aside from a few popular movies, a watch-list is a surprisingly individual trait (Narayanan and Shmatikov 2008). Adversaries can also use broad contextual knowledge to identify anonymous respondents. Academic publications often try to describe the research setting without identifying it.[9] While important for assessing generalizability of results, these details can also be used to identify the data collection setting, increasing the risk of de-anonymization. Knowing the data collection setting aids de-anonymization. Unique records with respect to age or occupation become more identifiable if the data are known to come from a particular city, school, or company.

Re-identification is the most nuanced threat to data security because it often depends on the extent to which researchers share their data, either in publications, as replication material, or even with their research partners. Some of the techniques commonly used to protect respondent privacywhen sharing these data are not always adequate protection against motivated adversaries.

Data Security with Research Partners

Researchers often work with partners and collaborators—people who are not themselves academic researchers but aid in collection of data either for employment or for mutual interest/benefit. Though some researchers work "solo" or collaborate only with other academics, a substantial number of scholars work with partners, especially to do field research (Kapizewski, MacLean, and Read 2015). Working with partners including NGOs, governments, companies, research assistants, translators, and enumerators or guides change the presentation of all three data security threats.

Theft may be easier if partners' computing and data storage systems are more vulnerable than university systems. Even many highly capable partner organizations (never mind individuals) may have poor digital hygiene/information security practices, making data that passes through their net- work more vulnerable to theft. Negotiating changes to information security practices or avoiding poorly secured networks all together, may be a difficult addendum to research agreements.

Partners may increase a project's vulnerability to expropriation if they need to maintain good relationships with governments where they work. Unlike researchers who may enjoy the freedom to "go home" from a research site, research partners could be subject to coercive pressure from government or, for organizations, their own funders. This exposure puts any data held by the partner at risk and may leave researchers with little leverage to fulfill their data security obligations.

Perhaps most importantly, partners are likely to be experts in the research context and thus particularly well-suited to identify individuals represented in the data that researchers collect.[10] This can complicate efforts to keep data anonymous. NGOs, governments, companies, and individuals are often valuable research partners *because* of their contextual knowledge, but the more they know about the context and the population being studied, the more points of external leverage they must re-identify individuals in de-identified records, quotations, or notes. When respondents share sensitive information with researchers, they may not want that information shared with a partner organization or members of the project team who reside locally. One common academic partnership arrangement, for example, is program evaluation (qualitative or quantitative) for a partner that serves the population that a researcher aims to study. If partners re-

[8] Re-identification technically refers to discovering respondent identity in data from which PII has been stripped. De-anonymization refers to inferring respondent identity even though the data never contained PII. I treat them together because, as I describe below, various examples have shown that people can be identified from data that are thought to be *anonymous*, not just de-identified.

[9] See, for example, the Facebook data from Lewis et al. (2008), which is no longer available because it was partially de-anonymized (Zimmer, 2008).

[10] I assume here that sensitive information needs to be protected against improper use by the partner, as well as by third parties.

identify data including negative attitudes or experiences related to the services, the consequences could be bad for respondents if local partners have leverage to retaliate against them. If, for example, a respondent admits to criminal activity and their response is re-identified by the research partner, the information could be used to deny the respondent benefits. In a real example from qualitative sociology research, disclosing data on informal economic activity to a gang "research partner" active in Chicago public housing allowed the gang to extract unpaid "taxes" from the respondents (Venkatesh 2008).

Preventing Re-Identification: Ideas for Improvement

This section introduces tools that might help scholars address the risk of re-identification, and the special risks that come from working with research partners.[11] The tools recommended here are not exhaustive, not necessarily appropriate for all research contexts, not "silver bullet" solutions, nor representative of the cutting edge in security research. Instead, they are meant to be *feasible* for most researchers. Data security practices only work when implemented, so I focus on measures that are inexpensive, non-time-consuming, and technically simple.

Data Minimization as a General Best Practice

The best way to protect respondent privacy is to *not collect sensitive information or the PII necessary to link it to individuals*. Variables like age, race, and location of residence affect many social science outcomes and must be measured. But many researchers, both in quantitative and qualitative research, feel pressure to measure everything possible, whether to respond to hypothetical reviewers or to "make something" from costly-to-collect data even when main hypotheses are unsupported.

A spartan impulse during research design addresses many key data security threats: data that are never recorded cannot be stolen, expropriated, or accidentally released.[12] "Data minimization," or "privacy by design" entails collecting the minimum amount (and minimum granularity) of both sensitive information and potentially identifying information necessary to test hypotheses plus the most likely alternative explanations. Though the specifics of data minimization would vary across projects, the general intuition should be widely applicable. A researcher designing an interview guide might ask themselves, for example: Can I articulate an analysis for which I will need this information? before asking respondents for personally-identifying information like their ZIP code, exact address, or date of birth.[13] For information that is unlikely to be included in the final analysis or write-up (i.e., where the researcher is more likely to list city or neighborhood than home address when quoting an interview subject), I argue that researchers would often do well to shed a "just in case" attitude about collecting additional information.

Data minimization comes with both benefits and costs. The most important benefit, I argue, is the potential to reduce risk to research participants. Even if other steps are taken to reduce the chance of data security failures like theft and expropriation, limiting the collection of sensitive or personally identifying data might mitigate some harm to participants if theft or expropriation were to happen. A second, smaller benefit accrues to the researcher: data that contain less sensitive or identifying information are easier to handle safely and easier to prepare for sharing.

There are several important costs associated with data minimization, though. For one, data minimization reduces a researcher's freedom to conduct exploratory analyses or find things the researcher was not expecting. If minimization makes the utility of a given data collection effort more narrow, one could say it means that researchers are spending participants' time less efficiently, which is not ideal.[14] Second and relatedly, data minimization reduces the re-usability of data. Conducting data collection is time and resource intensive, so many researchers try to use a single set of interviews, a single ethnographic site, or a single survey to produce multiple works. Data minimization might decrease the possibility of serendipitous spin-offs. Third, there might be professional costs to data minimization because having less information limits the researcher's ability to respond to comments or conduct additional analyses. The severity of this downside in practice likely depends on early adoption by more senior researchers, and integration of data minimization into already accepted norms like pre-registration.

With these costs and benefits in mind, when can researchers pursue a data minimization strategy? Three characteristics seem important for it to be feasible.

[11] Though the other threats discussed above—theft and expropriation—are also important, the ways to address them are much less generalizable because they vary so much with political and legal context.

[12] Un-recorded data can still be inferred by context experts, however.

[13] The intuition may be different in the special case of elite interviews, where potentially identifying information like specific job title might be a necessary part of the published analysis. In this special case, I would argue it is important to treat interviews as essentially "on the record," and affirmatively seek participants' consent to reprint identifiable quotes.

[14] This effect would hopefully be limited if data minimization decreases the length of participation by cutting questions/topics.

First, to accrue the harm mitigation benefits of data minimization, the data collection project needs to be more-or-less single purpose. If a single set of interviews (or an omnibus survey) seeks to test multiple theories about different phenomena, then "minimizing" with respect to those multiple objectives will not necessarily reduce the collection of sensitive information. Researchers who need to collect a wide range of information from the same participants may need to adopt other strategies for data security. Second, data minimization is probably only feasible for deductive, hypothesis-testing data collection. Adopting a data-minimization mindset for exploratory or inductive fieldwork (likely including a lot of critical and interpretive research) could impinge on a researcher's ability to find things they are not expecting. Third, data minimization will not be useful for projects where sharing identifying information like job title (with permission!) is important for establishing the credibility of the speaker. Minimizing other collection will not pay dividends for scholars conducting "on the record" elite interviews, for instance. Where the limitations of data minimization are tolerable, though, I argue it should be attractive to researchers because of its simplicity and relatively strong guarantees of success.

Preventing Re-Identification

Beyond data minimization, several methods are available to guard against re-identification specifically. Preventing re-identification is typically a priority when data are shared (in a manuscript or other public product), but as I discuss in a subsequent section, researchers can also take steps to prevent partners from re-identifying or misusing sensitive data before public release. I describe two techniques for preventing re-identification here: statistical disclosure control/*k*-anonymity and topic modeling for privacy protection.

Statistical Disclosure Control and *k*-anonymity:

Statistical Disclosure Control (SDC) and *k*-anonymity are concepts that come from the quantitative data security literature, but I argue that their shared, underlying intuition is also extremely useful for scholars analyzing, presenting, or sharing qualitative data. The idea behind *k*-anonymity, as proposed by Samarati and Sweeney (1998), is to modify data such that no value of any identifying attribute in the data is shared by fewer than *k* records (see also Sweeney 2002). If no individual value for "age" appears for fewer than three records, the dataset has 3-anonymity for age. This principal is more commonly implemented with respect to "quasi-identifier tuples," or combinations of attributes that could collectively lead to identification—for example, age-gender-ZIP code. K-anonymity is manufactured by suppressing values of identifiable attributes, or by generalizing values (i.e., converting birth years to birth decades).

K-anonymization has drawbacks. First, adversaries can still learn about individuals they know to exist *somewhere* in a dataset. Adversaries trying to learn the HIV status of "Steve"—male, age 35, ZIP Code 60637, known survey respondent—can look at HIV status for all records that match Steve's quasi-identifier tuple and infer the probability that Steve is HIV positive. Recent improvements at least make this risk easier to measure.[15] Second, *k*-anonymization is hard to implement in high-dimensional data, where the unicity of quasi-identifier tuples is remarkably high (de Montjoye et al. 2013). Finally, *k*-anonymization can change the distributional characteristics of data (Angiuli, Blitzstein, and Waldo 2015). K-anonymity is an attractive solution, though, because it is intuitive, relatively easy to implement, and widely used. A related tool, part of the broader research area around Statistical Disclosure Control (SDC), focuses on aggregation, limiting both the geographic and quantitative resolution at which data are reported. Like *k*-anonymity, aggregation eliminates unique records in data. This increases security at the cost of analytical value or informativeness. Aggregation necessarily obliterates high-leverage observations which may be major drivers of the results of statistical analysis.

How can the intuition behind these tools be applied to qualitative research? The intuition and the actual tools behind *k*-anonymity and statistical disclosure control can be a helpful rubric for deciding how to report the demographic identity of interlocutors in a variety of types of qualitative analysis, especially interviews and ethnography. Using tools demonstrated in the online appendix, scholars can empirically measure the relative identification risk of describing an interview participant as "female, age 45, from XYZ village" against the risk of describing that same participant as "female, in her 40s, from ABC district." Researchers trying to weigh the costs and benefits of providing more specificity in descriptions of the people they quote can simply make a spreadsheet containing the demographics they want to describe and then apply tools to measure and increase *k*-anonymity to find a privacy-preserving but still informative way to identify participants.

Maintaining Anonymity in Text and Other Qualitative Data:

Qualitative researchers often analyze sensitive data

[15] For a demonstration, see the online appendix: https://aidanmilliff.com/publication/data-security-agenda-for-improvement/QMMR_Appendix.pdf

that are either naturally represented in text (historical or legal documents, social media data), or can be coerced into text (interviews). Text data are often very easy to re-identify or de-anonymize given basic contextual knowledge. Text data can also be uniquely identifying in its pragmatics (context, implication) even if identifying data have been removed from the semantics (words) and syntax (organization of words). An increasing number of text studies use data that are semi-public (like tweets), or clearly private (like longer transcripts of interviews, which are traditionally analyzed qualitatively (but see Milliff, 2021)). For these applications, researchers need to pay attention to de- anonymization concerns when sharing data in manuscripts or in replication files. One novel method for privacy-protecting analysis of sensitive text, building on the user-friendly Structural Topic Model by Roberts et al. (2013), is demonstrated in the online appendix. Topic models are typically used for comparing documents in corpora of text that are too large to read. This new approach uses topic modeling to compare documents in a corpus that is quite small, but for which presentation of raw, high-dimensional data threatens the privacy of the speakers represented in the text.

Topic modeling helps here because it focuses exclusively on *morphologic* patterns (words and their meanings). The data format that topic models ingest (data that would be shared for replication) is a document-term matrix (DTM): a format which ignores word order, making it difficult to re-assemble the original natural language. For longer documents (such as multiple sentences containing multiple verbs or multiple subjects), re-assembling the original document from a DTM is practically impossible. A document-term matrix, so long as no terms are themselves identifiers, is hard to connect to a particular individual.[16]

Topic modeling, however, is not a silver bullet for portraying patterns in qualitative data. Three downsides are worth noting. First, because topic modeling is an "unsupervised learning" tool, researchers usually cannot pre-specify the topics they would like a model to focus on. There is no ironclad guarantee, in other words, that a topic model will return topic clusters that are relevant to the research question at hand.[17] Second, if raw text data contains identifying terms (i.e., proper names), the topic model will contain them as well. Researchers who want to use topic models for privacy preservation need to ensure before modeling that directly identifying terms are censored or replaced. Third, topic modeling is time intensive. Using this technique for interview data, for example, requires text transcripts that are either time consuming or expensive to make. Cleaning the data to get rid of identifiers is likewise time consuming (or computationally intensive). If researchers can produce clean, non-identifying text from their qualitative data, though, topic models offer an interesting new way to present privacy-preserving summaries of sensitive information.

Mitigating Threats from Partners

As noted above, working with research partners changes the threat of re-identification in both qualitative and quantitative data. As such, I argue that additional techniques to preserve data security might be necessary or useful when a researcher is trying to prevent disclosure or re-identification by partners *before* data are shared publicly. I describe two techniques here, both of which are aimed at "keeping honest partners honest" and erecting modest barriers to the misuse of data after it is collected. Neither is a substitute for up-front work to vet partners and ensure that research collaborators share a strong commitment to treating participants with respect and dignity.

One intuitive way to reduce the risk that partners re-identify respondents in non-public data is to guard against over-sharing. Partners, in many cases, only need access to a specific subject of project information in order to participate in a project. Sharing *necessary* rather than *complete* versions of information like lists of participants, interview notes/tapes/transcripts, or recruitment blasts will limit the ability of partners to use contextual knowledge to re-identify research participants. With some partners, negotiating an agreement that limits sharing of re-identifiable data is not difficult because practitioner partners are primarily interested in finished products, like internal reports created by the researcher, rather than raw data. If social scientists work proactively to identify products that the partner wants, they may be able to avoid sharing sensitive data. When the structure of a partnership requires sharing PII or sensitive data with a partner, sharing via cloud storage is a good way to keep honest partners honest. Cloud storage platforms like Dropbox allow file owners to monitor access and downloads, so that researchers can make sure raw data aren't being misused.

A second way to reduce the risk of re-identification is to practice a "hand tying" strategy when working with partners, simply taking the possibility of data sharing off

[16] Mosteller and Wallace (1963) find that it is sometimes possible to identify authors based on the rate at which they use common words. Unless adversaries are searching for a known author in a corpus analyzed using STM and have a substantial amount of "labeled" reference material, this seems like an unlikely vector for the re-identification of interview transcripts.

[17] New work by Eshima, Imai, and Sasaki (2020) may mitigate this downside, allowing researchers to specify keywords for topic formation.

the table. This strategy is likely more useful in situations where the partner has some leverage over the researcher. One new, simple technique uses PGP (pretty good privacy) encryption software to set up a "vault" for sensitive in- formation. Supplementary materials in the online appendix provide step-by-step instructions. Once researchers "deposit" information into the PGP vault and delete unencrypted copies, the information is inaccessible until the researcher can access the key. If the key is left in another location and is not internet accessible, the researcher has effectively *tied her hands*: she cannot access the data herself. Other methods, like mailing physical media, could theoretically serve the same purpose without using computer encryption. Hand-tying is fundamentally a short-term solution—the researcher will have to access the private key eventually in order to unlock the data.

These tools, which provide simple ways to manage the risk of re-identification by research partners, also have some downsides. Both tools, for one, are additional work and make collaboration less smooth. The researcher takes on something like a systems administrator role in order to structure and manage data access—this could consume a lot of time. Second, these tools must be applied carefully and tactfully. It could be detrimental to a research partnership if partners felt disrespecte by the systems a researcher put in place to ensure data security. This is especially a risk with hand tying. If a researcher took steps to be unable to comply with a request for data, it would likely jeopardize future work with the requesting partner. Finally, neither of these tools prevent people from knowing what they saw with their own eyes. Research assistants and translators especially will still be able to identify research participants because they will be present at data collection. None of the techniques here can supplant good leadership, communication of clear ethical standards, hiring well, and vetting employees.

Conclusion

This article has proposed new techniques for improving data security in qualitative (and quantitative) political science research. I have argued that the re-identification of individual research participants is a particularly important threat to researchers' ability to fulfill the promises they often make to participants and have identified some simple technical solutions that should help researchers fulfill their promises while still responding to professional imperatives to make qualitative researchtransparent when possible. The article has tried to show that it is eminently possible to reduce the risk of data security failures when gathering and storing sensitive data. Whether or not better practices are ultimately adopted, though, depends on whether social science disciplines incentivize good practices and tolerate the compromises that good security requires.

Ensuring the security of sensitive data is an evolving challenge that researchers will have to revisit regularly throughout their careers. By ignoring data security, researchers are allowing the (admittedly small) likelihood of failure to increase over time. As political scientists adopt new technology for collecting and storing data, new threats to the security of that data will arise as well and may catch researchers unprepared. Contemporary data security practices are not "future proof" in any meaningful sense, so it is important for researchers to update their knowledge and use of relevant data security tools regularly to prevent the pile of un-addressed threats from growing too large. As the likelihood of data security failure appears to increase, the expected consequences of failure are surely growing: The popularity of collecting and analyzing large, identifiable data is in- creasing, which means the ethical and professional consequences of a potential data breach grow as well. Examples from the academy in the last two decades (e.g., Venkatesh 2008; McMurtrie 2014) already hint at the grave consequences that the release of sensitive data can have for research subjects. With these examples in mind, political scientists should not be content to wait for an even larger crisis to prompt the re-examination of data security practices in their own research.

Taking more systematic steps to guard respondent privacy is important, but not without trade- offs and fundamental limitations. Researchers should be mindful of these limitations as they adopt new tools. First, increasing privacy via more robust data security impinges on transparency. Even in the best-case compromise, rigorous data security protocols might make it harder to detect dishonesty in research by limiting the amount of data that a curious reviewer can demand to see. Second, good data security practices are sure to vary widely across the incredible range of methods and contexts in empirical political science. It is up to scholars to weigh the risks and benefits of specific data security techniques before deciding what strategy is most appropriate for their work. Third, using new and more complex data security techniques increases the difficulty researchers face in explaining their security precautions to research participants, who need to be adequately informed about the privacy risks of participating in political science research. Finally, there is a risk that promoting new tools for privacy protection incentivizes riskier behavior to begin with. To end with a warning: none of the technical solutions presented here are as ironclad as simply declining to collect and store sensitive data. Because the data security challenge is fundamentally political and social, technical fixes can help, but are naturally incomplete.

References

Angiuli, Olivia, Joe Blitzstein, and Jim Waldo. 2015. "How to De-Identify Your Data." *Communications of the ACM*, 58, no. 12 (December): 48-55.

APSA. 2020. "Principles and Guidance for Human Subjects Research." Ad Hoc Committee on Human Subjects Research, American Political Science Association. https://www.apsanet.org/Portals/54/diversity%20and%20inclusion%20prgms/Ethics/Final_Principles%20with%20Guidance%20with%20intro.pdf?ver=2020-04-20-211740-153

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2020. "Keyword Assisted Topic Models." Last revised March 10, 2021. https://arxiv.org/abs/2004.05964.

Fujii, Lee Ann. 2012. "Research Ethics 101: Dilemmas and Responsibilities." *PS: Political Science & Politics* 45, no. 4 (October): 717–23. https://doi.org/10.1017/S1049096512000819

Gibney, Elizabeth. 2017. "Ethics of Internet Research Triggers Scrutiny." *Nature* 550 (7674):16–7. https://doi.org/10.1038/550016a

Identity Theft Resource Center. 2017. *Data Breach Reports: 2016 End of Year Report.* El Cajon, CA: Identity Theft Resource Center. https://www.idtheftcenter.org/wp-content/uploads/images/breach/2016/DataBreachReport_2016.pdf.

Jacobs, Alan M., Tim Büthe, Ana Arjona, Leonardo R. Arriola, Eva Bellin, Andrew Bennett, Lisa Björkman, et al. 2021. "The Qualitative Transparency Deliberations: Insights and Implications." *Perspectives on Politics* 19 (1): 171–208. https://doi:10.1017/S1537592720001164.

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

King, Gary, and Nathaniel Persily. 2019. "A New Model for Industry–Academic Partnerships." *PS: Political Science & Politics* 53, no. 4 (October): 703-09. https://doi.org/10.1017/S1049096519001021

Knerr, Charles R. Jr. 1982. "What To Do Before and After a Subpoena of Data Arrives," In *The Ethics of Social Research: Surveys and Experiments*, edited by Joan E. Sieber, 191–206. New York: Springer.

Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. 2008. "Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com." *Social Networks* 30, no. 4 (October): 330–42. https://doi.org/10.1016/j.socnet.2008.07.002

McCallister, Erika, Tim Grance, and Karen Scarfone. 2010. "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)." National Institute of Standards and Technology (NIST) Computer Security Resource Center SP 800-122. https://doi.org/10.1016/j.socnet.2008.07.002

McMurtrie, Beth. 2014. "Secrets from Belfast." *Chronicle of Higher Education.* January 26, 2014. https://www.chronicle.com/article/secrets-from-belfast/.

Menoret, Pascal. 2014. "Repression and Fieldwork," In *Joyriding in Riyadh: Oil, Urbanism, and Road Revolt*, 21-60. New York: Cambridge University Press.

Milliff, Aidan. 2021. "Facts Shape Feelings: Information, Emotions, and the Political Consequences of Violence." *Political Behavior*. https://doi.org/10.1007/s11109-021-09755-1.

de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. "Unique in the Crowd: The Privacy Bounds of Human Mobility." *Scientific Reports* 3 (1376). https://doi.org/10.1038/srep01376.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58, no. 302 (June): 275–309. https://doi.org/10.1080/01621459.1963.10500849

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets," In *2008 IEEE Symposium on Security and Privacy*, 111–25. Oakland: IEEE.

Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57:1701–1778.

Radden Keefe, Patrick. 2018. *Say Nothing: A True Story of Murder and Memory in Northern Ireland.* New York: Penguin Random House.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." Working paper, prepared for the 2013 NIPS Workshop on Topic Models: Computation, Application, and Evaluation.

Samarati, Pierangela, and Latanya Sweeney. 1998. "Protecting Privacy when Disclosing Information: *k*-Anonymity and Its Enforcement through Generalization and Suppression." *Technical Report SRI-CSL-98*-04 Computer Science Laboratory, SRI International.

Scarce, Rik. 2005. *Contempt of Court: A Scholar's Battle for Free Speech from Behind Bars*. Lanham: Rowman and Littlefield.

Shilton, Katie and Sheridan Sayles. 2016. " 'We aren't all going to be on the same page about ethics:' Ethical practices and challenges in research on digital and social media." In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS 2016),* 1909–1918. Kauai, HI: IEEE.

Summers, Scott. 2016. "Organising, Storing and Securely Handling Research Data." PowerPoint presentation, UK Data Service, Essex, England, June 15. https://dam.ukdataservice.ac.uk/media/604451/2016-06-15_storing_data.pdf

Sweeney, Latanya. 2000. "Simple Demographics Often Identify People Uniquely." Working paper, Carnegie Mellon Univeristy Data Privacy Working Paper Series.

Sweeney, Latanya. 2002. "k-Anonymity: A Model for Protecting Privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (5): 557–70. https://doi.org/10.1142/S0218488502001648

Traynor, Michael. 1996. "Countering the Excessive Subpoena for Scholarly Research." *Law and Contemporary Problems* 59, no. 3 (Summer): 119–48.

Venkatesh, Sudhir. 2008. *Gang Leader for a Day*. New York: Penguin Press.

Wood, Elisabeth J. 2009. "Field Research," In *The Oxford Handbook of Comparative* Politics, edited by Carles Boix and Susan C. Stokes. Oxford: Oxford University Press, Oxford.

Zechmeister, Elizabeth J. 2015. "Ethics and Research in Political Science: The Responsibilities of the Researcher and the Profession," In *Ethics and Experiments*, edited by Scott Desposato. New York: Routledge, London.

Zimmer, Michael. 2008. "More on the 'Anonymity' of the Facebook Dataset—It's Harvard College." Blog post. October 3, 2008. https://michaelzimmer.org/2008/10/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/.

QMMR Qualitative & Multi-Method Research

Symposium:

Author-Meets-Critic: James Mahoney, 2021. *The Logic of Social Science.* Princeton, NJ: Princeton University Press.

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448059

Applying A New Approach to Knowing the Social World

Jennifer Cyr
Universidad Torcuato di Tella

"*[M]ainstream social science methods depend on the assumed truth of essentialism.*" (Mahoney 2021, 5)

The Logic of Social Sciences is a tour de force. The book and its author are advocating for revolution—a revolution in the social sciences. I admire the author greatly for writing it.

I am also rather overwhelmed by this book. The need to *un-learn* how we undertake research and think about causality in the social sciences, in order to *learn* it all once more, is daunting. Indeed, the book sets out myriad tasks for us as potential teachers and practitioners of the kind of social sciences it promotes. At times I wondered if the book was more aspirational than applicable.

In this intervention, I consider what we must do to put into action the kind of social science that this book promotes. I consider the central arguments of the text before turning to some of its implications when it comes to the practicalities of teaching the book in a graduate seminar. I consider, as well, what an application of this kind of work involves for research and publication. I find the book's content to be provocative and worthy of and—indeed—necessary for debate. Yet, I ultimately wonder whom the book's disciples will be—who will assume the difficult task of utilizing the approach in their work, blazing the trail for others to follow.

The premises of this book are twofold. First, to fully grasp the way the world works, we must let go of our essentialist biases. As social scientists we have been taught to view the things we care about (political parties, peace, the U.S. Congress, democracy) as entities that "possess inner essences" (Mahoney 2021, 1), which allow us to confer properties of action onto them and infer relationships of causality. This understanding of the world is incorrect. The events, entities, and activities we study do not exist independently of us. They are, instead, products of the "collective understandings among communities of individuals located in particular places and times" (Mahoney 2021, 2). To treat them as independent of our minds is to essentialize them erroneously and deny the (inter-)subjective nature through which we came to see them as important to begin with.

Second, in shedding or unlearning one approach to the social sciences, the book advocates for another: that of scientific constructivism. The scientific-constructivist approach is committed to the pursuit of scientific-based truths while taking into account the mind-dependent nature of the things we study. The book argues that we avoid essentialism by seeing what we research as belonging to categories that we actively construct in our minds and then (re-)calibrate in response to how our shared understanding evolves. To this end, the use of set-theoretic analysis is appropriate. It forces us to make our understanding of the things we study more transparent, since we must be explicit about the categories we create. It also involves defining the logic and importance of any given causal relationship, as well as the sequence of events connecting the causal event to the outcome of interest.

In all, to be better social scientists—that is, to accurately pursue causal truths about the world around us—we must re-think how we do social science. We must re-evaluate the ontological and epistemological orientations that have traditionally guided our work (at least amongst more positivist scholars). We must resist the temptation to view the world we wish to understand as being fully independent from how we perceive that world in our mind. "The reality as we experience it is upheld by mostly unconscious collective understandings that strike us as brute facts about an objectively and independently existing reality" (Mahoney 2021, 18).

This overly brief and necessarily pared down rendering of the principal arguments will be intuitive to some. The book's message is elegant, convincing, and draws upon premises that will be familiar to all. Nevertheless, the book points us down an unfamiliar and potentially paradigm-shifting path—at least for those of us who do positivist work. And, while I feel strongly that all social scientists must read this book, I also question to whom this book is oriented. Who will follow Mahoney's lead and see and study the world as it really is?

In my case, I vacillated between vigorously nodding as I read the book's pages and feeling overwhelmed by my incapacity to escape my own essentialist biases. For example, the notion that the things we study as social scientists are dependent upon us for their existence is not always intuitive, although it can be. It is not a stretch to acknowledge that concepts like "peace" and "democracy" are constructed inter-subjectively. Peace means different things to different people (Firchow 2018), as does democracy. We struggle to offer universal definitions of both, because our understanding of each is deeply contextual.

Nevertheless, other entities—a political party or a piece of legislation or the US Congress—feel more tangible and therefore amenable to "objective" analysis. A law is a law. The 116th US Congress enacted 344 of them.[1]

Yet, no two political parties are the same. And the US Congress can be thought of, at any given time, as a legislative power, a group of lawmakers, or a polarized (or democratic or imperfect) institution. In other words, it can fit into multiple categories. Consequently, the US Congress—as with all things we study in the social sciences—is called, by Mahoney and others, a *human kind*, or an entity that lacks intrinsic properties and dispositions because it is ontologically dependent upon us for its existence. Human kinds are mind-dependent. (*Natural kinds*, by contrast, are ontologically prior to human beings and their cognitions. They are mind-*in*dependent (see e.g., 2021, 14-18.) Without human beings, the U.S. Congress, as a political entity, would not exist.

The social sciences, ultimately, embody the study of human kinds. The book asserts that a rigorous approach to studying human kinds demands that we acknowledge that the entities we care about are constructions. A law is only a law once we acknowledge that our understanding of it—e.g., laws shape human behaviors; laws are made to be broken; laws only protect the wealthy/white/male—is shaped heavily by our interaction with the world.

[1] (GovTrack n.d.)

The book is also, as the author tells us early on, "committed to science as a mode of discovering truths about the world" (Mahoney 2021, 2). This statement gave me pause. Can one advocate for constructivism, or the study of the mind-dependent nature of social science categories, and also believe that there are "truths" of any kind out in the world? I understand the book to mean that the pursuit of science is one of evaluating the approximate truths of our propositions. I understand, as well, that "approximate truths," as used in the book, is not a new term. Nevertheless, the term "truth," even when used to refer to *logical* truths, seems to edify or essentialize a set-theoretic relationship, even one that is semantically or contextually bound, in ways that seem to contradict the spirit of scientific constructivism.

Indeed, the book refers to truth-preserving methodologies with a skepticism that is based precisely on our inability to preserve truths. Social science modes of data analysis, it tells us, use "partial generalizations to reach uncertain conclusions" (Mahoney 2021, 69). Are we uncovering truths about the world or positing possible causal paths?

These (not so?) minor distinctions are salient for me as a potential teacher and practitioner of this kind of approach to our work. Indeed, key questions I ask upon reading *any* new methods text are: Can I teach this? How can I teach this? To answer these questions, I feel I need a deeper understanding of the implications of this book.

For example, the book is clear in its assertion that we need to re-think how we teach the social sciences. We need to teach students how to recognize the multiple layers of human kinds that help to constitute the (mind-dependent) phenomena we study. We need to rethink measurement and conceptualization so as not to fall trap to the property-possession assumption, or the belief that the instances of a category possess shared essential properties (Mahoney 2021, 323). *Un*learning is the first step in understanding this new approach to the social sciences:

> Letting go of essentialism involves letting go of both human intuitions and longstanding approaches to social research. (Mahoney 2021, 5)

So, how do we do this? As a starting point, we should assign this text and some accompanying bibliography either in a methods course or in a philosophy of science course for those graduate programs that have them. Even if we leave aside how to utilize scientific-constructivism in a research setting, this book will be valuable for putting into relief the mainstream approaches to knowledge accumulation in the social sciences. What epistemological and ontological assumptions underpin conventional causal work? Why are these assumptions problematic? How does the scientific-constructivist approach render these assumptions obsolete? I can imagine taking a classic text and unpacking the essentialist assumptions that underpin its arguments. Students could then evaluate those arguments from a scientific-constructivist perspective. By juxtaposing conventional with scientific-constructivist models of causality, students could better understand and apply both to their own work.

Nevertheless, the skeptical, rather cynical, and completely exhausted professor in me still has doubts. For one, most professors will be as new to this approach as students. We will be just as susceptible to, if not *more* susceptible to, the essentialist bias(es) that we must unlearn to truly take the scientific-constructivist approach seriously. How do we thoughtfully address students' questions about a new paradigm when most of us sit firmly in the current/dominant one?

Additionally, once we (teach our students to) unlearn, what happens to the wealth of knowledge already accumulated via other approaches? The book tells us that the most commonly used type of causality—the counterfactual model—relies on the assumption that "variables and units of analysis stand in an approximate one-to-one correspondence with entities in the natural world" (Mahoney 2021, 94). This assumption is not met, however, when we study human kinds. As such, the conventional approach to causality, as used by social scientists for decades, is inappropriate.

What do we do, then, with the extensive literature that relies on inappropriate causal logics to draw conclusions? Will we need to re-examine those causal relationships, or are we simply re-thinking how those relationships are uncovered? For example, should we re-consider the finding that democracies tend not to go to war with each other, because most studies utilize a counterfactual logic to draw the inference? Or are we simply re-stating the relationship to accommodate a set-theoretical logic (e.g., country dyads that are democracies are a subset of not war)? Ultimately, how does an alternative understanding of causality—one based on the logic of regularity, as promoted by this book—impact our existing knowledge of the world? Can we still stand on the shoulders of those social scientists who came before us?

(I am deliberately choosing to be hyperbolic here. But if I am asking these questions, won't students also ask them? It seems worthwhile to take the arguments of this book to their logical conclusion.)

Finally, when it comes to teaching this approach, there is also a more normative question at stake. In many ways, this book advocates for going against conventional social science and adopting a different approach to studying the world. The author is swimming against a very strong current. In addition, then, to asking *how* and *why* we teach this approach to social sciences, there is the very real question of *should* we be? As instructors for graduate programs, we help to shape the next generation of social scientists. They are a key target audience for "conversion" to this kind of logic, precisely because they are the future of the discipline. On the other hand, their initial position within the hierarchy of academia—at the very bottom of the pyramid—means that they already face serious structural and institutional hurdles to achieving the success necessary to assume their role as the next generation. I suspect they would be additionally hampered if they applied this logic to their burgeoning research agenda.

Indeed, the choice to publish using a scientific-constructivist approach, which would include adopting a particular model of causality while also justifying it using the logic proposed by this book, would seem to be risky for a lot of newer scholars. In addition to teaching this book, then, we must also consider the implications of it for our work as researchers. To be sure, the book focuses on how to apply this approach for case-study and small-N research. We learn what a scientific-constructivist approach to causality looks like. But my questions are a bit more practical: For example, how difficult might it be to publish scientific constructivist-based research in a major journal? Would journal editors know how to evaluate this kind of work? I can imagine, at least early on, that they might require an appendix with a more in-depth discussion of the approach—but what might this look like? I also could imagine more stubborn or less innovative reviewers pushing the author to adopt a more conventional (read: essentialist) method to their research question instead of or even perhaps *in addition to* the scientific constructivist approach, to show how or if the findings are similar. How does one get around these potential hurdles?

Of course, set-theory and its use in the social sciences is not *new*. Many qualitative scholars use it implicitly, as the text notes and as many of us teach. Its explicit use, however, is rarer and, because of this, riskier for scholars.

A reasonable question to end this text, then, is for whom this book is ultimately written. Younger scholars are not yet fully socialized into the academy and therefore may be less constricted by the expectations and demands of mainstream social sciences and the essentialist biases that underpin these. On the other hand, the costs they assume in pursuing a less conventional path to research may be too high. Older scholars like myself, by contrast, may be too stuck in our ways or too overwhelmed by work and life to dig in and unlearn one approach to research in order to learn something new.

I raise these questions as someone who recognizes, values, and is ultimately humbled by the visionary nature of this text. The content is extraordinary. Mahoney offers us a potentially paradigm-shifting work. It merits our careful consideration. As a discipline I hope we are up to the task of taking its content seriously.

References

Firchow, Pamina. *Reclaiming everyday peace: Local voices in measurement and evaluation after war*. Cambridge University Press, 2018.

GovTrack n.d. "Statistics and Historical Comparison." Accessed March 15, 2022. https://www.govtrack.us/congress/bills/statistics.

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448168

Did Mahoney Just Kill the "Comparative" in Comparative Historical Analysis?

Gary Goertz
University of Notre Dame

If one does a search for the word "comparative" in Mahoney's book there are not many hits. There are references to methodologies that have comparative in the name, such as comparative historical, or qualitative comparative analysis, but nowhere in the book is a comparative methodology presented. So has Mahoney killed off comparative, or, with a nod to Mark Twain, are reports of its death exaggerated?

This of course demands an answer to a conceptual and research design question: What is comparative case study methodology? Given current trends in causal influences and methods I think there is an answer to that question. But if one explores a great deal of current case study research only a small percentage of it implements a comparative case causal inference strategy. Mahoney's book signals a change to within-case causal inference and process tracing to the disadvantage of comparative methods. Scholars need to read his book because it contains the methods they really need to know, exactly because truly comparative methods have become rare for reasons I outline.

Comparative designs do causal inference by comparing cases. This is for example treatment versus control cases in experiments. In QCA, the core logical minimization procedure involves case comparisons. The clearest and most obvious choice today would be matching. In a matched pair design, one has treatment versus control with matching on confounders. This is clearly very similar to Mill's method of difference or the classic most similar systems design. Hence, it is not surprising that scholars have argued for its application in the qualitative methods space. Weller and Barnes (2014) as well as Nielsen (2016) have drawn attention to the fact that matching is the current methodology for doing paired comparisons. If the work in question is not doing this, then I do not consider it "comparative" in terms of causal inference.

I argue that most of those doing multiple case studies, basically more than two, are doing what I will call serial case studies. They are examining a theory or hypothesis across multiple cases and arguing that their theory or hypothesis works in these cases. It is serial because this is done one case at a time, often using the kinds of methodologies of process tracing that Mahoney so nicely describes in his book.

It is useful to go back to a classic in the comparative case study literature, the book that introduced or made famous focused case comparisons. This is of course the George and Smoke book on deterrence in American foreign policy (1974). What did they actually do in this book? They had a series of questions that they asked of the eleven cases of deterrence analyzed in the book. In modern language this would be roughly coding eleven cases on a variety of variables. In the conclusions and an important appendix, they ask about what kinds of generalizations they can arrive at given their analysis. This illustrates a serial case study: a series of hypotheses or theory is applied sequentially across multiple cases. There is no paired design at all, and hence no comparative causal inference.

Leaping forward several decades we can look at Fairfield and Charman's (2021) recent APSA paper entitled, "Bayesian inference with multiple cases: unifying process tracing and comparative analysis." Quite interestingly, they note in the first paragraph the move away from "comparative" case studies to within-case process tracing in general. They then want to integrate the two within a Bayesian framework. In their example they do serial case studies. As an example, they use Slater's theory (2009; 2010) applied to an initial scope of Southeast Asian countries. They start with the case of the Philippines; then they do Vietnam. They are concerned with generalization, so the final case is Argentina. It is natural to do serial case studies within a Bayesian framework because one can update after each individual case as you go along. Within a Bayesian framework, it would be natural to continue doing case studies until one reaches a pre-established confidence level (see Dion 1998 for a nice discussion of this).

What might explain this trend toward serial case studies?

Matching works really well when there is a clear univariate hypothesis. However, a large number of case study books published in recent years[2] involve a two-way table that lays out the basic theory. I've illustrated a very common one in Table 1, where the cell entries are the values for Y. Here, we have moved from one independent variable to two. One way to think about this is to ask what kind of Boolean theory can generate such a table? How would a reader of Mahoney's book interpret this table?

Table 1: Case studies and two-way tables

	$X_1 = 0$	$X_1 = 1$
$X_2 = 1$	0	1
$X_2 = 0$	0	0

One plausible interpretation would be that the two variables are individually necessary and jointly sufficient for the outcome. That would generate the pattern of zeros and ones in the table (there are other equations that can generate this table, but I stick with Mahoney's set theoretic approach). Here we have three hypotheses and four cells in the table. It is not obvious what the comparative paired analysis should be. One could do a paired comparison for each of the three hypotheses. I do not think I've ever seen that in practice. What happens in qualitative case study books is that the authors go through a series of cases and argue that their particular theory or model works in all of them.

Another prominent way to do comparative historical analysis is what I am beginning to call the Luebbert model (1991). This book is an undisputed classic of

[2] I have looked at many of them published by major university presses over the last couple of years, for example, there are a significant number in security studies published by Princeton and Cornell.

comparative historical analysis. What was he doing in this book? First, we can start from the title which indicates he has three dependent variables he is explaining, fascism, liberal democracy, and social democracy.

For each of these three kinds of regimes he has a fascinating causal model, with connections between the three different causal models, that include some of the same variables, e.g., failure or success of labor movements (Mahoney in fact gives a set theoretic interpretation of his theory on pages 136–37). The book then argues that his three models explain all the cases.

While these two-way tables are typically seen as having no temporal ordering, in a really important chapter Mahoney talks about sequencing: What happens if X_1 happens before X_2 or vice versa, considering various possible necessary or sufficient condition relationships? I think one strong conclusion from that sequencing chapter is that whenever one sees a two-way table one should ask sequencing questions.

The two necessary condition hypotheses in Table 1 lead naturally to counterfactuals, which is a core contribution of Mahoney's book. He also talks at length about sufficient conditions. One move is to replace real case comparisons with counterfactual ones. This has parallels in the statistical literature with the synthetic control method (Abadie, Diamond, and Hainmueller 2015). In both instances one creates counterfactual observations, which are compared to the real ones. Starting with Lewis's defining treatment of counterfactuals (1973) this becomes a discussion of possible worlds.

As seen in Table 1, there are two necessary conditions which normally produce counterfactual claims. In fact, one could focus mostly on the (1,1) cases—which is what people do in practice—and then do counterfactual analyses on the absence of the two necessary conditions. This is a matched paired comparison with one real case and two counterfactual cases one for each of the two hypotheses.

Mahoney's chapters on critical event analysis, counterfactuals, sequence analysis, etc., are specific techniques of process tracing. Hence, he provides a great toolkit for those doing case studies in case study and multimethod work. In an important sense, "comparative" has not disappeared at all but must be rethought with these new methodologies.

In short, "comparative" often means "does my theory apply to other or multiple cases?" Mahoney's book gives essential tools for answering that question.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59, no. 2 (April): 495–510. https://doi.org/10.1111/ajps.12116

Dion, Douglas. 1998. "Evidence and Inference in the Comparative Case Study." *Comparative Politics* 30, no. 2 (January):127–45. https://doi.org/10.2307/422284

Fairfield, Tasha, and Andrew Charman. 2021. "Bayesian Inference with Multiple Cases: Unifying Process Tracing and Comparative Analysis." Paper presented at the Annual Meeting of the American Political Science Association, Seattle, Washington, September 30- October 3.

George, Alexander L., and Richard Smoke. 1974. *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press.

Lewis, David. 1973. *Counterfactuals*. Cambridge: Harvard University Press.

Luebbert, Gregory. 1991. *Liberalism, Fascism, or Social Democracy: Social Classes and the Political Origins of Regimes in Interwar Europe*. Oxford: Oxford University Press.

Nielsen, Richard A. 2016. "Case Selection via Matching." *Sociological Methods & Research* 45, no. 3 (August): 569– 97. https://doi.org/10.1177/0049124114547054

Slater, Dan. 2009. "Revolutions, Crackdowns, and Quiescence: Communal Elites and Democratic Mobilization in Southeast Asia." *American Journal of Sociology* 115, no. 1 (July): 203–54. https://doi.org/10.1086/597796

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

Weller, Nicholas and Jeb Barnes. 2014. *Finding Pathways: Case Selection for Studying Causal Mechanisms in Mixed-methods Research*. Cambridge: Cambridge University Press.

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448178

Counterfactuals, Mechanisms, and Background Beliefs in *The Logic of Social Science*

Alan M. Jacobs
University of British Columbia

Elegant in its architecture and sweeping in its ambition, James Mahoney's *The Logic of Social Science* (2021) addresses deep philosophy-of-science foundations, set-theoretic methodology, and a suite of set-theoretic analytic tools. The text is exceedingly lucid and aided by visuals (Euler diagrams) that lend remarkable clarity to complex set-theoretic relations. Drawing on rich empirical examples, the book provides clear, actionable, innovative guidance on how to engage in case-level set-theoretic analysis of various forms, including counterfactual analysis, sequential analysis, and the analysis of critical events. Among the book's most enlightening features are the ways in which the it maps causal and inferential concepts native to other analytic frameworks into set theory. Perhaps the most remarkable of these translations is the book's set-theoretic rendering of Bayesian inference, in a chapter coauthored with Rodrigo Barrenechea. While I am entirely persuaded that Bayesianism assumes and requires a set-theoretic approach, as the authors claim, it is nonetheless striking to see how fully set-theory can represent a mode of inferential reasoning that we typically undertake in probabilistic terms.

I learned enormously from this book and have found it extremely fruitful to grapple with Mahoney's arguments, even when I did not entirely agree with them. I will use the remainder of this essay to frame two questions that the book raised in my mind. Both are questions that I think have significant implications for how we think about causality and causal inference within a set-theoretic framework. I raise them in constructive spirit and in the hope that I can learn more from Jim as he responds in his own piece.

First, to what degree do we have to sign on to the book's particular understanding of causality in order to employ and reap the benefits of its set-theoretic methodology and methods? One way to describe the book's structure is that it offers us a set of analytic strategies grounded in a methodology, which itself is placed atop a view of causality that is grounded in a particular ontology and epistemology. But how close are the logical relationships among these elements?

When it comes to causality, Mahoney pushes back against an understanding that has, over the last couple of decades, become pervasive in causal-inferential work in the discipline: the counterfactual model (Rubin 1974; Holland 1986). In the counterfactual view, X is a cause of Y in a given case if, under an imagined intervention that changed the value of X in the case (with all else of causal relevance to Y held constant), the value of Y would also change. Causes, on the counterfactual view, are "difference-makers." Mahoney argues, however, that the rival "regularity" view of causality is a better fit for causal inquiry in the social sciences.[1] In the regularity view, X is a cause of Y if (a) X precedes Y in time, (b) X makes direct or indirect contact with Y in time or space (i.e., via a mechanism), and (c) X is part of a minimized solution set that is constantly conjoined with Y (e.g., is necessary, sufficient, or an INUS or a SUIN condition for Y).

Mahoney's primary argument for employing the regularity over the counterfactual view is that the counterfactual view is inappropriate for studying relationships among "human kinds." Drawing on a distinction common in philosophy and the cognitive sciences, Mahoney (2021, 14) defines "human kinds" as entities that we mentally classify as similar "on the basis of characteristics that are not mind-independent properties," while "natural kinds" are entities that are ontologically prior to human beings and that we classify on the basis of shared, essential, mind-independent properties. While a revolution is a human kind, for instance, a photon is a natural kind. More generally, Mahoney argues, the entities we study as social scientists are typically human, not natural, kinds.

The central problem with using the counterfactual model in connection with causes and outcomes of the "human kind," according to Mahoney (2021, 94), is that the model "assumes and requires that variables and units of analysis stand in an approximate one-to-one correspondence with entities of the natural world," whereas human kinds are mental constructs. Mahoney argues that we can think of the problem of non-correspondence as a violation of the Stable Unit Treatment Value Assumption (SUTVA) central to standard approaches to causal inference. At the heart of the counterfactual model is an imagined

[1] Mahoney also compares the regularity view to the "causal power" view, which I do not address here.

change in X. The problem, Mahoney contends, is that a change in any human-kind treatment will always be ill-defined: the same constructed treatment category could, at the level of natural kinds, entail "a mostly unknown and unknowable change that is not constant across any two units" (2021, 94). Thus, for instance, "Democracy cannot cause economic growth across different countries in the ways proposed by counterfactual models because these categories do not map the structure of an objective reality." In fact, it is the heterogeneity of meanings of our human-kind categories across units, Mahoney argues, that explains much of the instability of empirical results (on topics like democracy's effects on growth) derived from counterfactual-model-based inquiry.

Mahoney also views the regularity model as encompassing a wider range of relationships that we would want to be able to think about as causal and that feature prominently in set-theoretic methods. In particular, INUS and sufficient-but-not-necessary conditions count as causes under the regularity view but are not difference-makers (i.e., removing them alone does not change the outcome).

But it is unclear to me how much is in fact at stake—for the methodologies and tools we deploy—in the distinction that Mahoney is making here. First, it is not obvious to me that the regularity view constitutes a distinctive *definition* of causality. While I agree that temporal priority and spatiotemporal proximity are relevant to causal inquiry, these seem more like *empirical criteria* that we use to identify a cause than like necessary components of the concept. Evidence that X happened before Y or evidence of the operation of a mechanism connecting X to Y constitutes empirical *support* for the claim that X caused Y. But do we need to specify these features as part of the definition? Put differently, suppose we know that X is a difference-maker. Would it then make sense to insist that X must additionally have occurred before Y and be connected to Y via a mechanism before we are willing to deem X a cause? Given a set of commonly held assumptions about how the world works (e.g., that the future cannot influence the past), it seems to me that we get temporal priority and spatiotemporal proximity "for free" —they are automatically satisfied—once we know X to have made a difference.

I also think we can see in actual research practice the ways in which the empirical examination of mechanisms can readily operate in support of a counterfactual view of causality. Chapter 5 of the book, on counterfactual analysis, presents an informative example. Mahoney and coauthor Barrenechea discuss Harvey's (2012) study of the origins of the Iraq War, focusing on the role of George W. Bush's election as President. In seeking to assess the causal role of Bush's election, Harvey gathers evidence on the causal process that played out under Bush's presidency, culminating in the invasion of Iraq. Importantly, he engages in this analysis of process to allow for *counterfactual* inquiry: understanding the process that in fact unfolded allows Harvey to ask how much of this process would likely have changed under a counterfactual Al Gore presidency. Together with evidence about "actual" Gore, this analysis points Harvey to the inference that the causal process would likely not have been very different under a hypothesized change in the 2000 election result.

Here the analyst is not studying mechanisms only to establish indirect spatiotemporal contact between Bush's election and the Iraq War, but to provide leverage on a question *about the case's potential outcomes*. The understanding of causality here appears essentially counterfactual, with evidence on mechanisms serving as empirical support for claims about what would have happened under the counterfactual.

Second, I am not sure how the regularity view performs better than the counterfactual view in addressing problems of non-correspondence. I may be missing something, but it seems to me that the claim that X is a necessary, sufficient, INUS, or SUIN condition for Y, across some universe of relevant cases, makes the same demands—in terms of the required homogeneity meanings of X and Y across units—as does the claim that X is a difference-maker for Y. To claim that X's presence always implies Y, for instance, do we not run up against the same issue of whether X means the same thing in all instances in which we think we have observed it? It is not clear to me how set-theoretic relationships get around the problematic, unstable mappings between our constructed categories and what is going on in our cases at the level of natural kinds.

Finally, I see the counterfactual model, understood in terms of the potential-outcomes framework, as just as capacious as the regularity view in the kinds of causes that it can accommodate. We can, for instance, readily represent a set of potential outcomes corresponding to sufficient-but-not-necessary causes. If we have, say, three potentially causal relevant variables—X_1, X_2, and X_3—we can represent X_3=1 as a sufficient but not necessary condition for Y=1 if by defining the potential outcomes Y(0,0,1)=1, Y(0,1,1)=1, Y(1,0,1)=1, Y(1,1,1)=1, and (say) Y(1,0,0)=1. It is slightly more complex, but no less logically straightforward, to write down a set of potential outcomes under which some condition W is an INUS condition (posit Y to be 1 under all permutations of conditions under which W and all other members of its sufficiency combination are present as all as all under all permutations in which all members of any other sufficiency combination are present, with Y posited to be 0 otherwise).

Overall, then, I wonder how much the book's methodological arguments and contributions actually hinge on the book's philosophy of science and attendant view of causality. If they don't, I think that would be good news. It would mean that the book's guidance on and innovations in the use of set-theoretic methods are of broader relevance and that these tools can readily be taken up, without philosophical contradiction, by the many in our field who subscribe to a potential-outcomes understanding of causality.

The second question I would like to pose is: how, in this framework, should we be grounding our case-level inferences in general causal knowledge? The book's core focus is on explaining outcomes at the *case* level, not on developing or testing general causal propositions. But if I am understanding correctly, we are intended to draw on general, background knowledge about logical causal relationships in the world when making case-level inferences. As Mahoney writes: "To excel at designing good set-theoretic tests, one must possess knowledge of relevant existing generalizations, perhaps established from studies of other cases" (2021, 135). In essence, the book offers us an analytic framework for combining general causal knowledge with evidence about specific cases to develop case-level (token) causal claims.

So far, so good. This logic, however, then raises the question of where our general, background knowledge of causal relations is supposed to come from. I believe part of the answer is that we can draw on tools like Qualitative Comparative Analyisis (QCA) that seek to test for general causal structures in a set of cases. But I am not sure if QCA could ever be enough.

Consider, again, the example of counterfactual causal-chain analysis in Harvey's study of the Iraq War. According to Mahoney and Barrenechea, the causal chain that Harvey assesses includes steps such as (where "**-S->**" indicates a relationship of sufficiency):

> *Iraq is a central foreign policy concern* **-S->** *UN inspectors are brought back to Iraq* **-S->** *faulty intelligence about* WMDs (2021, 165)

It is hard to imagine that QCA alone could ever yield credible knowledge about the near or probable sufficiency of the conditions here. It is surely unlikely that the cases exist to establish the general sufficiency of a foreign adversary being a central foreign policy concern for generating the return of UN inspectors to that country under circumstances "like" those prevailing in the Iraq War case.

Moreover, even where potentially relevant population-level QCA inferences exist, there is a judgment involved in deciding whether those inferences apply to the case at hand. Was the QCA sample sufficiently *like* the case we're trying to explain? Of course, this is a challenge for any inferential approach that involves applying population-level generalizations to specific cases.

It seems to me that the general knowledge required for inferences about token causation in this framework can only ever be empirically grounded *in part*. We will usually, I would think, need to draw on other sources: for instance, on logical reasoning (whether informal or instantiated in a formal or causal model) or expert consensus. This raises the question of whether there are better and less good ways of grounding our general beliefs. At a minimum, I would think transparency would be especially important here: we would want researchers to lay bare the foundations of the general causal beliefs they are mobilizing in a given case analysis—and perhaps even to undertake sensitivity analyses, showing the degree to which their case-level inferences are dependent on the choice of general beliefs. Readers might also want to understand case-level inferences in this framework as always being assumption-dependent—on the general beliefs being invoked—much as we need to interpret observational regression results as conditional on a set of model assumptions.

I would be interested in hearing more of Mahoney's thinking about the problem of background knowledge in this framework: on how we can or ought to form our general beliefs about set-theoretic relations; how we should map general beliefs into specific cases; and how the way we do these things should affect the presentation and interpretation of our inferences.

There is, of course, far more to *The Logic of Social Science* than I have touched on in this short comment. This is a volume packed with conceptual and methodological innovation and brimming with insight into the enterprise of causal explanation. Anyone interested in qualitative and case-study methods should read and contend with this magnificent book.

References

Harvey, Frank P. 2012. *Explaining the Iraq War: Counterfactual Theory, Logic, and Evidence.* Cambridge: Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–960. https://doi.org/10.1080/01621459.1986.10478354

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1037/h0037350

What Are the Boundaries of this Potential Revolution? Exploring the Shape of Mahoney's Scientific-Constructivist Social Science

Carsten Q. Schneider
Central European University (CEU), Vienna

Jim Mahoney has written an opus magnum. The breadth, scope, and potential implications of the system of thought proposed by Mahoney are such that any attempt—or at least any attempt by me—to do justice to all important aspects of this book is, unavoidably, doomed to fail. I will necessarily have to be very selective and focus on those aspects that I feel reasonably competent. And, even more, rather than commenting or responding to some of Mahoney's main arguments, I will mostly ask questions of clarification. These will be genuine questions, not rhetorical ones. I am curious to hear Mahoney's answers because—so my hope—those answers will further sharpen the boundaries of the revolution of the social science (singular!) that Mahoney is arguing for. I shall also disclose that I not only find this book mind-blowing and horizon-widening, but also largely agree with its main gist.

My questions (and, occasionally, some tentative answers) are structured in four groups. The first set aims at probing the difference in practice between, on the one hand, "scientific constructivism," Mahoney's core concept and favorite logic of the social science and, on the other hand, what he identifies as the current predominant logic of "essentialist" approaches. The second set of questions focuses on Mahoney's version of a regularity theory of causation. In the third set, I address the difference between fuzzy sets (an established term) and continuous sets (the term preferred by Mahoney). And the fourth set is a mixed bag of comments on issues that are less central to Mahoney's overall argument.

What is the Difference between Scientific Constructivist and Essentialist Approaches in Practice?

In Part I of his book, Mahoney makes the, I think, very convincing case that because social science is fundamentally different from (most) natural sciences, their methods must also differ. This difference stems from the fact that social scientists do research on social kinds. Unlike natural kinds, social kinds only exist in the researchers' (collective) minds. From this ontological position, Mahoney argues, radical consequences follow for the practice of how social science should be done.

For Mahoney, social science must be scientific constructivist. It is constructivist because social science concepts are mind constructed and need to be captured by assigning membership scores of cases in sets. This is an inherently interpretive act. The scientific part largely rests in the use of formal logic for analyzing relations between sets. With the following questions, I try to understand better what the practical implications for social science research are of Mahoney's position, according to which there is a sharp ontological division between scientific constructivist and essentialist research.

First, Mahoney explains in detail how the practice of calibrating sets —that is, establishing the membership of cases in mind-dependent social science concepts—is fully in line with the constructivist element of scientific constructivism. What I am wondering is whether other elements in the research process are equally constructed or whether they fall into the "scientific" domain of scientific constructivism. In particular, I am curious about the status of set relations. Arguably, identifying set relations of necessity and sufficiency (and some more complicated derivatives) is the goal of scientific-constructivist research. But are those set relations socially constructed or are they merely the result of applying the cold rules of formal logic? The book seems to allow for both answers. On the one hand, if set membership scores are constructed, set relations also ought to be constructed. On the other hand, formal logic and mathematical rules not only represent an important element of the scientific component of scientific constructivism, but, following philosophers like Leibniz, for Mahoney they also enjoy the elevated ontological status of absolute truth. This status of logic is remarkable, for truth is a scarce resource in a social (science) world in which things are made up by humans and therefore are contested and subject to change over time and space.

A second question probing the practical implications of the scientific constructivist revolution is this: Should scientific-constructivist researchers pay less attention to things that are currently associated with essentialist research, but which also feature high on the agenda of set-theoretic methods? Here I have in mind discussions on appropriate robustness tests for QCA results or the properties of different (minimization) algorithms for analyzing set membership data. My take on this would be that these more technical and computational problems pertain to the scientific aspect of scientific constructivism and should therefore continue to play an important role in refining and improving scientific-constructivist methods. In the book, however, there is little to no mention of such topics of applied empirical research and I am not sure if this is done intentionally or is simply caused by lack of space.

Third, Mahoney convincingly argues that it is wrong to take an essentialist perspective on social categories. How wrong, though? Mahoney himself writes (2021, 66) that there are two feasible ways of interpreting set membership scores: as facts (essentialist approach) or as societally agreed facts (constructivist approach). Whether one or the other approach is chosen does not seem to make any (important) difference in applied research. Mahoney even concedes that essentialist research can be (and often is) very successful in predicting social events—even if, according to Mahoney, by definition and default, it cannot establish causality. If my reading is correct, the question becomes: Does it matter in practical terms whether we assume essentialism or constructivism when analyzing sets?

Fourth, by design, scientific-constructivist research is about discrete categories of social phenomena and their set relations. My question is: Where, if anywhere, is there room for all those relevant questions that have at their core non-discrete phenomena and that are focusing on forms of associations other than set relations? For instance, in scientific-constructivist social science, can we continue to ask questions such as: Is economic performance related to political participation? or Does the amount of exposure to hate speech on social media increase the risk of political radicalization? Currently, such questions seem to dominate in essentialist empirical social research. Declaring (causal) research on them impossible would be quite a revolutionary step that might need some more explicit treatment and justification.

Fifth, and somewhat related to the last question: Can one imagine and design experiments that stay true to the principles and practices of the scientific-constructivist approach or would that amount to a contradiction in terms? If yes, what would such experiments have to look like? If no, what drives the incompatibility between scientific constructivism and experiments? Is it that the former is largely Y-oriented, whereas the latter largely X-oriented? Or is the incompatibility rooted at a deeper, ontological level?

Scientific-Constructivism and the Regularity Theory of Causation

Mahoney discusses three different theories of causation: causal power, counterfactual, and regularity (for details, see the very informative Table 3.1 on page 91). He identifies the latter as the most fitting for the scientific-constructivist approach. Mahoney's version of regularity theory of causality stipulates that cause X must (a) precede outcome Y in time; (b) make direct or indirect spatial contact with Y; and (c) be part of a minimized solution that is constantly conjoined with Y (2021, 91). This raises several questions of clarification for me.

First, the last criterion – that the cause is part of a minimized solution set – takes care of the question of causal relevance: Are all sets in a solution difference-maker causes? It leaves out, though, the question of causal completeness: Are all difference-making causes for the outcome included in the solution? This makes me wonder how in Mahoney's regularity theory of causation and, by extension, in applied scientific-constructivist research, the issue of model under-specification is dealt with.

Second, according to Mahoney, regularity models of token causality are best fitting for scientific constructivism. One of the most developed scientific-constructivist methods is the set-theory based method of Qualitative Comparative Analysis (QCA). My understanding of QCA is that it reveals type causality. If this is correct, I am asking myself: Does this make QCA incompatible with scientific-constructivist research? Does it prevent QCA from being able to reveal causality? And, in which way, if any, would either QCA and/or Mahoney's vision of social science need to be adapted to be fully compatible? Perhaps my next question provides a partial answer to this set of questions.

Third, I like Mahoney's interpretation of regularity theory of causation requiring spatiotemporal contact between X and Y. I read this the following way: For complete causal inference based on a regularity theory of causation one must include an analysis of the causal mechanism between X and Y that underpins a cross-case effect of X on Y. I am sure, many case-based researchers could not agree more. This reading would also solve partially my previous question on the causal status of cross-case patterns identified with QCA. To be causally interpretable, such cross-case pattern also need to be based on some evidence on within-case mechanisms. This is precisely what the literature on set-theoretic multi-method research is mostly about (e.g. Schneider forthcoming). My only question would then be this: Why do other contemporary proponents of regularity theories of causation not seem to attribute any importance or relevance to causal mechanisms (e.g., Baumgartner 2008)? In fact, most of them would probably explicitly deny any role for mechanisms in causal inference within a regularity theory framework. If the addition of mechanisms to this framework is an innovation by Mahoney, then it is probably worthwhile to point this out more clearly. Criticisms from other regularity theorists on Mahoney's requirement for a causal mechanism is likely to come his way and defending this addition is, I believe, a worthwhile effort.

Continuous vs. Fuzzy Sets

Mahoney replaces the established term "fuzzy sets" with the term "continuous sets." This is consistent with his earlier writings, in particular that with Gary Goertz in their seminal "Two Cultures" project (Goertz and Mahoney 2012). I have already expressed my uneasiness in a previous QMMR newsletter (Schneider and Wagemann 2013). The disagreement is not about which term to use or whether changing the term unsettles the semantic field and creates more confusion than necessary. The more important point is that the introduction of a different term seems to come with the introduction of a different meaning: fuzzy sets and continuous sets are probably not meant to mean the same thing. Let me explain what I think the difference is and why the meaning of continuous sets is potentially problematic for scientific constructivist research.

Fuzzy sets are sets. They first and foremost establish qualitative differences between members and non-members of a set. In other words, fuzzy sets categorize cases just like crisp sets do. With fuzzy sets, the distinction between members and non-members is established at the membership score of 0.5, the so-called point of maximum ambiguity (Ragin 2008).

Continuous sets also must establish such a qualitative distinction, else they are not sets. The question is where on the range of membership values between 0 and 1 is this qualitative shift located? The notion of "continuous" seems to rule out that the qualitative shift occurs at the 0.5 membership value. A more likely candidate is the membership value of 0. All cases that hold membership of higher than 0 are not only partial members of the set in question, but also qualitatively different from those that hold zero membership. For instance, in the set of tall person, someone with membership 0.1 would be qualitatively identical to someone with membership 0.9 but qualitatively different from someone with zero membership. As said, this is not how things are normally seen with fuzzy sets, where all cases below 0.5 are qualitatively different from those above 0.5.

Here is what I find problematic about a reinterpretation of where the qualitative anchor rests in continuous sets. First, if my interpretation about the location of the qualitative anchor is correct, it would need to be spelled out clearer than it is in the book. It represents a deviation from the common interpretation of fuzzy sets and triggers a series of (unintended?) consequences that I spell out in the following.

Second, imagine a case with, say, 0.3 membership in the set of "tall person." With continuous sets, it qualitatively counts as a tall person because its membership is higher than zero. The problem with this becomes apparent if we ask: What is this person's membership in the logical complement of "not-tall person"? The 1-x rule for logical negation yields a membership of 0.7 in the set of not-tall person. Hence, that very same person would also qualitatively count as a not-tall person. This is a contradiction in terms: one and the same person cannot count qualitatively as both tall and not-tall. Note that with fuzzy sets, this logical fallacy does not occur. With 0.3 membership in the set of tall person, the person in question qualifies as not-tall because their membership is below the qualitative anchor of 0.5. This classification becomes clearer if we calculate the person's membership in not-tall persons: 1 - 0.3 = 0.7, thus above the qualitative anchor of 0.5.

Third, because of its property to never classify cases as qualitatively belonging both to a set and its negation, fuzzy sets can be used in the analytic apparatus of QCA. At the heart of QCA-based research is the truth table. This table consists exclusively of 1s and 0s. Representing fuzzy sets in "crisp set-looking" truth tables can only work because the qualitative anchor in fuzzy sets is located at 0.5. With continuous sets and their alleged location of the qualitative anchor at 0, the current QCA principles and practices would need to be radically rethought and adapted. In the spirit of this reflection on Mahoney's book, I am turning this observation into a question: Am I right in locating the qualitative anchor in continuous sets at the membership value of 0? If yes, am I right in pointing out some problematic consequences of this redefinition of fuzzy sets? And if yes, how can these intended or unintended consequences be fixed?

Two Miscellaneous Observations

Mahoney states that the empirical importance of set relations can be captured by how close a given condition X comes to being both necessary and sufficient for an outcome Y. I fully agree. There is, however, also a second element of empirical importance that is not mentioned in the book. For necessity claims, importance also hinges on how big condition X is in relation to its logical negation ~X. In other words, if condition X is very big and therefore close to a constant, then ~X is very small. It is potentially trivial to claim that such a very big X is necessary for any given outcome Y, because it is virtually impossible for a very big set to not be a superset of whatever else set Y is. The QCA literature has developed the parameter of Relevance of Necessity (RoN) to capture both sources of empirical importance/relevance of necessity claims (Schneider and Wagemann 2012, chapter 9.2). Since not much attention is paid to this source of set-relational trivialness, I am wondering if this is because Mahoney does not deem it relevant for scientific-constructivist research or whether it has been de-emphasized due to lack of space or lack of importance.

For sufficiency claims, a similar problem of skewed set membership scores exists, but it is of practical relevance only with fuzzy sets. A condition X can be so small that it passes the sufficiency test for both outcome Y and its negation ~Y. Claiming that this X is sufficient for both outcomes would be logical nonsense and must be avoided. This problem is not fully addressed by Mahoney's conceptualization of empirical importance of set relations either. Charles Ragin has developed the PRI parameter to avoid the pitfall of such simultaneous subset relations (for details, see also Schneider and Wagemann 2012, chapter 9.2). Furthermore, with the traditional interpretation of fuzzy sets and their location of the qualitative anchor at 0.5, most of the dangers of these simultaneous subset relations of X vi-a-vis both Y and ~Y can be kept under control. With the notion of continuous sets and their qualitative anchor at 0, in contrast, this analytic problem seems to increase and strategies for containing it would become even more relevant.

My experience from many years of teaching set-theoretic methods is that a sizable chunk of participants tends to struggle when first exposed to a comparatively modest level of formal logic. Even more advanced students continue to sometimes mix up necessity and sufficiency when looking at set relational patterns in their data. My ad-hoc amateur evolutionary theory explanation of this has long been that (formal) logic does not seem to be hard-wired into the human brain because it was (and still is) not needed for survival. This, however, clashes with Mahoney's view, according to which logic is an essential tool in human reasoning. I would be curious to know how these seemingly opposing views and experiences could be reconciled.

A Concluding Praise

James Mahoney and his book deserve the highest praise. He has mine, not only for the vast knowledge and sophisticated mind that it takes to write such a text. I also admire the courage that is required to call for a revolution and to face some of the reactions that Mahoney's call to arms will (hopefully) trigger.

References

Baumgartner, Michael. 2008. "Uncovering Deterministic Causal Structures: A Boolean Approach." *Synthese* 170 (1): 71–96. https://doi.org/10.1007/s11229-008-9348-0.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Contrasting Qualitative and Quantitative Paradigms*. Princeton: Princeton University Press

Ragin, Charles C. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Schneider, Carsten Q. forthcoming. *Set-Theoretic Multi-Method Research: A Guide to Combining QCA and Case Studies*. Cambridge: Cambridge University Press.

Schneider, Carsten Q., and Claudius Wagemann. 2013. "Fuzzy Sets are Sets — A Reply to Goertz and Mahoney." *Qualitative and Multi-Method Research, Newsletter* 11 (Spring): 19–22.

Schneider, Carsten Q., and Claudius Wagemann. (2012). *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448191

Two Views of Within-Case Analysis: Ambiguities about Process Tracing in *The Logic of Social Science* and Beyond

Hillel David Soifer
Temple University

I am grateful to have been included in this conversation with esteemed colleagues about Jim Mahoney's important new book. Rather than using this opportunity to offer praise of the book (which would be easy to do) or criticism of an already-published work (which would be less than useful for the author, or for readers) I would like to use Mahoney's book as an opportunity to explore a tension that, in my view, underlies much of the contemporary scholarship on qualitative methods, and to suggest that the book itself is a bit at odds with itself on a core element of qualitative research in ways that point to some unresolved issues for us to address as qualitative methodologists.

Let me begin by pointing to the existence of a clear consensus among scholars of qualitative and mixed methods about the central place of within-case analysis, rather than comparison across cases, as the core of qualitative research methods. In arriving at this consensus, scholars have converged on the importance of so-called "process tracing" in the elaboration and evaluation of causal claims.[1] But this striking convergence on within-case analysis conceals what strikes me as a tension about what exactly is entailed in within-case analysis. Here, I see two distinct positions articulated.

On the one hand, some scholars explicitly argue that within-case analysis entails the tracing of causal mechanisms linking proposed cause to effect (Beach and Pedersen 2013). Bracketing the robust debate about what a causal mechanism is, we can see that scholars who take this position hone in on the central importance of identifying evidence that sheds light on Charles Tilly's (1996) invisible elbow, where cause becomes effect. On the other hand, some scholars articulate, more or less explicitly, the view that within-case analysis should leverage any evidence that can help arbitrate among various proposed explanations or shed light on the validity of a particular candidate explanation. This approach can be found most explicitly among Bayesians (Fairfield and Charman Forthcoming; Humphreys and Jacobs 2015) but is also articulated in work like the discussion of "causal process observations" by Collier, Brady and Seawright (2004), which includes information about context and other features of a case beyond causal mechanisms as important to within-case inference.

Sometimes this tension about what is at the heart of within-case analysis (or what ought to be—after all, methodology is intended to be prescriptive) is clearly and explicitly articulated. The forthcoming Fairfield and Charman book cited above, for example, makes a clear and explicit case for seeking any information about a case that can arbitrate among proposed explanations, and that process-tracing defined narrowly as evidence about causal mechanisms is too restrictive an approach to qualitative research.[2] But sometimes this tension lurks even where it is not articulated. And I think that it lurks in *The Logic of Social Science*. In particular, while the "regularity" model of causality advanced in Chapter 3 places causal mechanisms as a central element, later chapters in the book, even as they purport to build on this model, downplay it and take a broader or more eclectic view of the tasks at hand in within-case analysis.

In Chapter 3, Mahoney draws on approaches in philosophy to define a "regularity model" of causality, composed of three elements: temporal succession, spatio-temporal contact, and logical regularity. The first of these—that a cause must begin before an effect—is straightforward and need not detain us here. The third is intended to give us tools to tease apart spurious causes from important ones and is derived from the set-based approach, and the underlying scientific constructivist ontology that is central to the book as a whole. Because the book is oriented around the set-based approach that derives from the scientific constructivist ontology, the logical regularity component of causal appraisal follows from his ontological starting point.

Spatio-temporal contiguity, or contact (direct or indirect) between cause and effect, grounds the centrality of causal process in the regularity model. Mahoney uses the requirement of assessing spatio-temporal contact to bring causal mechanisms into the regularity model of causation. So here the book would seem to side with the first view of within-case analysis I presented above, in placing the analysis of causal process centrally. But the spatio-temporal contact component is not similarly grounded in the first principles of the ontological approach he develops. As a result, the reader is left a bit unsure about why and to what extent the centrality of causal mechanisms must be taken on board, or whether a broader view of within-case analysis could be consistent with the regularity approach and what Mahoney sees as its scientific constructivist underpinnings.

As the book proceeds, the centrality of causal mechanism to within-case analysis fades gradually away. In Chapter 4, for example, Mahoney discusses the evaluation of descriptive propositions. Here, he notes (2021, 122) that "one can ask about auxiliary traces that would have been left behind if a case had membership in the category of interest." This language of auxiliary traces sounds a bit more like the second approach to within-case analysis in its consideration of a wider range of information about a case. So far, though, Mahoney seems to want to reserve this approach for within-case analysis for descriptive propositions—his immediately subsequent discussion of causal propositions still emphasizes the primary place of causal mechanisms in evaluating causation in the regularity model.

But by the time we get to Chapter 6, the place of mechanisms is quite sharply downplayed. Here, the set-theoretic approach to sequence analysis does not incorporate mechanisms at all. Here, spatio-temporal contiguity is gone from the discussion of how we compare the importance of multiple causes. Chapter 7 on

[1] Much like Mahoney's book, I limit my discussion to approaches that are broadly positivist in nature, and that seek to achieve some element of explanation, rather than those oriented toward predictions or description.

[2] To put my own cards on the table, I tend toward this view as well (see Soifer 2020, 16).

Bayesian set-theoretic analysis unfolds with no distinction between observations about causal mechanisms and other kinds of observations about cases. By this point in the book, then, we've moved to a position quite closely aligned with the second view of within-case analysis I articulated above.

It seems, then, that the book is torn on the same issue that divides qualitative methodologists: should we prioritize evidence of mechanisms as our primary element of within-case analysis, or should the net of evidence we consider be cast more widely? Can we draw conclusions about a cause based on logical regularity alone? Or must the set-based approach that lies at the heart of Mahoney's logic of social science also bring along an emphasis on causal mechanisms? Here, the regularity model of causality is doing a lot of work in holding these two elements of causal appraisal together. In my view, the connection between these two elements is not sufficiently elaborated in Mahoney's book—the book simply states that the regularity model entails both of these elements, but the causal mechanism component, not being grounded in ontological first principles, feels like it has an ambiguous logical status in the framework of causal appraisal Mahoney proposes.

All this is intended not as a criticism of the book, but as a two-fold invitation. First, I encourage Jim and others to develop further for us the logic of the regularity model and to spell out how both of its components are necessary, and how they can be derived from the underlying ontology from which his logic of social science departs. Second, I invite all of us to make explicit the tension between the two views of within-case analysis that I have articulated here, and to engage with one another on this crucial issue rather than eliding it with the anodyne label of "process tracing" that can mean different things to different people. On this issue, and many others, *The Logic of Social Science* generates much food for thought and great payoffs to substantive engagement. I expect we'll be discussing many aspects of the book for some time to come.

References

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

Collier, David, Brady, Henry E. and Jason Seawright. 2004. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology." In *Rethinking Social Inquiry* edited by Collier, David and Herny E. Brady, 229-266. Lanham: Rowman & Littlefield.

Fairfield, Tasha, and Andrew Charman. Forthcoming. *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*. Cambridge: Cambridge University Press.

Humphreys, Macartan, and Alan M. Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109, no. 4 (November): 653–73. https://doi.org/10.1017/S0003055415000453

Mahoney, James. 2021. *The Logic of Social Science*. Princeton: Princeton University Press.

Soifer, Hillel David. 2020. "Shadow Cases in Comparative Research." *Qualitative and Multi-Method Research* 18, no. 2 (Fall): 9-18. https://doi.org/10.5281/zenodo.4046562

Tilly, Charles. 1996. "Invisible Elbow." *Sociological Forum* 11, no. 4 (December): 589–601. https://doi.org/10.1007/BF02425305

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448199

Author's Response: The Logic of Social Science and Contemporary Political Science

James Mahoney
Northwestern University

I would like to thank the five commentators in this symposium (Jennifer Cyr, Gary Goertz, Alan M. Jacobs, Carsten Q. Schneider, Hillel David Soifer) for their engagement with and thoughtful discussions of *The Logic of Social Science* (*LSS*). Their comments focus mainly on Part I (Ontology and Epistemology) and Part II (Methodological Tools) of the book, and I will also concentrate on these parts. For interested readers, Part III (Explanatory Tools) concerns theory building and formulating explanations.

Causal Heterogeneity: Ubiquitous and Inscrutable

LSS develops a positive argument about a new scientific-constructivist approach for social science research; it is mainly concerned with formulating tools to implement this approach in substantive research. To motivate the approach, however, the book argues in chapter 3 that unrecognized heterogeneity is a serious problem for counterfactual theories of causality in the social sciences (e.g., Rubin 1974; Holland 1986; Woodward 2003; Morgan and Winship 2015). *LSS* asserts that this heterogeneity is ubiquitous and inscrutable. According to *LSS*, unrecognized heterogeneity may well explain why social scientists have found it so difficult to generate stable and sound causal inferences using counterfactual theories of causality (unlike epidemiologists and natural scientists who study approximate natural kinds). *LSS* proposes that the solution is not to try to model this heterogeneity (because it is inscrutable), but to accept it as a basic part of what it means to study social categories.

LSS links causal heterogeneity to a referential mismatch between social categories (e.g., *democratic regimes*, *development*) and natural kinds (e.g., *sodium salts*, *ionization*). Causation occurs at the level of natural kinds, but this causation is disconnected from and not captured by our social science categories. Social science categories are mind-dependent entities in the sense that their existence as categories depends on individuals sharing an understanding of their meaning (von Wright 1971; Searle 1995). In this respect, social categories are different from natural kinds, whose existence is not dependent on human beings (or at least far less dependent on human beings) (Churchland 1985; Ellis 2001; Miller 2000). For example, the events that we call revolutions do not exist in the world in the same way that the chemical element copper exists in the world. Copper has certain properties (e.g., its atomic structure) and certain causal powers (e.g., copper is an electrical conductor) independent of human beliefs; copper possesses these properties and powers in an identical form across all human societies regardless of their specific beliefs and values.

The natural substances and properties that compose a social category are not homogeneous across the members of the category. Instead, at the level of their natural kind composition, social science categories are heterogeneous. For example, while all revolutions are constituted in part by hydrogen, oxygen, and other elements, the key defining similarities shared by all revolutions cannot be reduced to natural kinds. Revolutions do not have a one-to-one correspondence (either in the sense of bijection or surjection) with a particular set of defining natural substances and properties in the external world. Instead, certain entities are revolutions because (and insofar as) we share an understanding of the meaning of the category *revolution*. The categories we use to define *revolution* are social categories themselves that depend equally on our minds for their existence. Our definitions of social categories do not refer to or bottom out with natural kinds. Social categories are mind-dependent all the way down.

LSS briefly discusses the consequences of inscrutable causal heterogeneity for counterfactual theories of causality in terms of a violation of the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1974). These consequences could also be discussed in terms of massive unrecognized subgroup heterogeneity, which affects both experimental and observational research. In his commentary, Jacobs neither endorses nor disputes the argument that unrecognized heterogeneity is a serious problem for scholars who seek to make valid causal inferences using a counterfactual theory of causality. Instead, his approach is to argue that a regularity theory of causality, which *LSS* advocates, may be equally vulnerable to this problem.

Psychological Essentialism Disguises Category Heterogeneity

LSS argues that social scientists (like all human beings) experience the bias of psychological essentialism. Psychological essentialism is a human disposition in which we believe that the members of a category share underlying essences that endow them with an identity and a predictable nature. The scientific evidence in support of the proposition that human beings engage in psychological essentialism is extensive and convincing (see Gelman 2003; Newman and Knobe 2019). This bias causes us to perceive heterogeneous natural entities as homogeneous social entities. We are psychologically disposed to overlook heterogeneity among the members of a given social category.

Our psychological essentialism is highly functional and probably necessary for the existence of social institutions and human civilization. Because we are usually not aware that social institutions are dependent on our beliefs for their existence, we tacitly accept those institutions as basic facts about the world—we experience them as objective reality (Berger and Luckmann 1966). Psychological essentialism also underpins our capacity to make useful generalizations about social categories and their relationships (see Gelman 2003, 27-43). This orientation provides a basis for inductive inference: all entities of the same kind have similar natures because they share essential properties. Social scientists follow psychological essentialism when they understand regularities in terms of the efficacious properties possessed by social categories.

Yet the commonness and the utility of a psychological orientation do not establish its truth (Dennett 1987). Understanding reality often depends on departing from our commonsense orientations, helpful as they otherwise may be. *LSS* develops a scientific approach that aims to allow researchers to make inferences about social categories without engaging in psychological essentialism.

Why Category Heterogeneity Does Not Raise the Same Problem for a Regularity Theory of Causality

A regularity theory of causality avoids problems arising from category heterogeneity by assuming that causality is a logical and spatiotemporal relationship among social categories (see Mahoney and Acosta 2021). A regularity theory does not assume that a change on a causal factor will produce (probabilistically) any net change on the outcome—it rejects the basic starting point of a counterfactual theory of causality. A regularity theory also rejects the idea that causal factors are efficacious entities in possession of inherent causal powers. A regularity theory leaves the question of why a regularity exists "unexplained"; the theory does not require an account of why the regularity exists in order for the relationship to qualify as a causal relationship (Beebee 2006).

Nevertheless, a regularity theory sets up demanding criteria for a relationship to qualify as a causal relationship. It proposes that causality exists between social category *X* and social category *Y* if three conditions obtain: (1) temporal succession (*X* precedes *Y* in time), (2) spatiotemporal contiguity (*X* makes direct or indirect contact with *Y* in space and time), and (3) logical regularity (*X* is part of the fully minimized solution set that is constantly conjoined with *Y*).

The second component of this definition differentiates a regularity theory from counterfactual dependence theories that do not require as a matter of definition that cause and outcome be connected in space and time. Regularity theorists meet this requirement by focusing on the causal chain that connects a cause to an outcome (e.g., Glennan 2009; Hedström and Swedberg 1998; Mayntz 2004). With a regularity theory, each link in the causal chain is itself a regularity among categories; causation among temporally separated categories is a series of regularities that unfold over time. Thus, as Soifer correctly notes, the idea of a regularity is prior to the idea of a mechanism in this theory. However, a regularity theory insists on the identification of mechanisms—defined as intervening regularities—to demonstrate causality.

With the third component, the analyst identifies a fully minimized solution set consisting of all conditions and/or combinations of conditions that are *sufficient* for the outcome. A solution set is *fully minimized* if all redundancies are removed from both necessary conditions and sufficient conditions. This solution set is sometimes referred to as consisting of "minimally necessary disjunctions of minimally sufficient conditions" (Baumgartner 2008, 23). Every individual condition makes a difference to at least one aspect of the explanation of the outcome. Thus, every individual condition in the solution set is a *cause* of the outcome (assuming the other two criteria are in place). The need for solution sets that do not contain any redundant conditions connects a regularity theory to QCA and other methodologies that use logical minimization techniques to remove non-essential conditions and arrive at parsimonious solution sets (see Baumgartner 2008, 2013; Graßhoff and May 2001; Ragin 2008; Schneider 2018; Schneider and Wagemann 2012; Oana, Thomann, and Schneider 2021). In principle, these parsimonious solution sets weed out all spurious factors that do not play a role in the explanation of the outcome.[1]

This approach to causality might seem to be an inferior option when compared to sophisticated counterfactual theories of causality. Yet these sophisticated theories are always premised on the idea that variable values are homogeneous across cases. They always assume that a given unit change on a variable represents the same basic occurrence across cases. If we reject these assumptions for social categories, as I fear we must, we need to seriously reconsider the utility of counterfactual theories of causality for the analysis of social categories. As currently formulated, counterfactual theories of causality are appropriate for the study of natural kinds but not for most of the phenomena studied by social scientists.

A regularity theory of causality is appropriate for the analysis of social categories that refer to heterogeneous natural entities. For instance, consider the following Boolean solution set: $AB \vee CD \rightarrow Y$, where $\vee$ is the Logical OR and → is sufficiency. Let us assume that condition *A* is heterogeneous in the following way: $X \vee Z \rightarrow A$. The fact that condition *A* is heterogeneous in this way does invalidate or even raise any special concerns about the validity of the original solution. Certainly, we can rewrite the original equation to highlight the heterogeneity (i.e., $[(X \vee Z)\ \&\ \mathrm{B}] \vee CD \rightarrow Y$), but it is not necessary to do so to preserve validity. With a regularity theory of causality, findings are stable as one moves from the full population to subsets of cases. The original finding $AB \vee CD \rightarrow Y$ remains stable and applicable regardless of whether one looks at only cases with *X*, only cases with Z, or any other subset of cases. With a regularity theory, the researcher does not need to model or even know about subgroup heterogeneity when specifying a causal model.

The upshot is that a regularity theory of causality has significant advantages over a counterfactual theory in a context of massive *unrecognized* (and indeed *unrecognizable*)

[1] The ability of QCA and other large-N set-theoretic approaches to identify non-spurious regularities in the face of data with limited diversity is the topic of debate both within the set-theoretic community and between set-theoretic researchers and their critics (see Thomann and Maggetti 2020 for a literature review).

heterogeneity. Whereas this underlying heterogeneity does not affect the validity of findings using a regularity theory of causality, it is potentially devastating for the validity of findings using a counterfactual theory of causality. *LSS* provocatively suggests that these problems of heterogeneity explain why talented social scientists who use a counterfactual theory of causality with social categories (as opposed to categories corresponding to natural kinds) encounter great difficulty generating stable and consistent results that are widely accepted as true.

Note that the differing tolerance of a regularity theory vs. a counterfactual theory for unrecognized heterogeneity derives from their alternative conceptions of causality: a regularity theory sees causality as a logical and spatiotemporal pattern that exists between categories; a counterfactual theory sees causality as the "difference-making" effects of variables net of everything else. To be sure, a regularity theory of causality faces many very serious challenges to generating valid inferences in practice. As Schneider notes, the use of this theory of causality does not ensure that the analyst has correctly identified all important causal factors.[2] The argument of *LSS* is that a regularity theory of causality is appropriate as an *understanding and definition* of causality for research in the social sciences. It suggests that the same is not true of a counterfactual theory of causality.

Within-Case Analysis

Goertz is correct that most of the methodological tools developed in Part 2 in *LSS* focus on the within-case analysis of individual cases. I certainly hope that this fact does not kill my beloved comparative-historical analysis! After all, comparative-historical researchers depend primarily on within-case analysis for their causal arguments. As Goertz and I wrote in our *Two Cultures* book:

> In small-N qualitative research, the main leverage for causal inference derives from within-case analysis, with cross-case methodologies sometimes playing a supporting role.
>
> In large-N statistical research, the main leverage for causal inference derives from cross-case analysis, with within-case methodologies sometimes playing a supporting role. (2012, 88)

If *LSS* had focused on large-N analysis, it would have had to say much more about QCA and other cross-case methods. But the book explicitly focuses on case study and small-N methods. *LSS* combines "possible worlds" semantics, counterfactual analysis, and set-theoretic analysis as tools for analyzing regularities when only one case is under study. The book shows how assertions about causation necessarily invoke possible or counterfactual cases. By explicitly weighting these possible cases, the book shows how one can use a regularity theory of causality even when only a single actual case is under study.

The interesting approach described by Goertz in his commentary is a medium to large-N set-theoretic method that uses a regularity theory of causality. It resembles the method of analytic induction, as described by Charles Ragin in a new draft book manuscript. I note that while Goertz separates a regularity theory from a focus on mechanisms, this separation is not technically correct: Hume was clear that the analysis of causal links is an essential part of a regularity theory of causality (see Mahoney and Acosta 2021). Schneider points out that many scholars have lost sight of the importance of spatiotemporal connection as a defining component of a regularity theory of causality. I place some of the blame for this misunderstanding on Hempel's (1942) covering law model of explanation, which does not require any spatiotemporal connection between cause and outcome. The conflation of a regularity theory of causality with the covering law model of explanation is unfortunate.

In his commentary, Soifer raises important questions about the relationship between the specification of causal chains and the use of specific observations for evaluating propositions. He points out that methodologists who work on process tracing are pulling in two separate directions: (1) the study of causal chains, intermediary mechanisms, and causal flow processes; and (2) the identification of specific observations that carry substantial weight in the assessment of explanations—regardless of whether these observations are intermediary mechanisms. Soifer is correct that earlier discussions of process tracing often focused on (1) (e.g., George and Bennett 2005), whereas more recent discussions of process-tracing tests often emphasize (2) (e.g., Fairfield and Charman 2017). Soifer inquires about the relationship between these two directions, asking whether they are competing, separate, complementary, or even essential for each other.

To answer Soifer, I believe that the two kinds of process tracing are complementary and must be joined together to adequately assess causal arguments (see also Beach and Pedersen 2013). When using a regularity theory of causality, it is essential to pursue the study of causal chains and intermediary mechanisms. A good causal argument links causal factors to an outcome across time and space. One of the reasons why I find comparative-historical analysis so compelling is its orientation toward

[2] Going forward, machine learning and computation techniques could be used with QCA to select potential causal factors. The two methodologies nicely complement one another: computational methods help identify potential causal factors, and QCA methods remove redundancies to arrive at parsimonious solution sets (thanks to Qin Huang of Northwestern University for this insight).

sequential arguments that connect historical causes to more contemporary outcomes. At the same time, I believe that scholars must support their arguments about causal chains through the use of specific observations that have probative value in adjudicating among rival explanations. To evaluate the proposition that *X* causes *Y*, one normally asks questions about the intermediary events that should be observed (or should not be observed) if this proposition is true. In identifying these intermediary events, one simultaneously creates a causal chain argument and locates a critical observation for evaluating the truth of the proposition (see chapter 4 of *LSS*).

Set-Theoretic Analysis as a Constructivist Approach

LSS argues that a constructivist approach to social categories is essential for the social sciences. The book specifically recommends treating categories as sets in which other categories (also sets) can have membership, no membership, or partial membership.[3] The book contends that these sets are ultimately located in human minds as conceptual spaces (cf. Gärdenfors 2000; 2014). This approach creates an ontology in which social categories are inherently mind-dependent entities. While categories make reference to objective natural kinds in the world, those natural kinds are heterogeneous in their composition for any given social category. The one thing that all members of a social category have in common is their shared activation of conceptual spaces corresponding to the category within human minds.

Constructivist set-theoretic analysts do not arbitrarily categorize entities in the social world. Rather, these scholars establish boundaries and membership values on the basis of the *meanings* of social categories within one or more communities or societies. Constructivist set-theoretic analysts use a broadly *interpretive approach* to elucidate the meaning of categories within particular communities. Interpretation is needed for calibrating categories and for coding whether specific cases are members, non-members, or partial members of categories. The quest to understand the meaning of social categories in specific contexts often requires expert knowledge of the relevant communities and societies. The interpretive aspects of constructivist set-theoretic analysis link this approach to qualitative data collection techniques such as ethnography and interviews.

Yet the ultimate goal of constructivist set-theoretic analysis is not primarily interpretive; researchers do not stop with an analysis of meaning structures and semiotics. Instead, constructivist set-theoretic researchers ultimately seek to make generalizations about regularities that objectively exist among social categories within particular communities. Some generalizations concern combinations of conditions that are nearly sufficient for an outcome. Other generalizations concern categories that are important INUS conditions for an outcome, such as conditions that are frequently necessary and somewhat sufficient for an outcome. Constructivist set-theoretic analysts not only seek to discover regularities; they also use knowledge of regularities to explain occurrences in specific cases. For example, they may draw on the preexisting knowledge that membership in social category *X* is nearly necessary for membership in social category *Y* in order to explain why outcome *Y* occurred in an individual case.

Schneider asks where constructivism ends and formal logic takes over in constructivist set-theoretic analysis. The answer is that constructivism ends with the constitution and coding of categories and the creation of scope boundaries At this point, the logical machinery of set-theoretic analysis is used for the objective assessment of propositions and the discovery of regularities among categories. As the book makes clear, I treat first-order logic as an objective feature of reality that is essential for valid inference and reasoning. I explicitly reject radical constructivist views that see logic itself as a social construction. I believe constructivist set-theoretic analysts need to pay attention to robustness tests, the properties of algorithms, and other technical aspects involved in the scientific assessment of relationships among social categories. Schneider is right that students tend to struggle with formal logic, and I believe that logic (ideally with set theory) should be a basic component of graduate training in political science.

[3] Schneider argues that the label *continuous-set analysis* is problematic because it does not privilege the membership value of 0.5, which is the point of maximum ambiguity (the same could be said of the label *fuzzy-set analysis*). Schneider is correct that I believe the values of 0 and 1 are the qualitative anchors (see Wolff 2020 on the qualitative-quantitative distinction). And I am sensitive to the fact that a 0.5 threshold is essential for important QCA procedures; I agree that 0.5 is an *extremely useful* threshold for the purposes of substantive analysis. However, I think the label *fuzzy* is a disaster for set-theoretic analysis: the label is deeply misleading about category boundaries, which are sharp and bright and not at all blurry or hazy. There is nothing fuzzy about continuous-set measurement. I considered the label *permeable-set analysis*, but ultimately went with the more attractive *continuous-set analysis*. I hope some others will do the same. I originally encountered the label "continuous-set analysis" when reading McNeil and Freiberger (1993, 30).

Let me use this footnote to make one more point regarding Schneider's excellent comments: I disagree that a case with 0.3 membership in *tall person* should be classified as a tall person. That person is slightly tall or a little tall, such as a woman of 5 feet and 7 inches in the United States. Note that this person will have no membership in the category *short person*. The person has 100% membership in *slightly tall person* and 0% membership in *short person*.

The kind of constructivism endorsed in *LSS* differs from the constructivism endorsed by Alexander Wendt in his seminal *Social Theory of International Politics* (1999). Whereas Wendt adopts a critical realist position in which social categories are in part self-organizing entities (72-77), I embrace an experiential realist position in which social categories are not self-organizing entities.[4] A social category certainly refers to "physical" entities in the world (i.e., natural kinds), but these entities are heterogeneous in their composition; they require human minds to make them members of a given social category (von Wright 1971; Searle 1995; Reed 2008). Unlike Wendt, then, I argue that social categories are dependent on human minds for their existence *as particular categories* at all levels of analysis. Without human beliefs, we are left with natural kinds that do not group together in ways that overlap with our social categories. In this sense, I offer a stronger version of constructivist ontology than does Wendt in his magnificent *Social Theory of International Politics*.

The Science in Social Science

Science is an epistemology that consists of general and public procedures rooted in logic for using evidence to derive beliefs about the truth of propositions concerning the actual world. Constructivism is an ontology in which a social category is understood to be a mind-dependent entity; a social category refers to natural entities in the world, but those natural entities require human minds to become categories. *LSS* calls for an approach to social science that is both scientific and constructivist—that is, *scientific constructivism*.

Scientific constructivism is not common because many constructivists are skeptical about science (as conventionally defined), whereas many scientifically inclined social scientists are skeptical about constructivism. The mistake of many constructivists is to reject logic as subjective and optional; this mistake leads them to at times fall off the epistemological ledge into relativism about truth and reality. By contrast, the mistake of scientifically inclined social scientists is to believe that essentialism is appropriate for the study of social reality; this mistake leads them to reify social categories and to work under false assumptions.

LSS proposes set-theoretic analysis as a scientific-constructivist approach. With this approach, categories are sets, and sets are located in the mind, existing ontologically prior to the entities they classify. The boundaries of a set determine whether entities are members of the set; the properties of the entities do not determine the boundaries of the set. Membership boundaries can shift without any changes at all in the properties of the entities. Entities are similar or different *because of* their set membership. *LSS* shows how this approach to sets offers a constructivist alternative that avoids essentialism, recognizes the mind-dependent nature of categories, and lends itself quite naturally to scientific research (see chapter 2 of *LSS*).

The fact that social categories are dependent on subjective beliefs does not mean that social scientists cannot arrive at objective truths about propositions concerning those social categories (to respond to Cyr's concerns). To be sure, propositions using social categories are bound by scope conditions in which the categories carry a specific meaning. However, this dependence on semantic context does not make the truth of propositions relative to particular places and times. Instead, it makes the existence of the propositions themselves relative to particular places and times. Outside of certain context, a given proposition carries a different meaning, and thus it is not the same proposition.

Social scientists can never be *certain* about the truth of a proposition. However, this uncertainty is not an artifact of a relativism distinctive to the study of mind-dependent categories. Rather, uncertainty is inherent to scientific findings in general (Popper 1934/1968). Science can only deliver approximate truth or highly likely truth, not absolute truth. Well-formulated propositions are either 100% true or 100% false, but our *certainty* about whether they are true or false is never 100%.

Scientific Constructivism, Now and in the Future

Cyr asks about the reception of scientific constructivism within political science. Who will embrace the overall approach of the book? Who will follow *LSS* by explicitly using constructivist set-theoretic analysis? These questions are appropriate given a context in which: (1) a counterfactual theory of causality is almost universally embraced in quantitative political science; (2) set-theoretic analysis is at best viewed with suspicion and at worst dismissed in quantitative political science; and (3) constructivists are on the margins of quantitative political science. *LSS* argues that all three of these trends are unfortunate, but that does not make them any less serious as obstacles.

In the short run, set-theoretic researchers who already see themselves as employing interpretive analysis are good candidates for the explicit use of scientific constructivism. Likewise, constructivist researchers in IR with scientific leanings are good candidates; in fact, many of these scholars may already see themselves as scientific constructivists though not under that label. Likewise, many qualitative researchers are already employing set-

[4] Like Wendt (1999), I adopt a scientific realist position regarding *natural kind categories*. We both believe that a structured reality exists independent of all human observers, and we believe that natural science is successful because it at least partially models this reality.

theoretic tools in an implicit way (Goertz and Mahoney 2012); these scholars can use specific methodological and theoretical tools of *LSS* even if they do not identify with scientific constructivism. With respect to quantitative and experimental political scientists, I hope that the arguments about essentialism, heterogeneity, and set-theoretic analysis might merit some consideration and discussion. Quantitative scholars have much to give to the field of large-N set-theoretic analysis with respect to designing new tools and helping solve problems. These positive interventions would also allow qualitative researchers to feel more comfortable using set-theoretic analysis explicitly in their work. On the flip side, I think the ideas that animate set-theoretic analysis could be productively incorporated into a new potential outcomes framework that is connected to a regularity theory of causality rather than a counterfactual theory of causality (as Jacobs' comments hint at). This framework would require the efforts of our most talented quantitative methodologists.[5]

Cyr and I disagree on one point: the difficulty of pursuing scientific-constructivist work. I think she overstates this difficulty for many qualitative researchers. The main requirement of this kind of work is the treatment of social categories as continuous sets that exist in the mind and that are used to classify heterogeneous entities in the natural world. This approach is not totally different from what qualitative political scientists in the field of comparative-historical analysis are already doing (including Cyr herself!). If one accepts the argument that qualitative researchers are often "closet" set-theoretic analysts (at least some of the time), then it is not operationally difficult for these researchers to explicitly treat their categories as mental sets (at least some of the time). I do not think the move to an explicit scientific-constructivist approach involves a revolutionary new way of thinking about social reality for these particular scholars. And the advantages of explicitly conducting scientific-constructivist research are considerable: more valid and transparent conclusions by virtue of practicing better science through the self-conscious application of general rules and procedures.

The legacy of *LSS* will depend on how scholars react to its central arguments (or do not react), and how they choose to use (or not) its methodological and theoretical tools. If the book were to contribute to some larger reorientation toward scientific constructivism in political science, it would do so because it codifies principles and methods that some scholars in the discipline already embrace and use.

References

Baumgartner, Michael. 2008. "Regularity Theories Reassessed." *Philosophia* 36, no. 3 (September): 327-54. https://doi.org/10.1007/s11406-007-9114-4

Baumgartner, Michael. 2013. "A Regularity Theoretic Approach to Actual Causation." *Erkenntnis* 79, no. 4 (December): 85-109. https://doi.org/10.1007/s10670-013-9438-3

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

Beebee, Helen. 2006. "Does Anything Hold the Universe Together?" *Synthese* 149, no. 3 (April): 509-33. https://doi.org/10.1007/s 11229-005-0576

Berger, Peter L., and Thomas Luckmann. 1966. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. New York: Doubleday.

Churchland, Paul M. 1985. "Conceptual Progress and Word/World Relations: In Search of the Essence of Natural Kinds." *Canadian Journal of Philosophy* 15, no. 1 (March): 1-17. https://doi.org/10.1080/00455091.1985.10716405

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge: MIT Press.

Ellis, Brian. 2001. *Scientific Essentialism*. Cambridge: Cambridge University Press.

Fairfield, Tasha, and Andrew Charman. 2017. "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats." *Political Analysis* 25, no. 3 (July): 363-80. https://doi.org/10.1017/pan.2017.14

Fairfield, Tasha, and Andrew Charman. Forthcoming. *Social Inquiry and Bayesian Inference: Rethinking Qualitative Research*. Cambridge: Cambridge University Press.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge: MIT Press.

Gärdenfors, Peter. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambrdige: MIT Press.

Gelman, Susan A. 2003. *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford: Oxford University Press.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press.

[5] I see the potential outcomes framework as separate from a counterfactual or interventionist understanding of causality. I think much can be gained by finding synergies between possible world semantics as used in set-theoretic analysis and the potential outcomes framework as used in quantitative political science.

Glennan, Stuart. 2009. "Mechanisms." In *The Oxford Handbook of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Peter Menzies, 315-25. Oxford: Oxford University Press.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton: Princeton University Press.

Graßhoff, Gerd, and Michael May. 2001. "Causal Regularities." In *Current Issues in Causation*, edited by Wolfgang Spohn, Marion Ledwig, and Michael Esfeld, 85-114. Münster: Mentis Verlag.

Hedström, Peter, and Richard Swedberg, eds. 1998. *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.

Hempel, Carl G. 1942. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: Free Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, no. 396 (December): 945-60. https://doi.org/10.1080/01621459.1986.10478354

Mahoney, James, and Laura Acosta. 2021. "A Regularity Theory for the Social Sciences," *Quality and Quantity* https://doi.org/10.1007/s11135-021-01190-y.

Mayntz, Renate. 2004. "Mechanisms in the Analysis of Social Macro-Phenomena." *Philosophy in the Social Sciences* 34, no. 2 (June): 237-254. https://doi.org/10.1177/0048393103262552

McNeill, Daniel, and Paul Freiberger, 1993. *Fuzzy Logic: The Revolutionary Computer Technology that is Changing Our World.* New York: Touchstone.

Miller, Richard W. 2000. "Half-Naturalized Social Kinds." *Philosophy of Science* 67 (September): S640-S652. https://doi.org/10.1086/392852

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, second edition. Cambridge: Cambridge University Press.

Newman, George E., and Joshua Knobe. 2019. "The Essence of Essentialism." *Mind and Language* 34, no. 5 (November): 585-605. https://doi.org/10.1111/mila.12226

Oana, Ioana-Elena, Carsten Q. Schneider, and Eva Thomann. 2021. *Qualitative Comparative Analysis (QCA) with* R. Cambridge: Cambridge University Press.

Popper, Karl. 1934/1968. *The Logic of Scientific Discovery*. New York: Harper and Row.

Ragin, Charles C. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Reed, Isaac. 2008. "Justifying Sociological Knowledge: From Realism to Interpretation." *Sociological Theory* 26, no. 2 (June): 101-29. https://doi.org/10.1111/j.1467-9558.2008.00321.x

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66, no. 5: 688-701. https://doi.org/10.1037/h0037350

Schneider, Carsten Q. 2018. "Idealists and Realists in QCA." *Political Analysis* 26, no. 2 (April): 246-54. https://doi.org/10.1017/pan.2017.45

Schneider, Carsten Q., and Claudius Wagemann. 2012. *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

Searle, John R. 1995. *The Construction of Social Reality*. New York: Free Press.

Thomann, Eva, and Martino Maggetti. 2020. "Designing Research with Qualitative Comparative Analysis (QCA): Approaches, Challenges, and Tools." *Sociological Methods and Research* 49, no. 2 (May): 356-86. https://doi.org/10.1177/0049124117729700

von Wright, Georg Henrik. 1971. *Explanation and Understanding*. Ithaca: Cornell University Press.

Wendt, Alexander. 1999. *Social Theory of International Politics*. Cambridge: Cambridge University Press.

Wolff, J. E. 2020. *The Metaphysics of Quantities*. Oxford: Oxford University Press.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

QMMR Qualitative & Multi-Method Research

Longform APSA Awards (2021)

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6478359

Giovanni Sartori Award for Best Book on and/or using Qualitative Methods

Committee: Nicholas Rush Smith, CUNY (Chair), Gwyneth McClendon, NYU, Erik Martínez Kuhonta, McGill University

2021 Recipients

Diana S. Kim, *Empires of Vice: The Rise of Opium Prohibition across Southeast Asia* (Princeton: Princeton University Press, 2020)

Diana Kim's *Empires of Vice* is a phenomenal book that exemplifies the rich tradition of deep, historical work in political science. Through a comparative study of late colonial empires in Southeast Asia, Kim charts the efforts of state bureaucrats to bring to an end the extensive consumption of opium. Rather than drawing out a macro theory of state behavior, *Empires of Vice* pays close attention to the ideas, doubts, values, and anxieties of local civil servants who puzzled through the question of how to regulate opium consumption. What emerges is a very nuanced and complex picture of the underbelly of state bureaucracies, where a messy process of questioning and re-thinking the role of opium in society leads slowly to significant change. Kim finds that different concerns shaped the actions of local bureaucrats: anxieties over the moral effects of opium in British Burma, problems of fiscal dependency on opium in British Malaya, and illicit financial practices and accumulating debt in French Indochina. These concerns emerged gradually from the trenches, but in the long run shaped the views of higher officials and altered colonial policy regarding opium.

Kim has delved deeply into a wide range of colonial archives to explain how and why states do what they do. But she has done more than that: Kim has given meaning to the thoughts and everyday work of lower-level bureaucrats. It is this analytical contribution – of interpreting carefully actors' beliefs, writings, and action; of granting value to local officials who are often ignored or misunderstood; and then of structuring these actors' discourses within a broad, comparative-historical framework – that makes this study so outstanding.

Elegantly written and deeply researched, *Empires of Vice* is an original, compelling, and beautiful book that makes a powerful contribution to the fields of comparative-historical analysis, qualitative methods, and to political science.

Devorah S. Manekin, *Regular Soldiers, Irregular War: Violence and Restraint in the Second Intifada* (Ithaca: Cornell University Press, 2020)

Why do soldiers engage in violence against civilians? In answering this enduringly important question through analysis of Israel's counterinsurgency warfare in Palestine, Devorah Manekin's *Regular Soldiers, Irregular War: Violence and Restraint in the Second Intifada* is an impressive example of carefully crafted fieldwork in a highly sensitive context that will serve as a model for future researchers on violence. Manekin builds her analysis on nearly two years of fieldwork in Israel, in-depth interviews with approximately seventy former Israeli soldiers, and an online survey of combat veterans to understand why some units engaged in violence against civilians while other units refrained. Conceptually, Manekin develops a novel typology of violence to show how some violence against civilians is directed by military leadership, while other violence is entrepreneurial (violence intended to help the military mission, although not explicitly ordered by leadership) or opportunistic (violence with no clear military purpose). Her fieldwork reveals that both the level and form of violence deployed by units against civilians could best be explained by processes of organizational control, including unit socialization, a shared identity among soldiers, and the quality of leadership. The arguments have major implications for our understanding of patterns of violence during warfare, why soldiers engage in violence that violates the laws of war, how such violence can be reduced, and who bears responsibility for such violence.

In making these arguments, *Regular Soldiers, Irregular War* is a model of careful qualitative scholarship in a violent and politically sensitive context. Specifically, in a careful discussion of her methodological approach (Chapter 2), Manekin weighs the costs and benefits of understanding soldiers' accounts of violence against civilians, including the difficulty of assessing the veracity of soldiers' accounts, the possibility of legal implications or reprisal from commanders against combatants, and concerns that paying attention to soldiers' accounts of violence may unintentionally occlude victims' experience of it. In doing so, she attends to the fact that the meaning of violence often shifts during a conflict, in its wake, and even during the span of an interview and that these shifts in meanings have consequences for how violence is understood. Making matters more complicated, Manekin smartly notes that there are strong incentives for actors and authorities to conceal violence against civilians and that complex ethical obligations attach when investigating violence around subjects' ability to consent, subjects' privacy, and subjects' potential trauma stemming from having participated in violence. In carefully attending to how such complexities affect the study of a sensitive topic, *Regular Soldiers, Irregular War* is an example of superlative qualitative work.

2021 Honorable Mention

Janet I. Lewis, *How Insurgency Begins: Rebel Group Formation in Uganda and Beyond* (Cambridge: Cambridge University Press, 2020).

Janet Lewis' *How Insurgency Begins,* to which the committee awards an Honorable Mention, is a masterful study of how and when rebel groups are born. Lewis invites us to step back from the well-studied questions of when and how violent conflict with the state begins and interrogate the prior questions of how, and under what conditions, rebel groups are able to form in the first place. Set largely in Uganda, *How Insurgency Begins* focuses our attention on the role of rumors and uncertainty. Potential rebels who can keep secrets under wraps, and who can build support for their cause through well-placed rumors, are the ones that become viable. Where the state has tight control over the spread of information, by contrast, rebel groups are less likely to arise and develop.

The questions at the heart of Lewis' book are by their very nature difficult to study. Yet, Lewis' book marries innovative theorizing with a field-intensive empirical approach to develop a convincing account. The book offers a sterling example of mixed methods research, combining interviews with former rebels, government intelligence officers, and civilians, with archival work and an original field experiment that seeded benign pieces of information in two villages in rural Uganda in order to understand how such information spreads through social networks. The book breaks new theoretical ground, all while forcing us to rethink large-N data around rebellion and conflict onset and to see in stark relief the value of deep fieldwork.

Alexander George Award for Best Article or Book Chapter on and/or using Qualitative Methods

Committee: Ezequiel Gonzalez-Ocantos, University of Oxford (Chair); Martha Wilfahrt, UC Berkeley; and Rana Khoury, Northwestern University.

2021 Recipient

Emily Kalah Gade, "Social Isolation and Repertoires of Resistance," *American Political Science Review,* Volume 114, Issue 2, May 2020, pp. 309-325. DOI: https://doi.org/10.1017/S0003055420000015

Social Isolation and Repertoires of Resistance is a fascinating study of modalities of resistance in the West Bank. Why do some civilians support individual, often violent, resistance whereas others favor group efforts? Gade's answer is truly original: the architecture of occupation shapes repertoires of contention. In communities that live amidst a checkpoint infrastructure that isolates individuals and breaks social connections, hopelessness is pervasive, leading civilians to eschew collective action. By contrast, where checkpoints do not break intra-community ties in this way, civilians still have faith in the possibility of enacting change through coordinated modes of resistance. This intriguing proposition linking spatial fragmentation, emotions, and the dynamics of contention, is the product of inductive theorizing based on a well-crafted comparison of two communities with contrasting resistance profiles, where Gade conducted 71 interviews and life histories. Interview evidence is masterfully blended with contextual analysis to produce rich and compelling narratives. The article (and its methodological appendix) represents the gold standard of this kind of research, exuding both methodological rigor and ethical awareness. It is a must read for scholars and students interested in conducting similar qualitative work, for Gade's article is a testament to the value of deep and involved fieldwork for theory building and a prime example of how to collect and report interview data.

Kendra Koivu Award for Best 2019 APSA Paper on and/or using Qualitative Methods

Committee: Veronica Herrera, UCLA (Chair); Daniel Mattingly, Yale University; and Chloe Thurston, Northwestern University.

2021 Recipient

Sara Niedzwiecki and Jennifer Pribble, "Reconceptualizing Social Policy Expansion and Retrenchment: South America after the Commodity Boom."

This year's Koivu Paper Award is given to Sara Niedzwiecki (UC Santa Cruz) and Jennifer Pribble (University of Richmond) and their excellent paper, "Re-Conceptualizing Social Policy Expansion and Retrenchment: South America after the Commodity Boom." The authors examine ten presidential administrations in Argentina, Brazil, Chile and Uruguay in order to understand social policy expansion and retrenchment in the post commodity boom era. The paper finds that two types of welfare state change is more politically feasible: expanding existing transfer programs rather than existing social services, and backing retrenchment through covert rather than overt strategies. They differentiate between "easier" and "harder" strategies for social policy change and contribute to a rich welfare state literature. We welcomed the medium-N analysis, the strong conceptualization and careful attention to the political feasibility of difficult reforms, which has significant applicability outside of their empirical cases. Congratulations to Jennifer and Sara.

David Collier Mid-Career Achievement Award.

Committee: Melani Cammett, Harvard University (Chair); Alan Jacobs, University of British Columbia; and Jason Seawright, Northwestern University

2021 Recipient:

Hillel David Soifer, Temple University

Hillel Soifer has achieved distinction in all three areas that the David Collier Mid-Career Achievement Award recognizes: methodological publications, innovative application of qualitative and multimethod approaches in substantive research, and institutional contributions to this area of methodology.

To begin with Hillel's methodological work, Hillel's strong interest in qualitative research methods has been apparent since he attended the Institute for Qualitative and Multi-Method Research in 2004. Hillel has since produced a number of articles and chapters on methodological topics which collectively amount to an impressive body of work.

First, we note his work on critical junctures. His piece "The Causal Logic of Critical Junctures" in *Comparative Political Studies* won the QMMR section's Alexander George Award for the best article or book chapter developing or applying qualitative methods. The article provided a much-needed account of critical junctures, focusing on the distinction between permissive and productive conditions. Permissive conditions set the duration of the juncture, the time during which there is heightened contingency or increased causal possibility. Productive conditions determine the outcome that emerges from the critical juncture. Hillel provides an outstanding discussion, showing how his framework can be used to categorize a broad swathe of political science literature that uses critical junctures.

Hillel has also done important work on conceptualization and measurement. A leading example is his discussion of Michael Mann's concept of infrastructural power in his article "State Infrastructural Power: Conceptualization and Measurement in Empirical Analysis," published in *Studies in Comparative International Development*. In the article, Hillel unpacks infrastructural power, and shows how it can be viewed through three distinct analytical lenses: as the capabilities of the central state, as the territorial reach of the state, and as the effects of the state on society. The article also addresses measurement issues within each of these approaches.

Hillel continues to make impressive methodological contributions, and is currently working on a project investigating how social scientists should select units of analysis, the level of aggregation at which an analysis takes place. While some theoretical frameworks specify a particular unit of analysis that is relevant for operationalizing a theory, for others the choice is more ambiguous or uncertain, leaving researchers with more agency. Hillel notes that studies of aggregate phenomena using data at different levels of aggregation will produce different descriptive patterns, and lend support to divergent interpretations of the causal claims they evaluate. This lack of clarity can lead to studies of the same phenomenon producing different findings for the same variable when conducted at different units of analysis, and can cause scholars to conclude that different causal factors are important. This concern applies across the social sciences, as all of their core concepts could potentially be operationalized at various levels of analysis, and hence analyses based on these prior choices could produce inconsistent inferences. The book that should result from this project will thus serve as a major caution and corrective for a great deal of political science research.

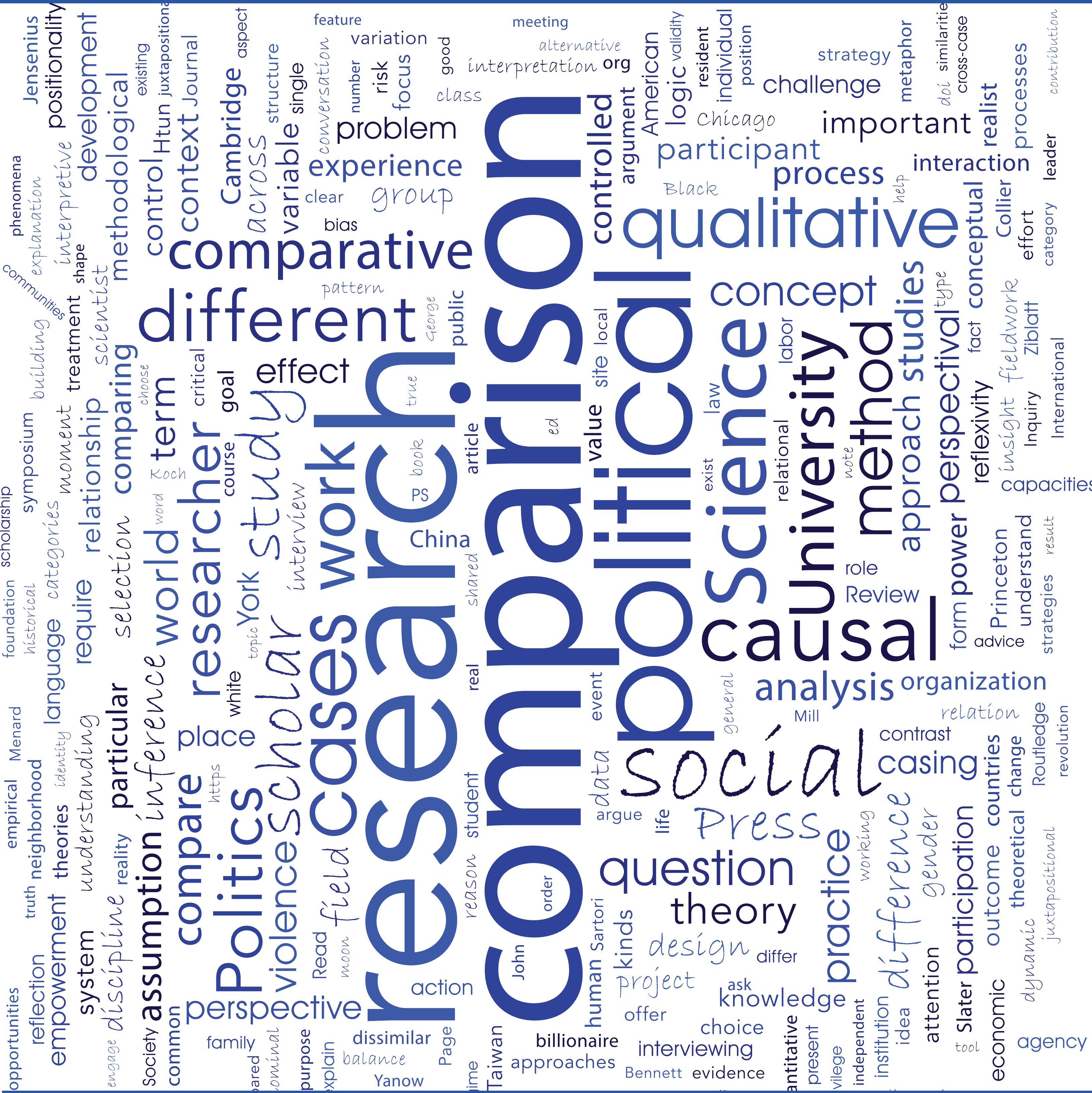
The most substantial piece of Hillel's applied research is his Cambridge University Press book, *State Building in Latin America* (2015). Based on extended and extensive archival research in multiple Latin American countries, the book draws on and skillfully deploys a wealth of historic documents to advance two novel explanations – one for variation in the emergence of state-building efforts (education, taxation, and conscription), and one for variation in their success – in Chile, Colombia, Mexico, and Peru. Described as "magisterial" by Miguel Angel Centeno (Princeton University), and "a model example of comparative historical social science" by Dan

Slater (University of Michigan), two leading voices in this area of scholarship, Soifer's book is a major contribution both substantively and methodologically.

In addition, Hillel has published multiple articles and book chapters focused on state-building, and types and levels of state power, strength, and capacity, in particular in the Andean sub-region of Latin America. From detailed historical analyses of education reform in Bolivia and in Chile developed through the intricate piecing together of meticulously researched historical evidence, to theoretical work with a much broader sweep, Hillel's scholarship is exceptional in its strong methodological foundations, its conceptual clarity, and its clear expression. Moreover, more than many scholars who produce both substantive and methodological work, there are strikingly clear synergies between the two spheres of Hillel's intellectual production: his methodological writings are based on solid research experience, and his research is stronger because he practices what he preaches.

Finally, we note the important institutional contributions that Hillel has made to the qualitative and multi-method research project, most notably in his multiple services to the QMMR section, including as: Annual Meeting program division chair; member of the Alexander George article/book chapter award committee; member of the APSA Qualitative Transparency Deliberations Working Group on Process Tracing and Comparative Case Study Research; chair of the Giovanni Sartori book award committee; and a member of the section's nominating committee.

In sum, it is the Collier Award committee's great pleasure to recognize Hillel's outstanding methodological and institutional contributions to the qualitative and multi-method research community in our discipline.



Qualitative &
Multi-Method
Research