

RESEARCH ARTICLE  

# Which Police Departments Want Reform? Barriers to Evidence-Based Policymaking

Samantha Goerger<sup>1</sup> , Jonathan Mummolo<sup>2\*</sup>  and Sean J. Westwood<sup>3</sup>

<sup>1</sup>Independent Scholar, <sup>2</sup>Department of Politics, School of Public and International Affairs, Princeton University, Princeton, NJ 08544, USA and <sup>3</sup>Dartmouth College, Hanover, NH 03755, USA

\*Corresponding author. Email: [jmummolo@princeton.edu](mailto:jmummolo@princeton.edu)

## Abstract

Political elites increasingly express interest in evidence-based policymaking, but transparent research collaborations necessary to generate relevant evidence pose political risks, including the discovery of sub-par performance and misconduct. If aversion to collaboration is non-random, collaborations may produce evidence that fails to generalize. We assess selection into research collaborations in the critical policy arena of policing by sending requests to discuss research partnerships to roughly 3,000 law enforcement agencies in 48 states. A host of agency and jurisdiction attributes fail to predict affirmative responses to generic requests, alleviating concerns over generalizability. However, across two experiments, mentions of agency performance in our correspondence depressed affirmative responses – even among top-performing agencies – by roughly eight percentage points. Many agencies that initially indicate interest in transparent, evidence-based policymaking recoil once performance evaluations are made salient. We discuss several possible mechanisms for these dynamics, which can inhibit valuable policy experimentation in many communities.

**Keywords** Policing; field experiments; selection bias; policy evaluation

The increasing availability of high-resolution data on human behavior and the development of field experimental methods in social science have made research collaborations with practitioners the gold standard in policy research (Cartwright and Hardie 2012). These partnerships – which span substantive arenas including poverty reduction (Alatas et al. 2012), political advertising (Gerber et al. 2011),

---

  This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the [Data Availability Statement](#).

This study was approved by Institutional Review Boards at Princeton University and Dartmouth College. A pre-registration plan for the national experiment appears in the Online Appendix.

Authors are listed in alphabetical order.

© The Author(s), 2022. Published by Cambridge University Press on behalf of The Experimental Research Section of the American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

and health care (Litvack and Bodart 1993) – offer numerous advantages, simultaneously leveraging access to otherwise restricted data, real-world settings, and rigorous experimental designs (Gerber and Green 2012).

But like any approach to research, partnerships with practitioners have drawbacks. Chief among them is the fact that the very organizations being studied decide whether research can proceed, and there are strong reasons to suspect this decision is associated with outcomes scholars wish to understand, like agency performance. Put differently, while many political elites have recently instituted calls for “evidence-based policymaking” (Orszag and Nussle 2017), such declarations may be cheap talk. The political risks associated with allowing outside experts to scrutinize organizational practices – for example the discovery of sup-par performance, or even misconduct – are substantial, especially for poorly functioning organizations (Carpenter 2014; Levine 2020; Moffitt 2010). And if poorly performing agencies are differentially likely to decline research partnerships, the body of evidence produced by one-off research collaborations could fail to generalize to organizations at large (Allcott 2017).

In this study, we assess the determinants and generalizability of research collaborations in the important policy domain of policing. A long history of allegations of racial bias (Alexander 2010; Gelman Fagan and Kiss 2007; Lerman and Weaver 2014), a recent string of high-profile police-involved killings (Edwards, Lee, and Esposito 2019), and growing concern over the use of excessive force and militarized policing (Gunderson et al. 2019; Knox, Lowe, and Mummolo 2020) have spurred numerous collaborations between academics and law enforcement agencies to detect inequity in police procedures (e.g. Goff et al. 2016) and test the efficacy of proposed reforms (e.g. Yokum, Ravishankar, and Coppock 2019). But the highly politicized nature of policing suggests many agencies will be reluctant to partner with researchers and that the ones who do are unrepresentative of the roughly 18,000 law enforcement agencies in the USA.

To evaluate the severity and nature of selection, we conducted two field experiments in which we sent offers to roughly 3,000 local police and sheriff’s departments to discuss a potential research collaboration with scholars at two East Coast universities and analyzed variation in responses. This design allowed us to assess both the correlates and causes of willingness to collaborate. Merging data on responses with records of jurisdiction demographics, local partisanship, department personnel, and agency performance, we first show that agencies open to discussing research collaborations are largely similar to those that declined our invitations. This finding bolsters the validity of the collaborative research approach and suggests findings emanating from one-off research partnerships are plausibly contributing to a generalizable body of knowledge. An exception is the population size of jurisdictions – we find larger jurisdictions are less likely to respond affirmatively to our outreach efforts. However, we find no significant associations between responses to our messages and numerous agency performance metrics, suggesting agencies arguably most in need of reform are not systematically averse to an initial discussion about a potential collaboration. However, across two experiments, including a pre-registered nationwide replication, a randomized mention of agency performance in our communications depressed affirmative responses by roughly eight percentage points. These negative effects hold even for top-performing agencies.

Agencies that initially show openness to research partnerships look broadly similar to those who will not consider them, but the willingness to partner with

academics for policy research is not as widespread as it appears. Once discussions move from the general to the specific and raise the prospect of performance evaluations critical to the field testing of any new policy, many agencies recoil. There may be several reasons for this reaction. Law enforcement may be averse to systematic evaluation – a sign that many agencies indicating willingness to discuss collaborations may be engaging in cheap talk. On the other hand, raising the specter of performance too early, before trust is established, may stymie an otherwise fruitful collaboration (Glaser and Charbonneau 2018). Regardless of the precise mechanism at work, this dynamic reveals a barrier to research collaborations that can preclude valuable policy experimentation in many communities.

### Experimental design

We began with a study in New Jersey involving 462 agencies, paired with newly released detailed data on the use of force (nj.com 2019). During April and May of 2019, we contacted police chiefs offering to collaborate on research that “aims to make both citizens and officers safer by reducing the frequency of violence during police-citizen interactions” (see Online Appendix section B2 for full text). We relied on a custom Python script to prepare and send our messages from a dedicated institutional email. These messages contained no deception; offers to discuss collaborations were sincere. We offered to work pro-bono and cover all research costs and added “We are not asking for a firm commitment now” but are simply asking whether the recipient is “interested in discussing a potential collaboration further.” Respondents could answer (via links in email and a URL provided in print letters) yes, no, or “I am not sure, but I would like more information.” Our primary outcome is a binary indicator of answering “yes,” with all other responses and non-responses coded as negative responses. If we received no response after three email attempts – spaced eight days apart – we sent a posted letter one week after the final email.

Agencies in the N.J. study were randomly assigned to one of four conditions.<sup>1</sup> All agencies received the information above, which served as the full text for those in the control condition. Three treatment conditions included language aimed at testing how common features of research collaboration requests affect agency responses. One treatment cell included a promise of confidentiality in any publication that resulted from a research partnership, which is a common practice in such settings and which we hypothesized would increase affirmative responses. A second “ranking” condition included mention of agency performance: the agency’s rank on uses of force per officer among contacted agencies. A third condition combined both the confidentiality and ranking treatments (with the order of the two treatments randomized within the text of the email across recipients).

Following the N.J. study, we deployed a second pre-registered experiment in 47 additional states during September and October of 2019, in which we attempted to contact approximately 2,500 local police and sheriff’s departments, a sample size we chose based on a power analysis in order to detect a possible interaction between the performance treatment and agency rank. We randomly drew our sample from a

<sup>1</sup>See Online Appendix Section D4 in Supplementary material for balance tests.

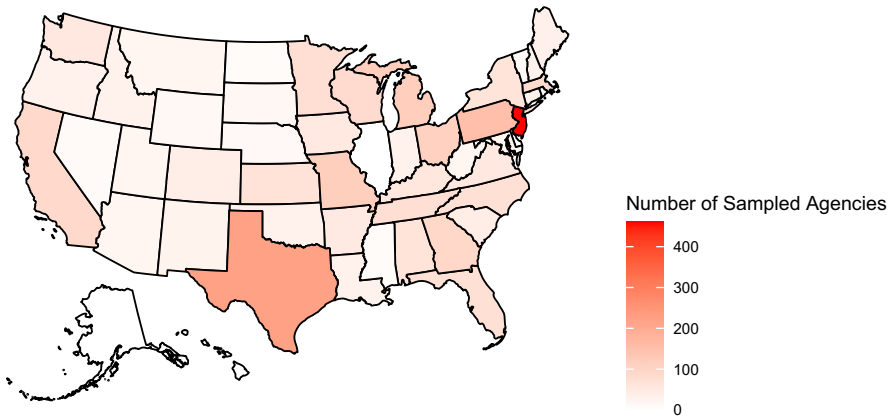


Figure 1.

Coverage of Field Experiments. Number of contacted agencies in each US state. Over 400 agencies were contacted in the N.J. study.

population of roughly 7,700 agencies that consistently report crime data to the FBI and for whom we could ascertain reliable contact information (these criteria excluded Alaska and Illinois; see Appendix section A1 for sampling details). Roughly 60% of the US population resides in these agencies' jurisdictions, according to FBI data. While we would ideally wish to sample from the entire USA, the population of agencies that remain after applying these filtering criteria are those with whom a productive research collaboration might plausibly occur. It would be difficult to form collaborations with agencies that do not regularly report basic crime data or publicize reliable contact information. Our sample is therefore a relevant one for applied researchers.

The design of this experiment was highly similar to the N.J. study with some exceptions. Two changes were aimed at maximizing statistical power. First, we retained only the ranking treatment and control conditions. Second, we employed a matched pair design (Gerber and Green 2012), in which agencies in the same state serving roughly the same population size were paired, and one agency was randomly assigned to treatment. Specifically, treated agencies were told how they ranked among the roughly 2,500 agencies sampled on the share of violent crimes "cleared" between 2013 and 2017 (crimes where an arrest was made and charge was filed) – a salient statistic for police agencies, and one on which journalists often focus (e.g. Madhani 2018). The use of two different performance metrics across these experiments helps to ensure the robustness of any observed treatment effects.

We hypothesized that mentions of agency performance would filter out "cheap talk" and depress affirmative responses on average, since making performance evaluations salient could cause agencies to consider the political risks associated with research partnerships. However, we anticipated that this negative effect would attenuate with agency rank, since agencies informed they were performing well relative to peers may be less likely to recoil at the specter of performance evaluations. Following both experiments, all contacted agencies were sent a debrief message

informing them of the purpose of the experiment and reinforcing that our messages contained no deception.<sup>2</sup>

Combined, contacted agencies serve jurisdictions that are home to close to 80 million people according to FBI data (see Figure 1), approximately one-quarter of the US population. These include large metropolitan police forces, mid-sized agencies, and small rural departments employing just a handful of officers.<sup>3</sup>

### Little evidence collaborating agencies are unrepresentative

To test whether willingness to collaborate systematically varies with agency attributes, we merged agency-level data on crime, fatal officer-involved shootings between 2015 and 2018, personnel, and jurisdiction demographics.<sup>4</sup> In total, 319 agencies indicated willingness to discuss a potential research collaboration – approximately 11% of our combined sample of 2,944 agencies across the two experiments – 238 agencies responded negatively to our message, and 2,387 agencies did not reply at all.

We estimated separate bivariate linear regressions predicting affirmative response as a function of each covariate, weighted by jurisdiction population.<sup>5</sup> We correct resulting *p*-values on all regression coefficients using the Benjamini-Hochberg method (Benjamini and Hochberg 1995), though this adjustment makes little difference to our overall conclusions. To avoid conflating the predictive value of a regressor with the effect of our randomized interventions, we confine this analysis to the roughly 1,400 observations assigned to control, of which 201 agencies responded affirmatively.<sup>6</sup>

Figure 2 displays the predicted change in the probability of an affirmative response when an agency is above (relative to below) the median on each trait. Across the 44 test results in this figure, only one covariate, population of the jurisdiction, was significantly associated with responses in either the control or treatment groups. Specifically, moving from below to above the median jurisdiction population in our sample is associated with a 14.84 percentage point decrease in the probability of an affirmative response (BH-corrected *p*-value = 0.01). However, none of the numerous measures of performance significantly predict responses. Overall, our analysis indicates selection bias – at least at the initial point of contact from researchers – poses a minimal risk in this setting.

Some may question whether we are missing meaningful associations in this analysis due to a lack of statistical power. But while additional data may allow us to detect correlations, we note that several features related to agency performance generate opposing signs, for example assaults on officers and a host of crime measures including murders and rapes per capita. While larger jurisdictions are less likely to show interest in research collaborations – an important limitation that could stymie

<sup>2</sup>Communication between contacted agencies could contaminate our results. This is less likely in the national experiment given its geographic spread. To the extent cross-talk occurred, it likely homogenized responses across treatment conditions and attenuated effects.

<sup>3</sup>Following our IRB protocol, we map agencies at the state level to maintain anonymity.

<sup>4</sup>See Online Appendix Section C3 in Supplementary material for details on data sources.

<sup>5</sup>Note: All continuous predictors are specified as indicators for being above the median value of the variable.

<sup>6</sup>For comparison, we conduct the same analysis for agencies in the treatment group in the national study; see Appendix Figure E1 in Supplementary material. We exclude N.J. State Police in this analysis due to difficulty in accurately pairing with U.S. Census data.

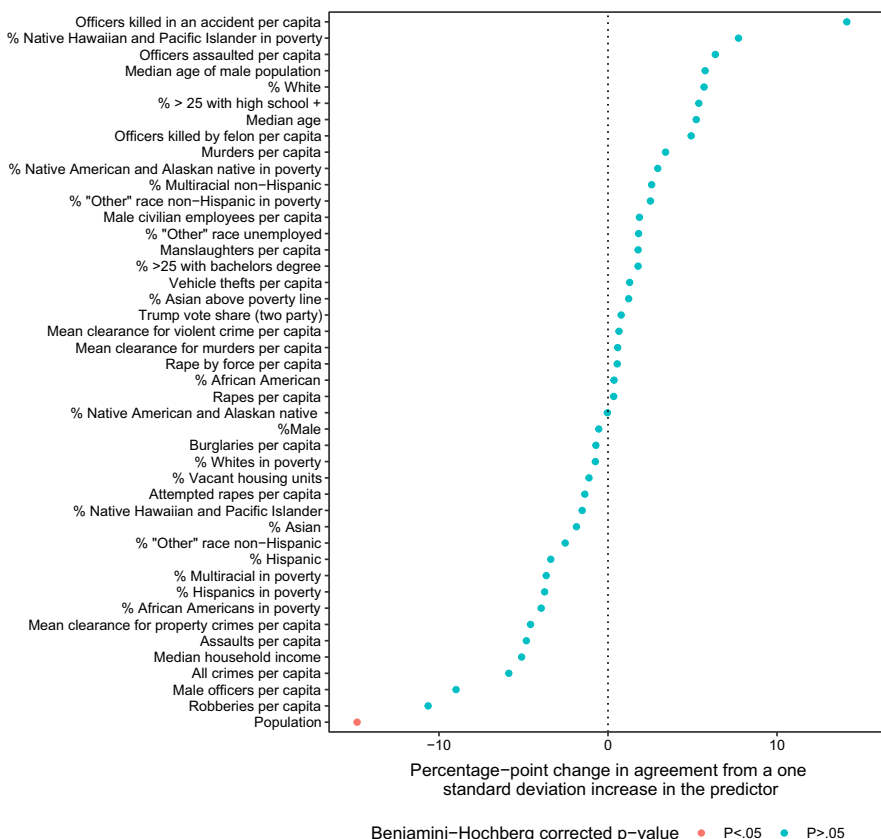


Figure 2.

Agency and jurisdiction attributes do not predict affirmative responses.

partnerships in major departments – the overall pattern of results does not indicate selection related to agency performance, suggesting agencies arguably in most need of reform are not systematically declining to collaborate.

### Mentions of performance evaluations inhibit collaborations

We now turn to assess the impact of our experimental interventions. Figure 3 displays the average effect of each treatment relative to the control condition estimated via linear regression. Because we cannot guarantee all messages were reviewed, these represent Intention-to-Treat effects (ITTs), understating the effect of universally received similar messages.<sup>7</sup> In the national experiment, our models include indicators for all matched pairs, with standard errors clustered by matched pair.

<sup>7</sup>Emails and post letters sent to four agencies were returned to us due to invalid addresses. We drop these and their corresponding matched pair from all analyses since the ITT interpretation is invalid.

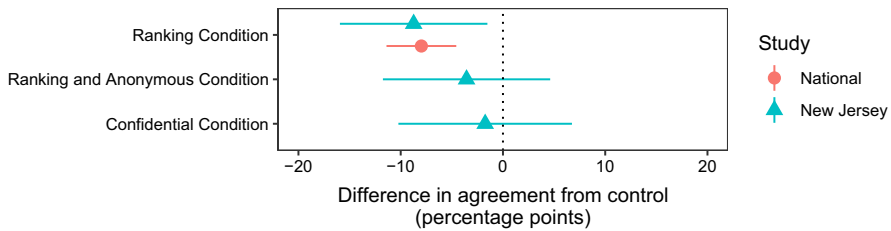


Figure 3.  
Mentioning agency performance lowers affirmative responses.

Turning first to the N.J. experiment, we find randomized offers to keep the identity of collaborating agencies confidential, including one version where a performance cue was also supplied, had no detectable effect on response rates ( $\beta = -0.02$ ,  $SE = 0.04$  and  $\beta = -0.04$ ,  $SE = 0.04$  respectively). This was surprising, as such confidentiality offers are often made to convey a sense of security and thereby increase the likelihood of collaborations. However, because such offers still rely on academic collaborators to keep their word and effectively safeguard agency identities, this promise may ring hollow, and additional assurances may be required before agencies will consider collaborations. However, recipients told their statewide rank on mean uses of force per officer (“Ranking Condition”) were roughly 9 percentage points ( $SE = 0.04$ ) less likely to respond affirmatively than agencies assigned to control where about 13% of agencies agreed. Strikingly, this effect was precisely replicated in the nationwide experiment: agencies told their rank on violent crime clearance rates were about 8 percentage points ( $SE = 0.01$ ) less likely to say they would discuss a potential collaboration.

Contrary to our expectations, additional tests interacting treatment assignment with agency rank show that these negative effects persist even among top-performing agencies (see Online Appendix Figure G1). This result is consistent with police agencies having a strong aversion to outside evaluation and suggests a powerful impediment to the formation of research partnerships. While many agencies indicate openness to collaboration, a large share recoils once the topic of agency performance is inevitably broached. This may be because agencies that performed well on a given metric in the past have no guarantee of positive results in the future, especially once outside scrutiny is allowed.

Of course, other mechanisms are also possible, and we have a limited ability to adjudicate between them with the data at hand. Prior work on collaborations with law enforcement agencies emphasizes the importance of researchers spending time with agency officials to educate themselves about the particulars of the institution and personnel (Glaser and Charbonneau 2018; Levine 2020). Conveying a measure of performance in an initial outreach message may have inadvertently sent the signal that the researchers had “jumped the gun” and evaluated the agency before doing proper due diligence. The specific metrics we conveyed in our interventions may also have depressed responses. Agencies may have felt other measures more accurately conveyed their level of performance and thus inferred the researchers contacting them were ill-equipped to assist them.<sup>8</sup>

<sup>8</sup>See Appendix 19 in Supplementary material for additional discussion of causal mechanisms.

However, regardless of the mechanism at play, the end result is the same: mentions of performance evaluations in outreach messages inhibit collaborations. Since evaluating the efficacy of reforms on agency performance is a central goal of these collaborations, we interpret these effects as evidence of a substantial barrier to policy experimentation, but one that may be overcome with a more measured approach to solicitation that works to establish trust with police administrators before discussing performance metrics.

## Discussion and conclusion

While they offer numerous advantages over other methods of inquiry, research collaborations with outside experts also pose political risks that may preclude partnerships in ways that threaten the generalizability of results. Despite a string of recent promising collaborations with individual agencies, researchers have understandably raised concerns over external validity. If agencies willing to collaborate with academics are unrepresentative of agencies at large, then collaborative field experiments, however carefully executed, may have little value outside the agencies in which they are conducted.

In this paper, we evaluated the nature and severity of selection into research collaborations with police agencies via two field experiments. Our results, precisely replicated across studies, offer several useful insights for applied researchers. First, we find little evidence that agencies which decline to discuss research collaborations are dissimilar to those that respond affirmatively across a range of agency and jurisdiction attributes. An exception pertains to the population size of jurisdictions, with larger jurisdictions responding affirmatively less often than smaller ones. We also find that the vast majority of agencies we contacted did not respond at all. This low response rate underscores the difficulty of initiating academic collaborations with practitioners and suggests the need to develop institutions that can facilitate such connections moving forward.

Our experimental results are also consistent with many agencies who profess an openness to evidence-based policymaking engaging in cheap talk, as a mere mention of agency performance substantially depresses affirmative responses. Our analysis is confined to the initial stage of contacting agencies to develop research partnerships. As this process unfolds and the possibility of negative publicity that sometimes results from transparent research is made more apparent, it is possible that even more agencies would be unwilling to collaborate on evidence-based policy research. However, we also recognize that alternative mechanisms may be at play. Specifically, mentions of agency performance in an initial outreach message may have decreased trust in the research team. This reaction need not be limited to policing collaborations: given these results, researchers seeking collaborations with schools, legislatures, and a host of other institutions may face similar hurdles if performance evaluation is mentioned too hastily. Our results suggest that a more cautious approach to solicitation of these partnerships that seeks to build a relationship over several interactions before discussing the details of performance evaluations may be more fruitful, though such an approach would add costs to the already burdensome process of establishing research partnerships with

practitioners. Future experiments could be deployed to disentangle these competing theories and to test their validity in other policy domains.

Increasing openness to evidence-based policymaking offers a valuable opportunity to generate effective reforms in a range of social institutions. However, we have little systematic evidence on the demand for such collaborations by practitioners (see Levine (2020)). This paper supplies such evidence and provides a replicable template for future work in a range of policy domains. Accumulating additional scientific knowledge on the scale and determinants of the willingness of practitioners to collaborate with academics can serve to streamline and accelerate the process of policy experimentation.

**Supplementary material.** To view supplementary material for this article, please visit <https://doi.org/10.1017/XPS.2022.21>

**Data availability statement.** Support for this research was provided by the Princeton University and Dartmouth College. The data, code, and any additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at: <https://doi.org/10.7910/DVN/IDAIUZ> Goerger, Mummolo, and Westwood (2022).

**Acknowledgements.** We thank Tori Gorton, Alexandra Koskosidis, Destiny Eisenhour, Krystal Delnoce, Grace Masback, and Madeleine Marr for research assistance.

**Conflicts of interest.** Samantha Goerger and Sean Westwood declare no conflicts of interest. Jonathan Mummolo works as a paid consultant for the American Civil Liberties Union, the NAACP Legal Defense Fund, and the US Dept. of Justice Civil Rights Division. He provides statistical expertise for litigation concerning discrimination in law enforcement and maintains confidentiality agreements pertaining to this work. These agreements do not apply to any data used in this paper.

**Financial support.** This research was funded by Princeton University and Dartmouth College.

**Ethics statement.** The protocols for this project were reviewed by the Princeton IRB (protocols 11921 and 11023). Dartmouth College deferred to Princeton as the IRB of record. We affirm that this research adheres to APSA's Principles and Guidance for Human Subjects Research. See section K of the online appendix for details.

## References

- Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken and Julia Tobias. 2012. Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review* 102(4):1206–40.
- Alexander, Michelle. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press.
- Allcott, Hunt. 2017. Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics* 130(3): 1117–65.
- Benjamini, Yoav and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1): 289–300.
- Carpenter, Daniel. 2014. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton University Press.
- Cartwright, Nancy and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press.
- Edwards, Frank, Hedwig Lee and Michael Esposito. 2019. Risk of Being Killed by Police Use of Force in the United States by Age, Race–Ethnicity, and Sex. *Proceedings of the National Academy of Sciences*, 116(34): 16793–8.

- Gelman, Andrew, Jeffrey Fagan and Alex Kiss.** 2007. An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias. *Journal of the American Statistical Association* 102(429): 813–23.
- Gerber, Alan S. and Donald P. Green.** 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.
- Gerber, Alan S., James G. Gimpel, Donald P. Green and Daron R. Shaw.** 2011. How Large and Long-Lasting are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment. *American Political Science Review* 105(1): 135–50.
- Glaser, J. and A. Charbonneau.** 2018. Making research matter: A Psychologist's Guide to Public Engagement. American Psychological Association chapter Law enforcement: Finding common purpose.
- Goerger, Samantha, Jonathan Mummolo and Sean J. Westwood.** 2022. Replication Data for: Which Police Departments Want Reform? Barriers to Evidence-Based Policymaking. *Harvard Dataverse*. <https://doi.org/10.7910/DVN/IDAIUZ>
- Goff, Phillip Atiba, Dean Obermark, Nancy La Vigne, Jennifer Yahner and Amanda Geller.** 2016. The Science of Policing Equity: Measuring Fairness in the Austin Police Department. Technical report Center for Policing Equity & John Jay College of Criminal Justice. [https://policingequity.org/images/pdfs-doc/Austin\\_PDI\\_Report\\_2016\\_Release.pdf](https://policingequity.org/images/pdfs-doc/Austin_PDI_Report_2016_Release.pdf).
- Gunderson, A., E. Cohen, K. Jackson, T. S. Clark, A. Glynn and M. L. Owens.** 2019. Does Military Aid to Police Decrease Crime? Counterevidence from the Federal 1033 Program and Local Police Jurisdictions in the United States. Working Paper. <https://tinyurl.com/y66j6s8g>.
- Knox, Dean, Will Lowe and Jonathan Mummolo.** 2020. Administrative Records Mask Racially Biased Policing. *American Political Science Review* 114: 619–37.
- Lerman, Amy and Vesla Weaver.** 2014. *Arresting Citizenship: The Democratic Consequences of American Crime Control*. University of Chicago Press.
- Levine, Adam Seth.** 2020. Why Do Practitioners Want to Connect with Researchers? Evidence from a Field Experiment. *PS: Political Science & Politics* 53(4): 712–17.
- Litvack, Jennie I. and Claude Bodart.** 1993. User Fees Plus Quality Equals Improved Access to Health Care: Results of a Field Experiment in Cameroon. *Social Science & Medicine* 37(3): 369–83.
- Madhani, Aamer.** 2018. Chicago Police Solved Fewer than One in Six Homicides in the First Half of 2018. *USA Today*. <https://bit.ly/2Q3yt1u>.
- Moffitt, Susan L.** 2010. Promoting Agency Reputation through Public Advice: Advisory Committee Use in the FDA. *The Journal of Politics* 72(3): 880–93.
- nj.com.** 2019. *The Force Report*. Technical report. <https://force.nj.com/>.
- Orszag, Peter and Jim Nussle.** 2017. Policymaking commission offers a glimmer of hope in hyper-partisan Washington. *The Hill*. <https://tinyurl.com/ry8hvax>.
- Yokum, David, Anita Ravishankar and Alexander Coppock.** 2019. A Randomized Control Trial Evaluating the Effects of Police Body-Worn Cameras. *Proceedings of the National Academy of Sciences* 116(21): 10329–32.