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

Samantha Goerger[1], Jonathan Mummolo[2], and Sean J. Westwood[3]

[1]Independent Scholar
[2]Princeton University
[3]Dartmouth College

December 11, 2021

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 are likely engaging in cheap talk, and recoil once performance evaluations are made salient. These dynamics can inhibit valuable policy experimentation in many communities.

Word count: 2,228

*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. We thank Tori Gorton, Alexandra Koskosidis, Destiny Eisenhour, Krystal Delnoce, Grace Masback and Madeleine Marr for research assistance. Authors listed in alphabetical order.

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), 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—e.g. the discovery of sup-par performance, or even misconduct—are substantial, especially for poorly functioning organizations (Carpenter, 2014; 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; Lerman and Weaver, 2014; Gelman, Fagan and Kiss, 2007), 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 more than 18,000 law enforcement agencies in the United States.

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. 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. This dynamic reveals a general 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.[a] 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 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 of Alaska and Illinois; see Appendix section A1

[a] See Online Appendix Section D4 for balance tests.

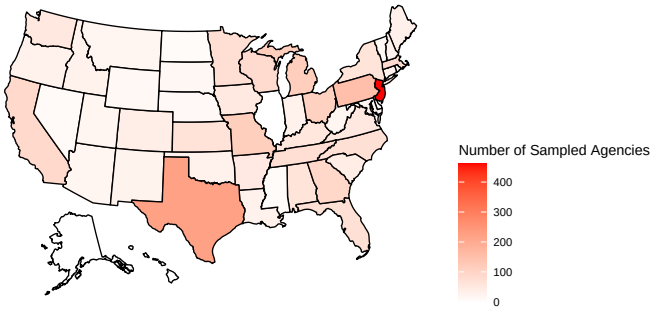
for sampling details). Roughly 60% of the U.S. population reside in these agencies' jurisdictions, according to FBI data. While we would ideally wish to sample from the entire U.S., the population of agencies that remain after applying these filtering criteria are those with whom a productive research 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 2013-2017 (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 generally, 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 spectre 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.[b]

[b]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.

Figure 1: **Coverage of Field Experiments**. Number of contacted agencies in each U.S. state. Over 400 agencies were contacted in the N.J. study.



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 U.S. population. These include large metropolitan police forces, mid-sized agencies and small rural departments employing just a handful of officers.[c]

Collaborating Agencies Are Representative

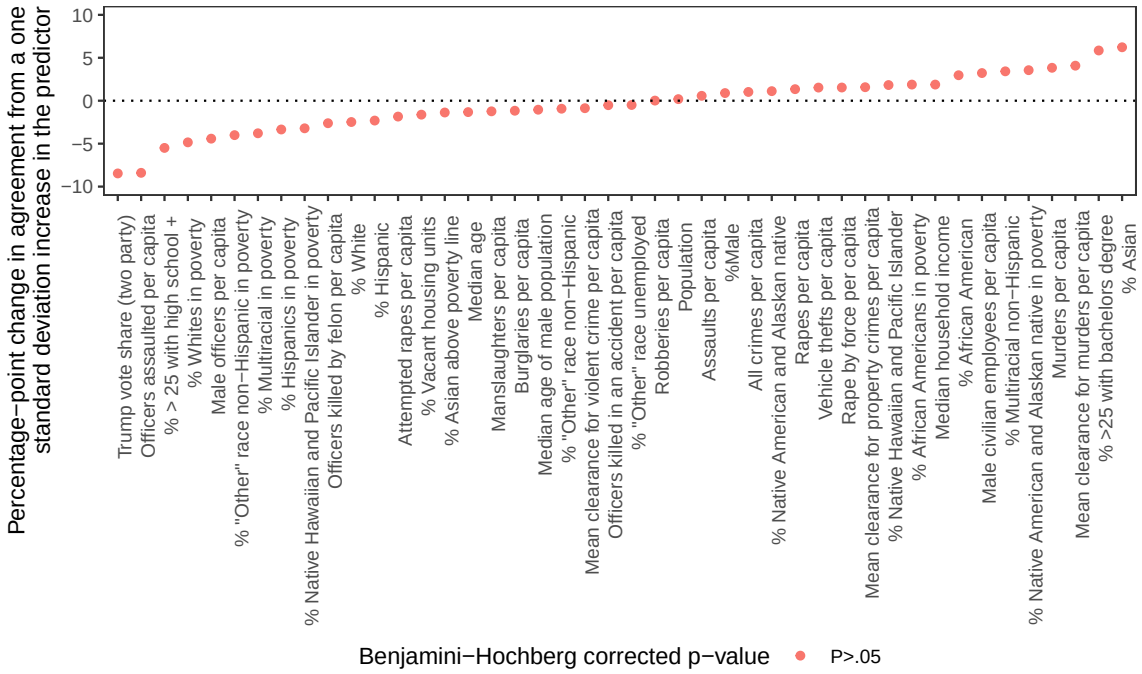
To test whether willingness to collaborate systematically varies with agency attributes, we merged agency-level data on: crime, fatal officer-involved shootings between 2015 and 2018, personnel, and jurisdiction demographics.[d] In total, 319 agencies indicated willingness to discuss a potential research collaboration—approximately 11% of our combined sample of 2,942 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. 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,

[c] Following our IRB protocol, we map agencies at the state-level to maintain anonymity.

[d] See Online Appendix Section C3 for details on data sources.

Figure 2: **Agency and jurisdiction attributes do not predict affirmative responses.**



of which 201 agencies responded affirmatively.[e]

Figure 2 displays the predicted change in the probability of an affirmative response given a one-standard-deviation increase in each predictor. Across roughly 50 tests, no covariate was significantly associated with 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, other features of the results belie the existence of meaningful relationships. For one, several features related to agency performance generate opposing signs, e.g. assaults on officers and a host of crime measures including murders and rapes per capita. Second, only one result, officers assaulted per capita, is statistically significant when we opt not to correct for multiple testing. Third, we find no significant associations even if we include data from all experimental conditions to maximize sample size (see Figure E2 in Appendix). The largest

[e]We exclude N.J. State Police in this analysis due to difficulty in accurately pairing with U.S. Census data.

estimated coefficient relates to Trump's share of the two-party vote in an agency's county, suggesting agencies in conservative areas may be less likely to collaborate. However, the overall pattern of results does not indicate selection related to agency performance, meaning agencies arguably in most need of reform are not systematically declining to collaborate.

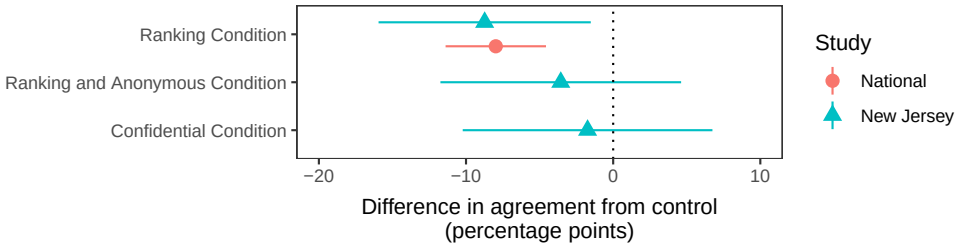
Performance Evaluations Inhibit Collaborations

We now turn to assessing 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.[f] In the national experiment, our models include indicators for all matched pairs, with standard errors clustered by matched pair.

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 ($\beta = -0.02, se = 0.04, p = 0.69$ and $\beta = -0.04, se = 0.04, p > 0.40$ 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 ($se = 0.004, p = 0.02$) less likely to respond affirmatively than agencies assigned to control where about 15% 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 ($se = 0.01, p < 0.01$) less likely to say they would discuss a potential collaboration.

[f] 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**



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 highlights the strong aversion of police agencies to outside evaluation, and suggests a general and powerful impediment to the formation of research partnerships. While many agencies indicate openness to collaboration, a large share recoil 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.

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 agencies which decline to discuss research collaborations are largely similar to those that respond affirmatively across a range of agency and jurisdiction attributes. However, our experimental results imply that many agencies who profess an openness to evidence-based policymaking are likely

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.

Increasing openness to evidence-based policymaking offers a valuable opportunity to generate effective reforms in a range of social institutions and fortunately, concerns over external validity appear to be overstated, at least in one important policy domain. But the political risks associated with performance evaluations impose a significant constraint, underscoring the inescapable politics of public policy.

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):1117–1165.

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 .

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):813–823.

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):135–150.

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.

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* 114:619–637.

Lerman, Amy and Vesla Weaver. 2014. *Arresting Citizenship: The Democratic Consequences of American Crime Control*. University of Chicago Press.

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):369–383.

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):880–893.

Mummolo, Jonathan, Erik Peterson and Sean Westwood. 2019. "The Limits of Partisan

Loyalty." *Political Behavior* pp. 1–24.

nj.com. 2019. "The Force 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.

Westwood, Sean and Erik Peterson. 2019. "Compound political identity: How partisan and racial identities overlap and reinforce." *Available at SSRN 3417476* .

Westwood, Sean J, Erik Peterson and Yphtach Lelkes. 2019. "Are There Still Limits on Partisan Prejudice?" *Public Opinion Quarterly* 83(3):584–597.

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):10329–10332.