Experiments on and with Street-Level Bureaucrats

Noah L. Nathan & Ariel White[1]

October 2019

Prepared for *Handbook of Advances in Experimental Political Science*, Cambridge University Press, James Druckman and Donald Green, editors

Abstract

We review recent experimental research on the behavior of street-level bureaucrats. These front-line government workers are tasked with implementing most government policy in both advanced democracies and developing countries, but their behavior is often difficult to observe. We highlight how experimental approaches have helped to address classic questions about street-level bureaucratic behavior, and then consider design challenges that arise in running experiments in this context. Finally, we raise several ethical concerns about experimentation on street-level bureaucrats, and propose strategies to minimize the social costs, and maximize the social benefits, of such research.

1 Introduction

Street-level bureaucrats – front-line government workers who interact directly with the public and implement policies – are a crucial part of the functioning of government. But it is not easy to study them, especially on a large scale; their behavior is notoriously difficult to measure, and it has also been challenging to evaluate programs intended to change their behavior. Field experiments have provided a valuable window into the actions of street-level bureaucrats, and show further promise as a set of tools with which to study these officials.

In this chapter, we begin by introducing several key substantive topics in the study of street-level bureaucracy, and review recent contributions that field experiments have made in these areas. We then turn to recommendations for researchers seeking to contribute to this literature. We discuss design considerations specific to experiments on

[1] Email: `nlnathan@umich.edu`, `arwhi@mit.edu`

this population and outline serious ethical concerns that may arise in such studies. Finally, we conclude by suggesting new substantive directions, alongside methodological and ethical best practices, for future researchers considering new field experiments on – or with – the street-level bureaucracy.

2 Substantive Contributions

Most interactions that citizens have with government do not involve politicians, but street-level bureaucrats, the frontline workers that implement most policies in both rich and poor countries. Street-level bureaucrats are gatekeepers to public benefits and other distributive programs, as well as enforcers of criminal law. They are often by far the most visible presence of the state in citizens' lives (Soss and Weaver, 2016). In the US, quotidian interactions with the nation's 3.2 million public school teachers, over 700,000 police officers, and nearly 300,000 public sector social workers are shown to significantly shape people's views of government and their willingness to participate in the political process (Soss, 1999; Bruch, Ferree and Soss, 2010). Yet despite the substantive importance of these actors, and the fact that their large numbers should facilitate systematic study of their behavior, they are subject to fairly little experimental inquiry relative to other governmental actors like politicians.

Such experiments stand to improve our understanding of bureaucratic behavior and its effects on citizens' experiences with the state. As Lipsky (1980) argues, street-level bureaucrats have a great deal of discretion in how they do their jobs, and often make many small decisions over the course of a workday that cumulatively can have large effects on citizens' lived experiences with government policy. These observations lead to a series of predictions about bureaucratic behavior that are important both to political scientists and to policymakers: that these workers might subvert the policy goals of their bosses, that they might incorporate personal biases into their decision-making, and that they might succumb to corruption or other professional wrongdoing. Each of these behaviors can be hard to observe, and it can be difficult to know whether interventions intended to change bureaucrats' behavior are working. Experiments have proven helpful in solving both of these research problems: providing richer measurements of bureaucratic behavior and providing better tests of the effectiveness of policies intended to change it. In the first part of this chapter, we discuss the contributions of experiments to three substantive topics in the study of street-level bureaucratic behavior.

Principal-Agent Problems A foundational issue in street-level bureaucracy is the principal-agent relationship between front-line government workers and their supervisors. Because principals – politicians or agency leaders – face serious monitoring challenges, street-level agents can undermine policy. They may shirk or undersupply effort, or serve some clients but not others. At the extreme, street-level bureaucrats may even work to sabotage policies. Most attempts to improve street-level bureaucrats' compliance involve increasing principals' oversight or monitoring ability. But increased supervision often fails, and sometimes even creates unintended consequences that can be worse than bureaucrats' initial deviations from stated policy objectives (Brehm and Gates, 1997).

Scholars of American politics and the political economy of development both increasingly employ field experiments to understand the causes of bureaucratic shirking and sabotage, as well as to evaluate interventions aimed at improving compliance. We highlight three examples, all focused on bureaucrats working in social welfare provision, but across very different settings. These studies demonstrate the particular value of experiments: they allow scholars to leverage randomization to observe forms of shirking and sabotage that would be difficult to measure directly; they also demonstrate limits of common tools used to mitigate bureaucratic non-compliance by providing cleaner causal estimates of the effectiveness of policy interventions.

Our first example uses randomization to observe shirking that would otherwise be difficult to observe. Slough (2018) examines two major nationwide social welfare programs in Colombia that are administered by municipal governments. In this context, local bureaucrats can use their discretion to give preferential treatment to clients who have political connections. In some municipalities, many poor residents remain un-enrolled in the programs despite meeting eligibility criteria; elsewhere, vastly more recipients are enrolled than should be eligible. Both indicate selective, possibly politically motivated, failure to enforce program rules.

Slough implements a phone-based audit experiment: research confederates with randomly-assigned personal characteristics called front-line service offices to ask about program eligibility.[2] Slough finds that front-line bureaucrats provide less information to lower class callers and to internal migrants (as cued by the vocabulary and regional accents of the callers), and shows that this bias is greatest in areas in which these marginalized constituents are less likely to have political ties to workers' political supervisors. Overall, bureaucrats appear to provide better services to constituents with the greatest ability to complain to the politicians that oversee them, while shirking from requests from eligible program beneficiaries who come from social groups that are typically less tied to politicians. Without being able to leverage randomization, these subtle forms of bureaucratic shirking – which have substantively important consequences for access to Colombia's welfare state – would have likely remained unobserved. Non-experimental research could have observed different rates of service provision to different groups, but would have struggled to rule out alternative explanations like underlying differences in eligibility or request rates. This paper highlights the value of experimentation for measuring bureaucratic behavior, and illustrates one way in which supervision of street-level bureaucrats – such as by local politicians – can create problems of its own, with bureaucrats pressured to selectively implement policies in politically-expedient, but unequal ways.

Rather than using experiments as measurement devices, our next two examples in-

[2] For more discussion of audit studies, see Butler and Crabtree's chapter in this volume.

stead use field experiments to test the effectiveness of policy interventions intended to improve street-level workers' behavior. Hess et al. (2016) study local compliance with the National Voter Registration Act (NVRA) in the United States. The NVRA mandates that local health and social service offices assist clients in registering to vote, with the goal of reducing persistent class disparities in voter participation. But the NVRA has faced widespread non-compliance from agency staff, with less than 1 percent of potentially eligible clients receiving registration assistance in some jurisdictions (Hess et al. 2016, 5). Hess et al. attribute this to a combination of shirking and sabotage: many local staff forego implementing the NVRA because they believe voter registration falls outside their agency mission; elsewhere, active hostility to the NVRA has blocked implementation. These forms of "agency loss" occur amidst a complex monitoring structure in which the principals with the greatest interest in seeing the NVRA implemented – state election officials – lack direct supervisory authority over the street-level bureaucrats in other agencies of the state government tasked with implementing it. This limits the available levers with which to threaten agents over non-compliance.

Hess et al. partner with election officials in two states to test politically feasible interventions that increase pressure on street-level bureaucrats to comply. In the first, Hess et al. randomly assign the timing of an online training module all social service staff are required to take. The training reminds bureaucrats of their responsibilities under the NVRA. In the second, Hess et al. instead assign offices to receive emails from state elections officials encouraging greater compliance. Rather than across the board improvements, Hess et al. find that each treatment only improves compliance among the agency offices that were already complying the most with the NVRA. These findings provide causal evidence demonstrating how small oversight nudges that may be easiest for principals to implement can be insufficient to overcome bureaucrats' incentives for shirking and sabotage.

The second example involves a much more extensive policy aimed at fighting shirking, and is representative of a large number of recent field experiments from development economics focused on bureaucratic principal-agent problems. Government clinics in Rajasthan, India provide free basic health services. Yet Banerjee et al. (2008) show that most poor households continued (as of the mid-2000s) to seek out costly private providers because government clinics were frequently closed due to non-attendance by staff, including nurses. Like many street-level bureaucrats across the developing world, nurses in rural Rajasthan are poorly paid and shirk from formal duties so they can supplement official incomes through other employment.

Banerjee et al. randomly assign clinics to an extended monitoring intervention in which nurses were required to punch-in attendance multiple times per day on tamper-proof machines. This information was passed to district health officials, who were tasked with docking nurses' pay for unexcused absences. This experimental treatment initially had large effects on compliance, nearly doubling nurses' daily attendance. But after one year, the intervention had failed spectacularly: nurses' attendance had become *lower* at treated clinics than those in the control group.

Banerjee et al. show that this reversal was likely due to adaptive behavior by street-level bureaucrats: over time, increasing numbers of time-card machines were reported as broken, preventing monitoring; moreover, district-level supervisors began granting a dramatically higher number of "excused absences" to treated nurses, allowing them to continue to miss work. This may have been because district health officials decided to go easy on their staff, or because nurses learned that they could cut their superiors in on the income gains from their outside employment in return for lax enforcement. Either way, Banerjee et al. echo Hess et al. in demonstrating the difficulty of overcoming principal-agent problems with street-level bureaucrats. Moreover, Banerjee et al.'s results demonstrate how principals sometimes repurpose opportunities for greater monitoring to achieve other ends, such as creating new points of leverage to extract illicit rents.

Bias and Discrimination Another set of experiments examines bias and discrimination by bureaucrats. Across contexts, citizens and scholars worry about whether policies are being implemented equitably: are street-level bureaucrats discriminating on the basis of personal characteristics like race, gender, or membership in ethnic, religious, or political groups? Because their jobs inevitably involve some discretion, workers may implement facially-neutral policies in ways that exclude outgroup members or privilege members of their own groups, either deliberately or unconsciously. Discrimination of this kind can be hard to observe, since bureaucrats themselves are unlikely to admit to biased behavior, and the targets of discrimination may not always realize that it has happened, especially if they were offered plausible reasons for not receiving government services. Administrative data can identify disparate outcomes across groups, such as showing whether one group has higher success rates in applying for benefits, but such disparities do not necessarily prove discriminatory behavior; it is possible that underlying differences in applicants are driving disparate outcomes, not discrimination by street-level bureaucrats. As such, it has been difficult to measure discriminatory behavior or to test whether attempts to reduce discrimination were working. Experiments can help with these problems.

One realm where neutral policy implementation is particularly important is election administration. In the United States, a prominent recent debate about elections focuses on voter identification laws and whether they disproportionately prevent minority voters from casting ballots. One way that this could happen is through discriminatory behavior by election officials, who might use their discretion to unevenly enforce voter ID laws or to help some people but not others learn how to comply with the law. But as noted above, it can be difficult to diagnose discrimination by these officials: it is hard to observe their behavior in any consistent way, and even harder to attribute disparate outcomes to discrimination.

As with the Slough (2018) study above, audit-style field experiments again provide a valuable means to overcome these inferential challenges by leveraging random assignment to uncover evidence of disparate bureaucratic responses to otherwise identical requests. In a paper with Julie Faller, we used an audit experiment to measure discriminatory behavior by the local election officials tasked with implementing most voting laws (White, Nathan and Faller, 2015). Before the 2012 election, we emailed thousands of local election officials in 48 states, randomly varying two characteristics of the messages: whether the name of the sender cued Latino identity or not, and whether the email asked about voter ID laws (as opposed to a generic question about voting). We then observed whether each email received a response over the next few weeks. We found that emails signed with Latino-sounding names like Luis Rodriguez were several percentage points less likely to receive responses than emails signed with putatively-white names like Greg Walsh.

Exploratory analyses also suggest that bureaucrats in jurisdictions that were federally monitored for discrimination under parts of the Voting Rights Act showed no bias against Latino emailers. This is consistent with predictions from the street-level bureaucracy literature and suggests that closer monitoring of how workers use their discretion could help to reduce discrimination. We note, though, that this additional evidence remains observational: while we can randomly assign the names used in an email, we do not get to randomly assign which places are subject to federal monitoring. This audit study is a good example of the use of experiments for measurement purposes: it allows us to observe discrimination by officials but does not allow us to causally test explanations for the presence or absence of that discrimination.[3]

Email is a convenient and consistent way of contacting officials, but in some contexts it may not be a realistic way of measuring bureaucratic behavior. Neggers (2018) takes a very different approach to measuring discriminatory behavior by election workers, combining surveys of voters and election workers with a natural experiment: the random assignment of Indian poll workers to polling stations in ways that yielded

[3] By contrast, key mechanisms in other substantive contexts might be subject to testing in an audit study framework by randomly assigning treatments with multiple sub-components. For an example of this approach, see Fang et al. (2019).

ethnically-mixed or homogenous teams. This approach yields an extraordinarily realistic look at discrimination in election administration, as it combines evidence from actual elections with survey experimental outcomes.

Neggers describes a system in which poll workers are randomly assigned to polling stations away from their home areas in order to prevent local election tampering. This process results in otherwise-similar polling stations receiving differently-composed teams of election workers: some contain members of an ethnic minority group, while others do not. The paper demonstrates that being assigned a team with at least one minority member can shift a polling station's voting outcomes several percentage points toward the minority-oriented coalition, with the effect concentrated in polling stations where many would-be voters do not possess voter identification cards.[4] These results suggest that poll workers may use their discretion to make it easier for co-ethnic voters to cast ballots or harder for out-group members to vote. A survey of voters takes further advantage of the pollworker randomization: in this survey, non-minorities report higher levels of polling station satisfaction and successfully being able to vote when their polling station was randomly assigned a team of workers with no minority members. Again, this pattern is concentrated among people without voter cards, suggesting that when pollworkers from the majority group are unconstrained by the presence of minority group members, they may use their discretion to advantage would-be voters from their own group and disadvantage or pressure others.

Neggers then provides a further test of this discrimination explanation with a survey of election workers that includes an embedded experiment: they are given vignettes of potential voters with names that cue majority or minority ethnicity. In this survey experiment, pollworkers are 25% more likely to judge their co-ethnics to be qualified voters compared to hypothetical out-group members with the same characteristics,

[4] As in the case of White, Nathan, Faller (2015)'s analysis of differently-monitored municipalities, Neggers (2018)'s analysis of places with higher and lower levels of voter identification possession is observational: a polling place's assignment to an ethnically-mixed team of pollworkers is random, but the proportion of voters holding identification cards is not.

consistent with pollworkers discriminating in actual election administration.

The approaches included in this paper have different strengths that help to bolster one another, showing the potential complementarity between survey and field and experiments. The real-world election results capture actual behavior in actual elections, but could potentially result from some mechanism other than in-group favoritism (such as actual differences in group members' eligibility to vote at a given polling place). Conversely, the survey experiment of election workers is less externally valid because the workers are responding to hypothetical vignettes, but it allows Neggers to control the information provided about hypothetical voters to be sure that the ethnicity of the voter is driving election officials' discriminatory behavior.

Malfeasance Field experiments have also been useful for studying other forms of bureaucratic malfeasance. We again highlight two examples from the American politics and political economy of development fields. The first reinforces how field experiments provide a strong means to measure sensitive behaviors that are otherwise difficult to quantify. The second experiment instead represents the more traditional use of randomized controlled trials to evaluate policy interventions aimed at changing behavior.

First, Bertrand et al. (2007) estimate the policy distortions of bribery in the street-level Indian bureaucracy, focusing on the provision of drivers licenses in Delhi.[5] As a clandestine, illegal activity, bribery is particularly tricky to measure. Difficulty measuring the extent of bribery, in turn, creates difficulty estimating its effects; if petty corruption is endemic in a street-level bureaucracy, scholars rarely have a valid counterfactual through which they can observe policy outcomes under less (or more) corruption.[6]

Bertrand et al. cleverly leverage randomization to overcome these challenges. A

[5] For further discussion of related research on corruption specifically, see Lagunes and Seim's chapter in this volume.

[6] For more work on experimentally measuring covert activity by government actors, see Pan's chapter in this volume.

sample of applicants already seeking drivers licenses were assigned to one of three conditions. In the first, applicants were offered a monetary bonus conditional on successfully obtaining their license in a short period of time. In the second, applicants were offered free driving lessons. The third group was simply tracked through the application process. The "bonus" treatment created an incentive for applicants to pay larger bribes than they otherwise would to obtain a license. The "lessons" treatment instead improved applicants' actual eligibility for a license. While this design raises ethical issues that we discuss below, it allows Bertrand et al. to back out both the marginal effects of bribery and the relative effects of bribery versus legitimate eligibility without having to observe bribes directly.

Bertrand et al. find that bribery works: applicants in the "bonus" treatment were 24 percentage points more likely to receive licenses during the 32-day window of the study and the marginal return to bribery was greater than the marginal return to improving an applicant's real eligibility. But most importantly, they causally demonstrate the serious policy distortions of bribery: the "bonus" treatment increased the proportion of applicants who obtained licenses despite not knowing how to drive, as indicated in a post-experiment driving test. Bertrand et al. supplement their main findings with an audit experiment in which paid confederates attempted to apply for licenses under different, randomly assigned pretenses. This reveals the mechanics of how the corruption unfolds, as the confederates were almost uniformly steered to fixers offering to secure licenses through informal channels for a fee, which was ostensibly shared clandestinely with the license agents.

Second, in the American context, Yokum et al. (2019) use a field experiment to evaluate a politically-charged policy initiative now being adopted by many local governments to reduce malfeasance by street-level bureaucrats. Amid public outcry over police shootings of unarmed Black men and persistent allegations that a culture of silence among rank-and-file police officers creates impunity for misconduct, some police departments have begun requiring officers to wear body cameras to better monitor their day-to-day interactions with citizens. Reformers argue that cameras will both allow departments to better hold officers accountable for misconduct and have a deterrent effect on officers' behavior.

Yokum et al. (2019) partner with the police department in Washington, DC to assign all active duty, street-level officers in some precincts – over 1100 officers – to wear body cameras while on duty. Officers assigned to the control group continued without cameras, the status quo policy prior to the study. The authors then tracked several indicators of potential malfeasance over an 11 month period, including uses of force, civilian complaints about officers' conduct, and the rate at which prosecutors proceeded with charges based on each officer's arrests.[7]

Contrary to expectations, Yokum et al. (2019) find consistent null effects of body cameras across all indicators of police behavior. In the face of widely-held priors that body cameras would have an impact, the ability of a field experiment to produce direct causal evidence of their ineffectiveness provides useful information to advocates and reformers: it demonstrates that efforts to change police practices would be better concentrated elsewhere. It also has academic value. As noted above, classic theories of street-level bureaucracy predict that enhanced monitoring should have a deterrent effect on malfeasance (Lipsky, 1980). But the ineffectiveness of body-worn cameras to deter misconduct suggests that police officers face a more complex incentive structure. Explaining why these cameras are ineffective can help refine theories of bureaucratic behavior.

3 Design Considerations

In addition to their substantive contributions, these studies raise experimental design challenges that are particularly salient to the study of street-level bureaucracy. We outline four design challenges in this second section that future researchers must consider,

[7] When prosecutors frequently decline to press charges, it implies that an officer has been committing misconduct in the process of arresting citizens.

referring back to specific issues that come up in the studies featured above.

Power (part 1): the risks of a small, and spoilable, subject pool Much like studies of legislators or other political elites, experiments on street-level bureaucrats can run into problems of scale: there are only so many people to include in a given study. In White, Nathan and Faller (2015) for example, the relevant population consisted of "local election officials in the United States," of whom there are fewer than 10,000. This limited pool brings an obvious challenge of limited statistical power: at some point, one cannot gather more observations because there are no more people to include in the study.[8] For this reason, we encourage researchers to think about how to design experiments as efficiently as possible – for example, by including prognostic covariates in analyses to improve precision or blocking carefully before randomization. But it also comes with the less obvious difficulty of the pool being at risk of spoilage if multiple studies are run on the same participants. If officials learn about a study that has been run in the past, for example, they might be quicker to identify future treatments as being part of a research project, in ways that could ruin the design of the study.

Risks of discovery or unintended spillovers The size of the subject pool is not the only thing that makes discovery especially likely in experiments with street-level bureaucrats. Colleagues in the street-level bureaucracy often confer with one another about how to handle tasks, which can lead to SUTVA violations as well as possible discovery of the entire study. In the case of audit experiments, a worker in one office may forward an email to someone in another office, who notices that they received the same request signed with a different name. In the case of experimental tests of policy interventions, treating a given worker may inadvertently "treat" other

[8] In some extreme cases, such sample limitations may mean that an experiment is not feasible without redefining the population of interest. For example, if one were interested in the behavior of a population like "election administrators in Delaware" because of some specific characteristic of that state, we would advise against running an experimental study on these officials; there are simply too few of them to ever have the statistical power to measure even very large effects.

workers who come into contact with them.

Experimentalists have developed many ways to guard against these kinds of spillovers or accidental discovery. These include cluster-randomized treatments that limit the risk of interaction between treatment and control subjects. Careful consideration of street-level bureaucrats' daily activities and spheres of contact can help to determine the correct level of treatment assignment to avoid inadvertent spillovers.[9]

In the audit experiment context, there are a number of ways to guard against discovery. For one, we encourage researchers to design treatments that are naturalistic and will not raise bureaucrats' suspicions.[10] And technological fixes can help make it less likely that treatment conditions will look exactly the same if viewed side-by-side, such as when researchers create many treatment texts or scripts with small permutations.[11] Below, we also discuss ways to avoid using deception altogether, which is ethically preferable.

Concerns about a finite pool of participants – and of experiments being discovered while in the field – are not abstract. In 2016, four years after we had run the audit study reported in White, Nathan and Faller (2015), we learned that two different teams of researchers had begun fielding similar studies to ours in the weeks before the 2016 election. How did we learn this? We were contacted by one state's secretary of state's office to ask if we knew anything about a series of emails received by local election officials there! They explained that they had received a surprising amount of email traffic, including some similar emails signed with different names, and a web search for the names and email addresses used had led them to our published study.[12]

We share this anecdote to underscore several points. Having multiple researchers

[9] For more on these tools, see Aronow's chapter in this volume.

[10] Designing contacts that are low-cost for bureaucrats, such as asking them brief and frequently-asked questions, is not only a good way to avoid wasting their time– it also helps make it less likely that they will have to go ask someone else for help.

[11] For more detailed design suggestions, see Butler and Crabtree's chapter on audit studies in this volume.

[12] One team had even registered the same web domain we had used to send our original emails, which made it even easier for officials to connect the new emails to our original audit study through Google.

trying to experiment on a limited pool of subjects raises obvious threats of discovery; if people are suddenly inundated with contacts, that may raise suspicion. Incautious study design, like contacting bureaucrats multiple times with similar requests or using aliases or treatments that are exactly the same as those used in previous studies, can exacerbate these risks. But each additional study contact with a given pool of bureaucrats carries some risk of spoiling the pool and making them suspicious of future contacts. This means that we should balance the benefits of the knowledge derived from a particular study against not only the immediate costs of that study, but also the knowledge we may forgo if other studies on this population become impossible.

There is currently no centralized repository of all experiments that are currently in the field or that have been run on a given population; we echo Grose (2014)'s call for such a "clearinghouse" for experiments on elites. However, we note that recent moves towards experimental pre-registration may provide a partial solution in the absence of such a clearinghouse (Humphreys, de la Sierra and van der Windt, 2013). We encourage researchers pondering a research project to skim existing preregistration databases like EGAP and the AEA RCT registry to identify other studies that may target the same population of bureaucrats, to register any studies they run, and to "gate" their preregistrations as little as possible.[13] As we note below, this recommendation has both practical goals – preventing discovery or spoilage of the pool – and ethical ones.

Measuring bureaucratic responses and outcomes Experiments on street-level bureaucracy also involve tough choices about how to measure bureaucratic behavior. We highlight two challenges. First, many bureaucratic behaviors that are meaningful to citizens can be hard to quantify in the large-N framework necessitated by experiments. For example, it matters whether bureaucrats treat citizens with dignity and respect (Soss, 1999; Soss and Weaver, 2016), but these concepts are hard to operationalize objectively. In our experiment, we measure the friendliness of election

[13] For further discussion of preregistration and study repositories, see Boudreau's chapter in this volume.

officials' responses to putative constituents based on whether they contained basic salutations like "Dear [NAME]" or sign-offs like "Have a great day" (White, Nathan and Faller, 2015).[14] But we recognize that this is at best an imperfect measurement necessitated by the difficulty of interpreting tone in a consistent manner across thousands of responses.

Second, and closely related, many of the outcomes that can be most directly examined in field experiments on street-level bureaucrats are several steps down a causal chain from our real substantive outcomes of interest. For example, the real motivation of Slough (2018) is to understand which citizens receive welfare benefits; whether the bureaucrats who answer the phone at social service offices provide information is a (potentially minor) antecedent outcome that is not proven to affect the ultimate outcome. Similarly, in our study, information provision about voter ID laws is used to stand in for broader concerns over whether Latino citizens are discriminated against throughout their other (unobserved) interactions with the electoral system (White, Nathan and Faller, 2015).

In both cases, the focus on proxy outcomes is necessitated by the experimental design: manipulating more complex interactions between citizens and bureaucrats can be substantially more difficult. Still, scholars employing these types of research designs must be careful about over-claiming about the external validity of their findings. For example, even if one were to find no discrimination by street-level bureaucrats in responsiveness to some form of simulated citizen outreach, this does not mean there will be no discrimination in the actual distribution of benefits by these same bureaucrats, and *vice versa*. The prospect that the outcome in an audit experiment may not stand in well for the real outcome of interest is particularly acute when these interactions

[14] Separately, there are estimation issues that must be considered when examining second-order experimental outcomes (e.g., friendliness) that are only defined based on realized values of initial outcomes (e.g., whether there was any response). Coppock (2019) explains how results for second-order outcomes will be biased if conditioned improperly on the first-order outcome. The earlier chapter in this volume on audit experiments details statistical approaches for addressing this bias.

afford bureaucrats different degrees of discretion. It may be easier to be discriminatory in a highly discretionary context like responding to an email, even as the actual delivery of services is more rule-bound.

As a result, we believe there will be large returns to future experimental scholars of street-level bureaucracy who can pair strong experimental designs with more creative and innovative measurement strategies aimed at overcoming these challenges. To the extent we can benchmark the outcomes studied against the true outcomes of interest, our experimental designs will be more credible and our conclusions more compelling.

Collaborating with bureaucracies A final design consideration is whether to collaborate with bureaucracies when conducting experiments on street-level bureaucrats. Partnering with government agencies to study their own behavior offers multiple benefits from a design perspective.[15] Most importantly, direct partnerships greatly expand the range of experimental treatments that are feasible, allowing researchers to use experiments to explore a broader range of theoretical questions. For example, the studies described above by Hess et al. (2017), Yokum et al. (2019), and Banerjee et al. (2008) would only have been possible with such a partnership. Partnerships are also an excellent means to maximize the policy relevance of research.

But there are also design risks that must be weighed against these benefits. First, some topics, especially those related to discrimination (e.g., White et al. 2015) or corruption (e.g., Bertrand et al. 2007) may become off limits when collaborating with government agencies. Researchers interested in bureaucratic malfeasance may explicitly want to avoid collaboration so that they can maintain the necessary independence.

Second is that collaborations can create a new principal-agent problem between researcher principals and bureaucratic agents that can be just as difficult to manage as the bureaucratic principal-agent problems being studied. Even if agency heads agree to a research collaboration, their staff may not faithfully implement the researchers'

[15] For further discussion of research partnerships, see Levine's chapter in this volume.

experimental protocols, creating non-compliance problems that bias the experiment's findings. Ensuring compliance requires properly monitoring and managing bureaucrats' incentives, which can be difficult for outsiders who do not yet fully understand the bureaucracy with which they are collaborating and often have no means to observe bureaucrats' day-to-day behaviors.

Well-designed interventions can fall apart if these incentives are not taken seriously. For example, the monitoring intervention for nurses in Banerjee et al. (2008) became ineffective because the middle-tier bureaucrats responsible for implementing Banerjee et al.'s monitoring treatment began undermining it. Rather than docking nurses' pay for missed work days, as designed, nurses' intermediate supervisors instead started granting excessive numbers of excused absences. These mid-level bureaucrats were not experimental subjects, and the study was not designed with their incentives in mind. Yet they ended up being crucial for interpreting the results. If the goal of the experiment were purely policy evaluation, this is not an issue: the treatment's failure provides a valid test of what scaling up such a policy would look like in practice, revealing the compensatory behaviors likely to emerge in response to real-life monitoring interventions. But this kind of non-compliance limits the theoretical knowledge that can be gained: because the treatment was not consistently implemented, this study cannot evaluate what effective monitoring would accomplish.

4 Ethical Considerations

These new experimental studies also raise challenging ethical considerations that are especially salient when studying this population.[16] The costs of this kind of experimental research are often understated or poorly understood, and they are often not borne by those who stand to benefit the most from the research. This disparity is growing as online technologies have made audit experiments on bureaucrats increas-

[16] For a broader discussion of the ethics of experimentation, see Teele's chapter in this volume.

ingly straightforward for researchers to run at large scale. We believe that experimental research on the behavior of street-level bureaucrats carries real value and can uncover otherwise-inaccessible knowledge. However, we encourage researchers to more carefully consider the costs of such research before embarking on it and to make design choices that minimize those costs wherever possible.

We describe three ethical issues in this section, again referring back for examples to the individual studies featured in the vignettes in the first section above. In each sub-section we propose best practices for future scholars.

Deception and consent Experiments on street-level bureaucrats rarely obtain consent from research participants in advance, sometimes use limited deception in order to uncover socially-undesirable behaviors (e.g., racial discrimination), and generally do not debrief bureaucrats that they have been included in a study. These approaches have received mixed reception from the research community. Many economists do not condone any deception or the waiving of consent, and some political scientists also fall into this camp; Teele (2014) concludes that experiments that do not gain informed consent "are unethical under the principles of the *Belmont Report*."

McClendon (2012) provides a spirited defense of the ethics of waiving consent and using deception when studying the behavior of public officials, noting that knowledge of their behavior is of high social value and can sometimes be gained only through deception. Still, McClendon acknowledges potential externalities from such an approach, describing risks to subjects as well as to future research, and proposes a rule to "internalize[..] some of the potential costs": researchers should only experiment on public officials they anticipate studying again in future. The hope here is that concerns about officials' willingness to participate in future research (or, in the case of audit studies, their willingness to believe a given treatment in a future study) will lead researchers to be thoughtful about what they do.

McClendon (2012) also discusses the potential costs and benefits of debriefing public officials who have been part of an experiment, noting that such approaches may well create negative feelings or even hostility in public officials, but arguing that a dialogue-based, educational debriefing is part of "fully respecting subjects." We think the costs of debriefing public officials about experiments may be especially high when considering the social costs that could arise from debriefing; these include the possibility that street-level bureaucrats will not respond to genuine citizen requests in the future out of concern that they might be part of an experimental treatment. It is also not clear that debriefing would change these officials' or citizens' assessment of whether the underlying study was ethical; in a survey of the general public, debriefing did not change perceptions of whether an audit experiment on politicians was "acceptable" (Desposato, 2018).

A key concern in this debate is the use of deception. Some recent work has highlighted the potential to avoid deception by partnering with "confederates who might sincerely undertake the activity at the heart of the research" (Findley and Nielson, 2016). Recent experiments on US legislators highlight both the logistical challenges of such an approach and its promise for maintaining realism while avoiding explicit deception (Bergan, 2009; Bergan and Cole, 2015; Butler, Karpowitz and Pope, 2012). Particularly instructive is Kalla and Broockman (2016), which takes on the question of whether members of Congress are more accessible to campaign donors than to other constituents. Rather than running an audit experiment with deceptive emails, the authors partnered with a political organization to contact congressmembers on behalf of real constituents in their districts, who would actually attend a meeting if one were scheduled (Kalla and Broockman, 2016).[17]

This approach – of working with real confederates who have standing to make requests of government – also has promise in the study of street-level bureaucracy.

[17] Rather than inventing campaign-donor status, the project began with a pool of campaign donors and then the experimental manipulation took the form of selectively revealing that status for some constituents and not others.

Indeed, the study of drivers' licenses discussed above uses this approach in its main experiment: rather than only sending actors to these offices to pretend to want drivers' licenses, the researchers randomized real people actually seeking licenses into groups with more or fewer qualifications for a license (via driving lessons) and more or less incentive to pay bribes (via a bonus if they got a license) and observed the bureaucratic outcomes (Bertrand et al., 2007).

We acknowledge that it will not always be feasible to mobilize members of the public to contact public officials,[18] but we think this approach of working with real people has promise for the study of street-level bureaucrats as well as other public officials.

Burdens to citizens: opportunity costs and spillovers Institutional Review Boards (IRBs) mainly (or exclusively) consider possible risks to experimental subjects when evaluating an experiment's ethics. But experiments on street-level bureaucrats carry another set of risks that may be more consequential: risks to *non-subjects*, such as other citizens who interact with the bureaucrats included in an experiment. Risks to non-subjects come in two forms: as opportunity costs, with experimental treatments diverting scarce bureaucratic time and resources away from real constituents; and as spillover effects, with behavioral changes induced by the experiment carrying over to bureaucrats' interactions with others.

Opportunity costs are a particular concern in audit experiments. Time spent responding to queries from the audit is time not spent serving real citizens. Many audits, including our own (White, Nathan and Faller, 2015), attempt to mitigate these costs by making the experimental interactions extremely brief. But opportunity costs can grow quickly if the experimental design calls for repeated contacts, or if the same small pool of subjects is studied multiple times. Indeed, even a very simple audit experiment can create real time burdens on bureaucrats if the experiment's deception is uncovered.

[18] We also note that mobilizing members of the public can carry other ethical questions, like whether it is acceptable to induce people to take risky actions or break the law (see below).

Returning to the anecdote from above, when a Secretary of State's office contacted us because two studies by other scholars had re-treated the same pool of local election officials and been discovered, it could no longer be plausibly claimed that these new experiments had only taken a negligible amount of time away from official activities. Kovaleski (2016) reports that the FBI and officials from thirteen different states became involved in the investigation. They spent valuable effort trying to figure out if the experiments were a form of election interference less than a week before the 2016 election, at a time that their offices were busy with real duties.

There are also risks that experimental treatments affect bureaucrats' interactions with other citizens. In the context of an audit experiment whose cover is blown, this may manifest in bureaucrats being less responsive to real constituents in the future because they now believe they are also fictitious. But this second concern extends beyond audits.

Consider Bertrand et al.'s (2007) experiment on driver's licenses. Their study entailed some direct risk to subjects: the "bonus" treatment incentivized participants to commit the (minor) crime of bribing officials.[19] Potentially more challenging is that such a study could affect non-participants who also seek licenses from the same offices. The experiment gave 268 participants in the "bonus" condition a financial incentive to pay bigger bribes at four specific offices over several months. The researchers took care to space treated participants out, staggering them into waves every two weeks. But a similar experiment conducted at a larger scale, or in a narrower time frame, might change the local market for bribes by flooding offices with applicants willing to pay more than usual. Bureaucrats could respond by raising bribe prices, reducing access to licenses among non-participants. Or an uptick in bribery caused by the experiment could attract attention, sparking a crackdown that ensnares bureaucrats

[19] There should be debate over whether such a treatment is ever ethical. But we see how a good-faith defense could be made: participants gave informed consent and were not explicitly told or forced to pay bribes. Bribery was also endemic already – from the control group it is clear that many treated participants would have paid bribes anyway.

or delays non-participants' ability to get licenses. To be clear, we have no evidence that either happened during Bertrand et al.'s experiment. But the study design points to the possibility of such spillovers.

We see several steps that researchers could take to mitigate both sets of risks. First, field experiments on street-level bureaucrats should be expected to provide an explicit ethical defense of why the benefits of the experiment are worth the risks. This defense should not focus only on risks to subjects, but also to non-subjects. Journal editors may be those best positioned to enforce such an expectation by requiring it for publication of these types of experiments.

Slough (2018) provides a model for what this can look like, at least for concerns about wasting officials' time. By timing the phone calls in her audit, she directly calculates the bureaucrats' time that she has used.[20] Drawing on publicly available wage data, she then estimates the direct labor cost of this time as just $2,644 USD total, spread across more than 600 municipalities. This type of explicit calculation is at best a first cut, as it includes only the most directly quantifiable costs of the study. Tough questions remain, especially about who gets to decide that the unquantifiable, non-monetary benefits of the study exceed this cost, and about how to factor in the less-easily-quantified risks we discuss above. But even basic and incomplete cost estimates like this are useful, as they force scholars to begin engaging explicitly with the burdens their experimental design may impose.

Further, we encourage researchers doing explicit cost-benefit calculations to be both concrete and realistic about the expected benefits of their research. If a study is going to be the fifth or tenth of a particular phenomenon, what is the marginal knowledge to be gained from the study, and does it warrant the potential costs to the public? And if there is substantial knowledge to be gained from such a study, how will it be disseminated such that the people bearing the risks of the research will be able to benefit from the study's findings? Researchers should clearly state any anticipated benefits to,

[20] For more on such calculations of elite time, see Grose's chapter in this volume.

or planned communication with, people in the specific experimental contexts, rather than imagining that everyone generically benefits from academic knowledge – especially given how inaccessible most political science research is to policy-makers and the public.

One potential best practice, which Slough (2018) admirably employs, is to collaborate with the leaders of the bureaucracy being studied; this will both allow agency leaders to make their own informed decisions about what costs are acceptable and will set researchers up to disseminate their findings directly to the actors most likely to be able to use them to change substantive policy outcomes. However, if agency leaders decline an invitation to participate, scholars must carefully weigh their ethical and legal obligations to respect agency leaders' wishes against the potential public benefits of information that could be gleaned from still continuing with a study.

Second, researchers must be thoughtful about the statistical power needed for their studies, as well as their decisions to re-examine the same limited populations of bureaucrats as earlier studies. Better statistical power is generally a good thing, and replication is essential to the scientific process. But these benefits must be weighed against the risks of "spoiling" the pool for future research (McClendon, 2012), and, more importantly, changing the behavior of bureaucrats in ways that affect non-subjects.

In particular, because the pool of relevant bureaucratic subjects can sometimes be quite small, experimental researchers studying street-level bureaucracy can confront a thorny collective action problem: even if each individual study carries little risk, the cumulative risk of multiple studies on the same population may become unacceptably large. There are no easy solutions to this collective action problem. At a minimum, as described above, we believe future researchers should take advantage of the growing use of pre-registration (e.g., Humphreys et al. 2013) to check existing experimental registries carefully for other studies on the same population before beginning any new experiment, especially an audit experiment that employs deception. This will allow for better assessment of relevant risks and allow scholars to adapt their new research plans to reduce these risks.

Further, we encourage researchers to use their pre-experimentation power analyses to think more carefully about how big a sample they really need for a given study. In laboratory experiments with paid subjects or face-to-face survey interviews in developing countries, this kind of calculation is common: when each additional observation is expensive, researchers are incentivized to collect enough observations for a credible study, but not too many more to avoid inflating the study's budget. But where adding additional observations is relatively low-cost to researchers, as in an email-based audit experiment, experimenters might simply collect as many observations as is feasible, without considering whether they are "overpowered" for the question at hand. We urge researchers to think about the social costs of additional observations in field experiments just as they might think about budgetary costs in other forms of research, and avoid running experiments that are larger than needed to measure effects of an expected size. This will help limit ethical risks.[21]

The ethics of collaborating with bureaucracies Above, we discuss the pros and cons of collaborating with bureaucracies from a design perspective. Collaboration also has ethical benefits and drawbacks. Humphreys (2015) provides a framework through which to weigh collaboration in field experiments by considering the separate and overlapping "spheres of ethics" between researchers and cooperating agencies. If collaborating agencies genuinely consent to partnering for a study, Humphreys (2015) argues that primary ethical responsibility for the experimental intervention will rest with the government agency implementing it, not the researcher.[22] This can have clear benefits. Agency supervisors are in a far stronger ethical position to decide whether it is appropriate to experiment on their employees than outside researchers and university

[21] Future work could also explore adapting sequential hypothesis testing approaches that have been used to limit sample sizes in medical trials of risky drugs or devices, though we note concerns about bias from these approaches (Liu and Hall, 1999; Tartakovsky, Nikiforov and Basseville, 2014).

[22] Humphreys (2015) notes the risks, however, that government partners may feel pressured into implementing a treatment by offers of valuable funding from the researchers, or may not fully understand what the researchers are proposing when they agree to collaborate. If either is the case, researchers still hold primary ethical responsibility for the intervention.

IRBs. Moreover, perhaps the most legitimate people to decide what opportunity costs and burdens to citizens are acceptable are the political leaders accountable to those citizens.

But Humphreys (2015) also warns against passing off too much ethical responsibility to collaborating institutions. Researchers are still complicit in all aspects of field experiments conducted on their behalf, and should not hide behind a "spheres of ethics" argument if the intervention implemented by a government partner would not otherwise be ethical under the standards of the research community. We do not believe any of the example studies above cross this line, though readers may disagree. But for several, we can see a slippery slope in which a more involved intervention, or a similar intervention viewed from another perspective, could seem unethical.

For example, Yokum et al. (2019) collaborate with a major US police department with the goal of reducing citizen complaints and cutting back on officers' misconduct. This seems valuable, particularly given that it would be nearly impossible to test the effects of a policy like body cameras without such a collaboration. Yet similar collaborations with police could quickly become more fraught, especially if some citizens perceive that researchers are helping the police become more effective at punishment or repression. Similarly, at first blush the ethics of collaboration in the intervention targeted at Indian nurses in Banerjee et al. (2008) seem fine: by partnering with a district health ministry, the researchers gained a unique opportunity to evaluate and refine a new policy aimed at improving service delivery for the rural poor. But to a union representative or labor activist, this same partnership could seem more compromising. One of the underlying reasons nurses' initial attendance was so low is because they are poorly paid and seek outside employment to supplement their public sector incomes. Partnering with the nurses' employer to perfect a punitive incentive structure that docks pay for low attendance may also seem like unethically helping the government better exploit their staff while letting politicians off the hook for underfunding the bureaucracy in the first place.

Ultimately, we see no easy answers to which collaborations are ethical and which are not. But we urge researchers considering government collaborations for their design benefits to also consider how the ethics of these partnerships appear from other perspectives. This stance is particularly important when the street-level bureaucrats in question, or the people they serve, are from communities of which the researcher is not a part.

5 Conclusions and Future Directions

We have described recent advances in field experimental research on street-level bureaucracy conducted by scholars of American politics and the political economy of development. Despite their different research environments, scholars in both areas have found field experiments increasingly useful as measurement devices, and as means to test theories of bureaucratic politics. Scholars in both sub-fields also face similar challenges in designing effective experiments and must weigh many of the same ethical trade-offs.

Going forward, we see several substantive areas that would benefit from greater focus from experimental scholars. First, as summarized above, much of the recent experimental literature focuses on monitoring interventions aimed at mitigating the principal-agent problem or reducing discrimination and malfeasance. These studies are rooted in a theoretical expectation that too much discretion for street-level officials allows for shirking and malpractice (Lipsky, 1980). But recent non-experimental scholarship, such as Honig (2018), increasingly emphasizes the *virtues* of discretion, arguing that successful policy implementation requires giving frontline agency staff the flexibility to innovate and adapt in response to conditions on the ground, free from one-size-fits-all monitoring criteria. The literature would benefit from similarly rigorous experimental tests of arguments like Honig's (2018) to better spell out the pros and cons of discretion in different domains.

Second, very little recent experimental work focuses on another key determinant of street-level bureaucrats' behavior: *who they are.* But a large body of non-experimental scholarship in both American and Comparative politics suggests that the identities of frontline workers – their ethnicity, gender, and partisanship – can be crucial to understanding bureaucratic outcomes (Krislov, 1974; Keiser et al., 2002), especially in settings where discretion and monitoring are difficult for principals to change (Hassan, 2020). Indeed, recent research on bureaucracy in the developing world emphasizes that state elites often seek to solve principal-agency problems and mitigate shirking not through enhanced monitoring, but through the strategic placement and rotation of bureaucrats with different identities into jurisdictions that require different types of behavior (Landry, 2008; Bhavnani and Lee, 2018; Carter and Hassan, 2019; Hassan, 2020). We see fertile ground for more direct experimental tests of these and related theories.

Finally, our discussion suggests several methodological and ethical "best practices" for future scholars in both sub-fields. In particular, we believe that scholars will benefit from:

1. Checking registries of Pre-Analysis Plans, such as the EGAP repository, before beginning new experiments to assess the risks of over-studying the same populations of bureaucrats.

2. Investing in greater methodological effort to better quantify the most relevant outcomes of interest in order to more fully realize the scholarly benefits of an experiment.

3. Where possible, partnering with real citizens to implement experimental treatments, rather than using deception via fictitious contacts.

4. Explicitly defending their ethical choices in pre-analysis plans and published papers.

5. Avoiding "over-powered" studies by better internalizing the inferential and ethical

costs of “spoiling the pool”, diverting bureaucratic resources, and/or creating spillover effects on non-participants into decisions about sample sizes.

6. Where feasible, collaborating directly with the bureaucratic agencies being studied, including by allowing appropriate local partners to make their own ethical judgments about what risks are acceptable *vis-a-vis* the potential scholarly benefits of an experiment.

References

Banerjee, Abhijit V., Rachel Glennerster and Esther Duflo. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." *Journal of the European Economic Association* 6(2):487–500.

Bergan, Daniel E. 2009. "Does grassroots lobbying work? A field experiment measuring the effects of an e-mail lobbying campaign on legislative behavior." *American politics research* 37(2):327–352.

Bergan, Daniel E and Richard T Cole. 2015. "Call Your Legislator: a field experimental study of the impact of a constituency mobilization campaign on legislative voting." *Political Behavior* 37(1):27–42.

Bertrand, Mariane, Simeon Djankov, Rema Hanna and Sendhil Mullainathan. 2007. "Obtaining a Drivers' License in India: An Experimental Approach to Studying Corruption." *Quarterly Journal of Economics* 122(4):16391676.

Bhavnani, Rikhil R. and Alexander Lee. 2018. "Local Embeddedness and Bureaucratic Performance: Evidence from India." *Journal of Politics* 80(1):71–87.

Brehm, John and Scott Gates. 1997. *Working, Shirking, and Sabotage: Bureaucratic Response to a Democratic Public.* Ann Arbor, MI: University of Michigan Press.

Bruch, Sarah, Myra Ferree and Joe Soss. 2010. "From Policy to Polity: Democracy, Paternalism, and the Incorporation of Disadvantaged Citizens." *American Sociological Review* 75(2):205–226.

Butler, Daniel M, Christopher F Karpowitz and Jeremy C Pope. 2012. "A field experiment on legislators home styles: service versus policy." *The Journal of Politics* 74(2):474–486.

Carter, Brett L. and Mai Hassan. 2019. "Regional Governance in Divided Societies: Evidence from the Republic of Congo and Kenya." Forthcoming, *Journal of Politics*.

Coppock, Alexander. 2019. "Avoiding Post-Treatment Bias in Audit Experiments." *Journal of Experimental Political Science* 6(1):1–4.

Desposato, Scott. 2018. "Subjects and Scholars Views on the Ethics of Political Science Field Experiments." *Perspectives on Politics* 16(3):739–750.

Fang, Albert H, Andrew M Guess and Macartan Humphreys. 2019. "Can the government deter discrimination? Evidence from a randomized intervention in New York City." *The Journal of Politics* 81(1):127–141.

Findley, Michael and Daniel Nielson. 2016. Obligated to Deceive? Aliases, Confederates, and the Common Rule in International Field Experiments. In *Ethics and Experiments. Problems and Solutions for Social Scientists and Policy Professionals*, ed. Scott Desposato. pp. 151–70.

Grose, Christian R. 2014. "Field experimental work on political institutions." *Annual Review of Political Science* 17:355–370.

Hassan, Mai. 2020. *Regime Threats and State Solutions: Bureaucratic Loyalty and Embeddedness in Kenya.* New York: Cambridge University Press.

Hess, Douglas R., Michael J. Hanmer and David W. Nickerson. 2016. "Encouraging Local Compliance with Federal Civil Rights Laws: Field Experiments with the National Voter Registration Act." *Public Administration Review* 76(1):165–174.

Honig, Dan. 2018. *Navigating by Judgment: Why and When Top Down Management of Foreign Aid Doesn't Work.* New York: Oxford University Press.

Humphreys, Macartan. 2015. "Reflections on the Ethics of Social Experimentation." *Journal of Globalization and Development* 6(1):87–112.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt. 2013. “Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration.” *Political Analysis* 21(1):1–20.

Kalla, Joshua L and David E Broockman. 2016. “Campaign contributions facilitate access to congressional officials: A randomized field experiment.” *American Journal of Political Science* 60(3):545–558.

Keiser, Lael R., Vicky M. Wilkins, Kenneth J. Meier and Catherine A. Holland. 2002. “Lipstick and Logarithms: Gender, Institutional Context, and Representative Bureaucracy.” *American Political Science Review* 96(3):553–564.

Kovaleski, Tony. 2016. “FBI probes emails sent to county clerks across Colorado and 13 other states.” *The Denver Channel* .

URL: *https://www.thedenverchannel.com/news/investigations/fbi-probes-emails-sent-to-county-clerks-across-colorado-and-12-other-states*

Krislov, Samuel. 1974. *Representative Bureaucracy.* Englewood Cliffs, NJ: Prentice-Hall.

Landry, Pierre F. 2008. *Decentralized Authoritarianism in China: The Communist Party’s Control of Local Elites in the Post-Mao Era.* New York: Cambridge University Press.

Lipsky, Michael. 1980. *Street-Level Bureaucracy : Dilemmas of the individual in public services.* New York: Russell Sage Foundation.

Liu, Aiyi and WJ Hall. 1999. “Unbiased estimation following a group sequential test.” *Biometrika* 86(1):71–78.

McClendon, Gwyneth H. 2012. “Ethics of Using Public Officials as Field Experiment Subjects.” *Newsletter of the APSA Experimental Section* 3(1).

Neggers, Yusuf. 2018. "Enfranchising Your Own? Experimental Evidence on Bureaucrat Diversity and Election Bias in India." *American Economic Review* 108(6):1288–1321.

Slough, Tara. 2018. "Bureaucrats Driving Inequality in Access: Experimental Evidence from Colombia." pp. 1–55.
URL: *http://taraslough.com/assets/pdf/JMP.pdf*

Soss, Joe. 1999. "Lessons of welfare: Policy design, political learning, and political action." *American Political Science Review* 93(2):363–380.

Soss, Joe and Vesla Weaver. 2016. Learning from Ferguson: Welfare, Criminal Justice, and the Political Science of Race and Class. In *The Double Bind: The Politics of Racial and Class Inequalities in the Americas, A Report of the Task Force on Racial and Social Class*, ed. Juliet Hooker and Jr Alvin B. Tillery. Washington, DC.: American Political Science Association.

Tartakovsky, Alexander, Igor Nikiforov and Michele Basseville. 2014. *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC.

Teele, Dawn Langan. 2014. "Reflections on the ethics of field experiments." *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences* pp. 115–140.

White, Ariel R., Noah L. Nathan and Julie K. Faller. 2015. "What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials." *American Political Science Review* (February):1–14.

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.