Can the Government Deter Discrimination? Evidence from a Randomized Intervention in New York City

Albert H. Fang
Postdoctoral Associate
Institution for Social and Policy Studies
Yale University
77 Prospect Street
New Haven, CT 06520
albert.fang@yale.edu

Andrew M. Guess
Assistant Professor
Department of Politics and Woodrow Wilson School
Princeton University
Robertson Hall
Princeton, NJ 08544
aguess@princeton.edu

Macartan Humphreys
Professor
Department of Political Science
Columbia University
420 W. 118th Street, 7th Floor
New York, NY 10027
mh2245@columbia.edu

December 4, 2017

Abstract

Racial discrimination persists despite established anti-discrimination laws. A common government strategy to deter discrimination is to publicize the law and communicate potential penalties for violations. We study this strategy by coupling an audit experiment with a randomized intervention involving nearly 700 landlords in New York City and report the first causal estimates of the effect on rental discrimination against Blacks and Hispanics of a targeted government messaging campaign. We uncover discrimination levels higher than prior estimates indicate, especially against Hispanics, who are approximately six percentage points less likely to receive callbacks and offers than whites. We find suggestive evidence that government messaging can reduce discrimination against Hispanics, but not against Blacks. The findings confirm discrimination's persistence and suggest that government messaging can address it in some settings, but more work is needed to understand the conditions under which such appeals are most effective.

KEYWORDS: government communication; discrimination; political economy of race; behavioral policy compliance; field experiment

Supplementary material for this article is available in the appendix in the online edition. Replication files are available in the JOP Data Archive on Dataverse (http://thedata.harvard.edu/dvn/dv/jop). This research and the authors' involvement with the City of New York in this research were reviewed and approved by Columbia University Morningside IRB (Protocols IRB-AAAI4007 and IRB-AAAN4059). A pre-analysis plan for this study was submitted to the Experiments in Governance and Politics (EGAP) registry on April 3, 2013, while data collection was ongoing but prior to any analysis (ID 20130403AA). This study was conducted with the generous support of the New York City Commission on Human Rights, the Russell Sage Foundation (RSF Award No. 87-13-02), and the Columbia Center for the Study of Development Strategies.

The passage of civil rights laws in the 1960s marked a turning point in the development of racial politics in the United States. Against the historical backdrop of slavery, Jim Crow, longstanding racial inequalities, and the marginalization of non-whites, the legal prohibition of discrimination on the basis of race, color or national origin signaled an important political shift toward a more egalitarian racial order (King and Smith 2005; Massey and Denton 1993). In addition to examining the political conditions leading to the passage of civil rights laws, scholars examining the politics surrounding the formal end of Jim Crow have focused on its aftermath. Scholars have documented how racial inequalities have persisted, and have offered numerous policy-centered, political economy explanations for their persistence (e.g., Alexander 2012; Forman 2012). Central to these explanations is the argument that racial disparities persist because racial discrimination is both pervasive and persistent despite established civil rights laws (Dawson 2016; Dawson and Francis 2015; Bobo, Kluegel and Smith 1997). This calls into question the role of government in enforcing these laws, as well as the downstream consequences of government action (or inaction) for racial inequalities, the political economy of race, and the development of racial orders.

Understanding this puzzle requires furthering research around two related lines of inquiry. The first asks: *To what extent does racial discrimination persist and why?*[1] The first-order challenge of measuring discrimination levels is not trivial. For its potential targets, discrimination is difficult to detect because in many cases it is impossible to observe how a counterfactual individual would be treated in a market interaction. Going beyond first-hand accounts, relying on third-party reports is imperfect because they require unobtrusive observation and, as prior psychological research has documented, many individuals are unwilling to "see" or make attributions to discrimination even when confronted with direct evidence of it (Major and Dover 2016). To address this challenge, governments use enforcement audits to collect evidence of discrimination on a case-by-case basis, and both scholars and governments have relied on audit studies to measure the aggregate-level

[1] We focus on understanding the incidence of discrimination and set aside questions about its causes, which have been widely studied in economics and sociology (e.g., Heckman 1998; Heckman and Siegelman 1993; Bertrand, Chugh and Mullainathan 2005; Pager and Karafin 2009).

incidence of discrimination in employment and housing.[2] Despite their proliferation, high-quality audits are both costly and relatively rare. Most audit studies focus on discrimination in early-stage market interactions but do not measure end-line outcomes (such as being offered a job or a housing unit), which are important to measure because they are key indicators of disparate impact. The second, related line of inquiry asks: *Which government strategies are effective at reducing discrimination and why?* Despite an abundance of theoretical work on policy enforcement in general and research on the primacy of legal strategies in efforts to enforce anti-discrimination laws since their initial passage,[3] to our knowledge no work exists that theorizes the conditions under which governmental strategies to reduce discrimination are effective and that experimentally tests hypothesized expectations. Answering this second question is doubly difficult as it requires measuring a hidden behavior and causal inferences around those behaviors.

In this article, we address both of these major questions and provide the first attempt in the literature to experimentally test the effects on racial discrimination levels of any government strategy to reduce it in the United States. We study the effectiveness of pre-emptive strategies to deter racial discrimination using official communication campaigns encouraging compliance with anti-discrimination law. We examine two commonly employed but understudied appeals: making the law itself salient and making the costs of violating the law salient. We work at a large scale which offers a good handle on discrimination levels although, given fundamental measurement and inferential challenges, our treatment effect estimates are still measured with considerable uncertainty.

Our analysis examines discrimination and interventions in field conditions, in the context of government efforts to enforce fair housing law in the New York City rental market. We focus on the rental market as it is the segment of the housing market in which reported discrimination is the most pervasive.[4] We partnered with the New York City municipal government to implement

[2] See Bertrand and Duflo (2017), Pager and Shepherd (2008), Turner et al. (2013), Edelman, Luca and Svirsky (2017).

[3] See Rosenberg (1991), Cover (1995), Epp (1998), Frymer (2003), and Skrentny (2002).

[4] Of the 18,978 housing discrimination complaints reported by private fair housing groups in 2014,

a large-scale field experiment that began in 2012 and lasted 20 months. We assess the effects on racial discrimination levels against Black and Hispanic rental applicants (as compared to whites) of government appeals that are delivered via targeted and personalized phone calls to landlords and brokers[5] who interact with confederates posing as rental housing applicants. The city randomly assigned to nearly 700 landlords either (1) a targeted live phone call from the city that drew attention to fair housing law and implicitly signaled increased government monitoring of housing agents (a "monitoring" condition), (2) a targeted live phone call with the contents of the monitoring message and additional information about the potential pecuniary punishments for discriminatory practices (a "punitive" condition), or (3) no call from the city (a pure control condition). The city then sent matched triples of confederates who varied by race (Black, Hispanic, and white) to visit the same advertised unit—meeting in person with the same landlord—and to collect detailed qualitative notes on their interactions over the course of the housing search process. The enhanced audit design embedded in the experiment provides leverage above and beyond existing audit methodologies to measure both subtle and direct forms of racial bias that pervade the housing market. Additionally, in contrast to prior housing audit studies focusing on differential treatment in landlords' responses to initial inquiries about rental listings, this study focuses on two outcomes occurring toward the end of the housing search process that are consequential for downstream housing inequalities: discrimination in receiving a callback following an in-person meeting with the landlord to view the listed unit and discrimination in receiving a rental offer from the landlord after the appointment.

The theoretical framework motivating the experiment is relatively simple. Drawing on existing research, we expect that the monitoring condition should decrease racial discrimination levels relative to control, because receiving a personally targeted monitoring signal from a government known to actively enforce fair housing law is expected to reduce landlords' perceived benefits from

16,694 (or about 88%) occurred in the rental market (National Fair Housing Alliance 2014).

[5] For simplicity, we refer to subjects, defined as landlords or brokers associated with a sampled rental housing ad and who interact with a matched team of testers, simply as "landlords."

discriminating and to increase their perceptions of the probability of punishment if they discriminate. By contrast, existing research offers mixed expectations about the conditions under which coercive sanctions induce pro-social behavior, with some proposing possibly adverse effects; thus we are agnostic about the expected direction of the punitive messaging effect.

We find evidence of substantial baseline levels of discrimination in the New York City rental housing market, particularly against Hispanics: they are 28% less likely to receive a callback (6.1 percentage points, 95% C.I. = [1.00, 11.19]) and 49% less likely to receive an offer for an apartment than whites (5.7 percentage points, 95% C.I. = [1.32, 10.16]). This result is striking when compared to discrimination levels reported by the 2012 US Department of Housing and Urban Development (HUD) audit of racial rental discrimination in the New York metropolitan area, which found a lack of discrimination against minorities in receiving callbacks (discrimination in receiving offers was not studied).[6] We show that racial discrimination in fact persists in the New York City rental market. We also find suggestive evidence that treatment messages reduce levels of discrimination in receiving callbacks, though these effects are statistically weak and are observed only for discrimination against Hispanics. There is no evidence that discrimination against Blacks is affected by these government messaging interventions. We also find no evidence that the punitive messages are more effective than simple monitoring messages. Our results suggest that a bundled governmental messaging strategy that invokes fair housing law and makes salient the costs of punishment for lawbreakers can be effective at reducing racial housing discrimination, but not under all circumstances. Taken together, this study makes the case that evaluating government efforts to reduce racial discrimination is central to understanding the political economy of race, and motivates the need for additional future research on the conditions under which government efforts to enforce anti-discrimination laws and reduce discrimination are effective.

[6] The estimates from both samples condition on matched testers making an appointment, but the samples also differ. While our sample is larger, the 2012 HUD study had a larger initial sampling frame (including Northern New Jersey and units found through non-Craigslist means) whereas ours only includes New York City rental listings from Craigslist.

Theoretical Framework and Expectations

Government Strategies to Reduce Racial Discrimination in the United States

Government strategies to enforce anti-discrimination law may be classified into one of two broad types: reactive or pre-emptive. Reactive strategies such as “fire-alarm” and “police patrol” models of policy enforcement (McCubbins and Schwartz 1984) are those in which governments investigate and litigate potential discriminators as an *ex post* response conditional on citizens and governments, respectively, observing and reporting discrimination in the first place. Given the challenges with observing discrimination (as previously discussed) and the expectation that pursuing litigation may incur disproportionate financial and psychological costs on the targets of discrimination relative to defendants (Feagin 1991; Donohue and Siegelman 1991; Galanter 1974), reactive strategies are limited in their ability to identify, much less punish, most who discriminate.

In contrast, pre-emptive strategies are those where official messaging campaigns attempt to deter discrimination *ex ante* by advertising existing anti-discrimination laws and making various appeals that attempt to induce compliance with the law. Such campaigns complement reactive strategies and are attractive in part because, if effective, they could offer a lower-cost method to induce behavioral change (as compared to deploying an intensive enforcement audit program). Existing government communication campaigns typically involve bundled messages and appeals that aim to reach multiple potential audiences including the targets of discrimination, discriminators, and third-party observers. Evaluating existing bundled government campaigns does not provide leverage to understand why enforcement efforts are effective, however, because it would not be possible to infer whether a potential lawbreaker decides to comply with the law because they perceive the costs of violating the law to outweigh the benefits (Wilson 1980). To disentangle the effects of different governmental appeals on individuals whose behavior fair housing laws ultimately aim to change, we focus on testing whether governmental appeals that directly target potential discriminators (i.e., landlords and brokers) and that are designed to alter their expected returns to discriminating are effective at inducing behavioral change.

When Do Messaging Campaigns Reduce Discriminatory Behavior?

Under what conditions might government messaging campaigns be effective at reducing discriminatory behavior? Existing research undertheorizes why official appeals would be effective at reducing discrimination. We develop a simple decision-theoretic model to clarify the basic logic. In the context of a housing market interaction, the individual of interest is the landlord or broker who interacts with housing applicants and may potentially engage in discriminatory behavior. Discrimination (as an individual-level behavior) simply refers to the differential treatment of two individuals who vary on a relevant attribute such as race. This definition is agnostic as to why people discriminate, which may arise because people hold group-specific prejudices and a taste for discrimination (Becker 1957) or because people engage in statistical discrimination by relying on average group stereotypes to make inferences about a person perceived to belong to that group (e.g., Phelps 1972; Arrow 1973; Aigner and Cain 1977). Let individual i's utility from discrimination given government policy x be expressed as $u_i(x) = b_i(x) - p_i(x)c_i(x) + e_i$ where b_i denotes marginal benefits from discriminating, p_i the subject's expectation of punishment when she discriminates, c_i the expected severity of the punishment, and e_i individual heterogeneity. If $b_i(x) > p_i(x)c_i(x)$ then i will discriminate.

Building on this framework we aim to estimate both the baseline level of housing discrimination against Black and Hispanic renters as well as the causal effects on racial housing discrimination of governmental messaging that encourages landlords and brokers to comply with fair housing law. We investigate these quantities separately for Blacks and for Hispanics because in multiracial contexts, there are multiple dimensions of social difference that are potentially salient and the nature of bias against a racial group is expected to vary as a function of how each group is perceived to deviate from a majority reference group (e.g., Dovidio et al. 2010; Pehrson, Vignoles and Brown 2009).[7] We examine two distinct types of policy: a monitoring signal that invokes the law and implicitly signals increased government monitoring of housing agents, and a punitive appeal that

[7] Our study therefore addresses the need for additional research on discrimination against Hispanics (Dovidio et al. 2010) and the need to study discrimination against multiple groups in the same study

additionally primes the costs of discrimination. For practical reasons—in particular to satisfy our implementing partner's preference for treatment realism and to avoid depleting statistical power—the experiment is designed to test realistic, bundled treatments, but does so at the expense of testing specific parameters. We are cognizant that this is a major limitation of this study and recommend future work to test treatments that target each parameter separately. This framework is nevertheless useful to develop expectations about why the appeals we test would affect discrimination.

We test the hypotheses that sending monitoring messages alone decreases rental housing discrimination rates against Blacks and Hispanics (hypothesis family *H1*). There are two possible reasons for why it might. First, simply contacting landlords and making the law salient may raise expectations of sanctioning, p, even if this is not explicitly invoked and no information on sanctions is provided. Given the context of this study, the mode of treatment delivery, and the particular messenger of the treatment appeal, we argue that this is plausible in this experimental setting. Because the treatment calls in this study are personalized and targeted and because they are sent by a city government known to have the capacity and willingness to monitor and enforce fair housing law, receiving such a call provides a strong and credible signal to the landlord that the city is already monitoring them. Thus receiving any targeted call from the city may increase p above zero.

Increasing the salience of injunctive norms about fair housing in the landlord's thinking may also reduce b. Simply invoking the law may activate compliance norms and may also render compliance focal (McAdams and Nadler 2005; Tyler 2006). For discrimination in particular, Mendelberg (2001) and others have highlighted how *explicit* priming of discrimination considerations can result in less discrimination (relative to implicit priming) due to an invoking of social nondiscrimination norms. If housing agents perceive discrimination to be a descriptive norm, then these perceptions may drive beliefs that the law is illegitimate and result in less compliance (Tyler 2004, 2006). But appeals to injunctive norms have been theorized to be effective at crowding out descriptive norm perceptions used to justify noncompliance (Weaver 2015). Our estimates of the monitoring effect thus capture joint effects on p and, to the extent that preferences are not fixed, b.

to understand discrimination in a multiracial context, which is rare in the literature.

In addition, we study the effect of altering *c* by assessing the effect of sending "punitive" messages, relative both to the control condition (thus the combined effect of monitoring and punitive content) and to the monitoring condition (thus the additional effect of punitive content as compared to the monitoring condition). Thus in practice we assess the hypotheses that sending punitive messages decreases rental housing discrimination rates against Blacks and Hispanics relative to the baseline condition and to a monitoring condition (hypothesis family *H2*).[8] Although we have clear expectations about the marginal effect of costs, *c*, we are cognizant that our punitive treatment may also have the effect of altering *b*, relative to the monitoring treatment.

Existing research suggests mixed expectations about whether amplifying the salience of the severity of punishment will be effective at reducing discrimination. On the one hand, theoretical work on the politics of inducing behavioral policy compliance argues that when individuals face insufficient incentives to comply or perceive complying as costly, a potentially effective strategy to induce compliance involves increasing and making salient the costs of noncompliance such that they are perceived to be greater than the costs of compliance (e.g., Dixit 2006; Hadfield and Weingast 2014; Weaver 2014, 2015). Prior research evaluating the effectiveness of similar official communication campaigns has shown that appeals highlighting the costs of noncompliance can, under certain circumstances, induce greater compliance with the law in domains such as paying taxes (Blumenthal et al. 2001; Slemrod, Blumenthal and Christian 2001; Iyer, Reckers and Sanders 2010, but see Dunning et al. (2015)) and paying delinquent fines (Haynes et al. 2013). On the other hand, a growing line of social psychological research challenges this expectation and argues that coercive sanctions to induce behavioral change may backfire and instead crowd out social norms needed to motivate pro-social behavior (e.g., Bowles 2016; Gneezy and Rustichini 2000; Cardenas, Stranlund and Willis 2000). Thus while we gather direct evidence on the effect of the punitive treatment, this can be interpreted as the effect of costs, *c*, only under the assumption of no *differential* effects, relative to the monitoring treatment, on *b* and *p*.

[8] The latter comparison can be interpreted as the pure effect of explicitly priming punishment.

Experimental Design

Partnering with the City of New York, we designed and analyze data from a field experiment where the city randomly sent targeted messages to landlords associated with a specific advertised rental unit, which is pursued by a trio of matched testers who vary by race. The matched audit provides leverage to measure racial discrimination, which we operationalize as differential landlord behavior toward testers by the tester's race for the same advertised unit. This section describes the experiment's setting, design, implementation, and analysis.

Study Context

New York City is a useful political setting for testing the effectiveness of governmental campaigns encouraging compliance with fair housing law. It is well known for having one of the strongest anti-discrimination laws in the country (Gurian 2005). The agency enforcing the city's anti-discrimination law, the New York City Commission on Human Rights, is also well known among real estate professionals for having the capacity and willingness to enforce fair housing law. Thus, receiving any targeted and personalized treatment phone message from this city government is arguably perceived as a credible signal of increased government effort to monitor landlords and enforce fair housing law. We focus on rental listings posted on the online classified advertising site Craigslist. While Craigslist is one of multiple sources of classified rental advertisements, it is one of the primary forums used to post and pursue rental listings in general and in New York City.

Audit Design

We briefly summarize how rental listings are sampled from Craigslist and how they are pursued by matched auditors.[9] On each day of the study's implementation, a set of rental housing ads from the current day was selected from Craigslist using an automated script. First, using keywords, a list

[9] Full details about the sampling, audit, and field procedures are provided in the online Supplemental Information (SI).

of "likely discrimination" (LD) ads were identified and selected with 100% probability with the goal of increasing statistical power if baseline discrimination levels were low.[10] Second, among the remainder of ads posted on Craigslist that day (excluding those identified in the LD search), a sample of ads was randomly drawn in a way that was representative of the distribution of advertised vacant rental housing stock by New York City borough.[11] Only advertisements that invited housing seekers to reply by phone were pursued.[12]

Testers responded to sampled ads posing as individuals interested in renting the listed apartment. A project manager pre-screened all sampled ads against a master database of prior audits to ensure that there were no duplicate landlords or brokers in the sample of ads ever pursued by testers. We refer to all ads pursued by testers as the "audit sample." Each ad was pursued in a randomized order by a matched team of three testers of the same gender who varied by race: one white, one Black, and one Hispanic.[13] By extensively controlling for aspects of testers' assumed biographies within each trio, we employ a design-based approach to reduce the likelihood that observed racial discrimination levels are driven by statistical discrimination. Assigned biographies logically consistent with the rental price and size of the advertised unit were automatically generated at the time of ad sampling.

Upon reaching an individual when replying to an ad, testers were instructed to provide *limited*

[10] All LD ads were pursued (n=156) and of these, 44 were associated with landlords admitted into the experiment. To address concerns that their inclusion increases discrimination levels in the control group and decreases effect magnitudes if landlords in the LD subsample are more likely than non-LD landlords to discriminate, we find that the main results are not materially affected when the LD subsample is excluded. See the SI for full results.

[11] 35% Manhattan, 30% Brooklyn, 20% Queens, 10% Bronx, and 5% Staten Island. These are based on the distribution of ads by borough as identified in a pilot study.

[12] In a pilot study, we found that initial contact rates for emailed inquiries were extremely low.

[13] We matched testers' assumed credit score range, income level, household composition, occupation, perceived employment stability, gender, interpersonal skills, and perceived age.

information about themselves over the phone, including their assumed name,[14] interest in pursuing and renting the unit, their availability for an in-person viewing, and their financial qualifications to rent. Testers were instructed to schedule an in-person appointment on the same day at the earliest convenience of the individual with whom they were speaking. If testers were asked about other biographical attributes, they volunteered that information accordingly.

When all three testers in a matched team successfully secured an appointment to view the same advertised housing unit with the same landlord, the landlord associated with that unit was admitted into the experiment (i.e., the "experimental sample") and randomly assigned to a treatment condition. We refer to the landlord-housing unit (and thus the unit of randomization) as a "case." Testers then made their individually scheduled appointments, viewed the unit, interacted with the subject, and recorded their interactions afterwards. Testers were blind to treatment, but they were not blind to the city's interest in assessing discrimination as one of numerous characteristics of the rental market. As such, they were extensively instructed and coached not to fish for particular reactions, not to let their personal opinions about landlord behavior interfere with their ability to continue interactions during the audit, and not to raise suspicions of an audit.

Randomization Procedure and Definition of Treatments

Landlords admitted into the experiment were randomly assigned to one of three conditions: a control condition where no message was sent, a monitoring messaging condition, or punitive messaging condition. A block randomization procedure was used where blocks were defined by the

[14] Testers had an assumed first name and an assumed last name for the duration of the study. Both were randomly drawn from a database of names tagged with racial and ethnic identifiers (see the SI for details). While testers were screened at the time of hire such that their manner of speaking did not strongly signal a particular racial identity over the phone, testers' assumed names did contain a signal of their race (Butler and Homola 2017; Fryer and Levitt 2004).

ad's stratum[15] and by treatment regime.[16] Table A1 in the SI shows the distribution of cases across blocks and arms.[17]

Treatment scripts contain official language that the city uses to communicate aspects of fair housing law to the public.[18] In both the monitoring and punitive conditions, a city employee delivered the assigned appeal via a personalized, targeted phone message to the landlord. Calls were sent about two hours after testers successfully scheduled appointments over the phone and about two hours before the first scheduled appointment. In the monitoring condition, the treatment script verified that the subject was on the line, informed the subject that the call was from the New York City Commission on Human Rights *"as part of an ongoing informational campaign to remind landlords and brokers of their obligations under fair housing law,"* and provided the Commission's web address for more information. While the monitoring message does not provide specific information about what exactly is illegal and is in fact designed to avoid priming subjects to think about racial discrimination in particular, our prior is that most of the subjects already knew about fair housing law in the absence of treatment.[19] Thus the monitoring message can be interpreted an intensive intervention that makes salient injunctive norms surrounding fair housing law simply by invoking it. In the punitive condition, the treatment script was the same as the monitoring treatment script but includes, prior to providing the Commission's web address, the

[15] New York City borough or the LD oversample.

[16] Defined as a distinct design and randomization procedure. See Figure A2 in the SI for details.

[17] We verified the randomization was valid using randomization inference. The probability of obtaining a log-likelihood statistic (from a multinomial logistic regression of treatment assignment on 122 pre-treatment covariates) at least as large as the observed test statistic is p=0.97. Balance tables are presented in the SI.

[18] See the SI for the full text of the treatment scripts.

[19] This is because brokers, who comprise nearly 85% of experimental subjects, are required to demonstrate knowledge of fair housing law to be licensed in New York, and because Craigslist shows advisory language stating that discrimination is illegal to those posting rental ads.

following advisement emphasizing the punitive power of the state and the potential pecuniary costs of violating the law: *"It is illegal to discriminate against a person seeking housing due to their membership in a protected class. If you are found to have broken the law, you may be ordered to pay damages, provide reasonable accommodation, or incur civil penalties of up to $250,000."*

Data and Measurement

We combine the following data: (1) scraped data on advertised rental listings, (2) automated assumed biographical and treatment assignment data, (3) data on treatment delivery and compliance, and (4) detailed field reports testers recorded about their interactions with landlords *prior to the visit* (pre-treatment); *during a housing unit visit* (post-treatment); and *after the visit* (post-treatment).[20] We construct case-level measures of net discrimination, which is defined as the difference in favorable treatment between the majority and minority group tester in any matched pair. We apply this measurement strategy to construct two objective measures of discrimination, which we pre-specified as our primary post-treatment outcome measures: differential treatment by race in landlord or broker efforts (i) to call back and follow up with testers after the appointment and (ii) to offer the unit to the tester. These measures are computed for each majority-minority pairing (white-Black, white-Hispanic, and Black-Hispanic) and can take three values at the case level: -1 if only the minority tester is treated favorably, 0 if both the minority and majority testers are treated equally, and 1 if only the majority tester is treated favorably.[21] When examining average levels of the net measure, 1 means 100% net discrimination against the minority group, -1 means 100% net discrimination against the majority group, and 0 means that the two groups are treated equally.[22]

[20] Full details about data collection and measurement procedures are provided in the SI.

[21] If the landlord did not honor the appointment, the net discrimination measure is coded 0 (i.e., both in any pair were treated the same as neither received a callback or an offer).

[22] A concern is that estimates of discrimination levels are driven by the composition of testers in each racial group. To address this, we estimate a non-nested hierarchical model regressing a landlord-tester level indicator for favorable treatment on pre-treatment covariates and tester, tester race,

We also construct pre-treatment measures of discrimination that occurs over the phone between the time testers initially contact the landlord when inquiring about the ad and the time of random assignment. We measure early-stage discrimination for all cases testers pursue to describe baseline discrimination levels and to include as covariates in the experimental analysis.[23]

Sample Definitions

The study was in the field from April 13, 2012 to December 20, 2013. We focus on two key samples for the analysis: the audit sample, which contains 2,711 cases, and the experimental sample, which contains 653 cases. Figure A1 in the SI presents a flow diagram summarizing the sample filtering procedure and Figure A2 in the SI summarizes the cumulative number of cases admitted into the experimental sample over this period and by treatment regime. The characteristics and geographic distribution of the rental housing stock in our samples appear to be broadly similar to the rental housing stock in the New York City rental housing market during this period.[24]

Estimation and Inference

We use both design-based and model-based approaches to estimation and inference, with both approaches producing near-identical results.[25] For design-based (non-parametric) inference, we

and tester team gender random effects. We estimate this model for each of the objective outcome indicators among the control group and among the experimental sample. The estimated variance of the varying tester intercepts is negligible; thus we infer that the composition of testers in each racial group does not drive discrimination estimates. See the SI for details.

[23] Early stage discrimination indicators include differences in making initial contact; scheduling an appointment; how many biographical attributes the landlord asks about; the number and share of attributes discussed where testers receive either positive, negative, neutral, and skeptical feedback; and whether any negative or skeptical feedback is received.

[24] See the SI for summary statistics of the housing stock characteristics of listed units by sample.

[25] Additional analyses are described and reported in the Discussion and in the SI.

estimate sample intent-to-treat (ITT) effects[26] of government messaging on discrimination levels as the weighted average across blocks of differences in net discrimination levels between treatment groups within blocks, where weights are equal to the inverse of the probability of assignment to the condition to which the case was actually assigned. For this non-parametric analysis, standard errors are calculated using the conservative weighted Neyman estimator. In addition, we estimate effects using the following linear model:

$$Y_{ib} = \alpha_0 + \beta_1 T_{ib} + \gamma_b + u_{ib} \tag{1}$$

where *i* indexes landlords and *b* indexes experimental block, *Y* is the net discrimination outcome measure, and *T* is treatment assignment, variously defined when comparing mean outcomes between monitoring versus control, punitive versus control, or punitive versus monitoring. γ is a full set of block fixed effects and *u* is a disturbance term. We estimate Equation 1 on the subset of the data assigned to each pair of treatment arms being compared; β_1 is the effect of the treatment group relative to the comparison group. For this analysis we again use inverse propensity weights. We calculate *p*-values corresponding to a one-sided test of the null hypothesis of equality of means for the monitoring-control and punitive-control comparisons. We calculate *p*-values corresponding to a two-sided test of the null hypothesis of equality of means for the punitive-monitoring comparison and for all analyses involving net discrimination against Hispanic (vs. Black) testers because we have no strong priors about the expected direction of racial bias and treatment effects on racial bias when comparing the experiences of Black and Hispanic testers.

[26] We observe some noncompliance because some landlords assigned to a call hung up mid-call or because the city staffer delivering calls could not reach the landlord before tester appointments. Because the estimated share of Compliers is relatively high (between 71-81%), Complier Average Causal Effect (CACE) estimates are similar to ITT estimates and are shown in the SI.

Results

Figure 1 summarizes the main findings and shows, separately for each outcome measure: (1) levels of favorable treatment for different racial groups (top left quadrant); (2) differences in favorable treatment rates between groups (i.e., net discrimination levels) by treatment assignment (lower left); (3) differences in favorable treatment rates across treatment conditions for the same group (top right); and (4) the effects of treatment assignment on net discrimination levels relative to the control or monitoring comparison group (lower right).[27] We focus our discussion on the lower-left and lower-right quadrants in particular, which present estimates of baseline net discrimination levels and of ITT effects of messaging on net discrimination levels, respectively. We also draw attention to our most important findings in the figure by highlighting key estimates.

[FIGURE 1 HERE]

Baseline Discrimination Levels

First we assess baseline levels of discrimination in the outcome variables by examining the control group mean levels of net discrimination, which are defined as the control group mean difference in favorable treatment rates between groups, as shown the lower-left quadrants in Figure 1. Here we find statistically significant differences between Hispanic and white testers. We estimate that Hispanic testers were less likely than white testers to receive a callback from a landlord or broker—in 15.4% of cases compared to 21.5%, a difference of 6.1 percentage points ($p = 0.019$). They were also less likely to receive an offer for an apartment—in 6.1% of cases compared to 11.8%, a difference of 5.7 percentage points ($p = 0.011$). We find smaller, and statistically weaker, differences between Black and white testers. We estimate that Black testers were less likely than white testers to receive a callback from a landlord or broker—in 16.8% of cases compared to 21.5%, a difference of 4.7 percentage points ($p = 0.107$). Black testers were also less likely to receive an offer—in 9% of cases compared to 11.8%, a difference of nearly 2.9 percentage points ($p = 0.239$).

[27] Estimates corresponding to the information presented in Figure 1 are shown in the SI.

We can compare baseline levels of discrimination in callbacks to estimates of differential treatment in receiving a follow-up from the 2012 HUD audit of ethnic and racial discrimination in rental markets.[28] Our results run counter to the New York metropolitan area estimates from the 2012 HUD audit, which found that Hispanic testers received a follow-up from agents more frequently than white testers (in 6.9% and 5.4% of cases, respectively, a difference of -1.5 percentage points, $p = 0.804$) and that Black testers received a follow-up from agents more frequently than white testers (in 10.2% and 3.9% of cases, respectively, or a difference of -6.3 points, $p = 0.096$) (Turner et al. 2013, pp. 157-158). These differences in apparent baseline levels of discrimination might be due to two factors. First, the two estimates may be comparing discrimination baseline levels for two different populations in the New York area rental market. The HUD estimates for the New York City area correspond to the New York/Northeast New Jersey metropolitan area, and not to New York City specifically. Our estimates are specific to New York City, and in particular to landlords and agents who post ads on Craigslist and who schedule appointments with all three testers pursuing the listing. Second, there is less variability in our estimates than in the New York area estimates from the HUD audit since the sample size of our control group is nearly twice the size of the entire New York HUD audit. When comparing our estimates to national estimates of discrimination in follow-ups from the 2012 HUD study, the estimates are similar in that Black and Hispanic testers were less likely to receive follow-ups from agents than white testers: 10.5% for Black testers versus 11.0% for white testers, a difference of 0.6 percentage points that is not significant at the 0.05 level; and 7.9% for Hispanic testers versus 11.2% for white testers, a difference of 3.3 percentage points that is significant at the 0.1 level (Turner et al. 2013, pp. 44, 47).

Does Government Messaging Reduce Racial Discrimination?

Next, we report estimates of the effect of messaging on net discrimination levels, as shown in the lower-right quadrant of Figure 1. This section of the figure presents weighted nonparametric estimates (open markers) and regression estimates adjusted using block fixed effects and inverse

[28] No comparable measure of discrimination in offers exists in the HUD audit study.

probability weighting (filled markers; our preferred specification) with 95% confidence intervals.[29]

When compared to a pure control condition, sending a monitoring message decreases net discrimination against both Black and Hispanic testers (versus white testers) across the objective outcome measures, but mean effect estimates are substantively negligible and not statistically distinguishable from zero. Sending a monitoring signal decreases net discrimination against Black testers in receiving a callback ($-0.002; p = 0.486$) and in receiving an offer ($-0.003; p = 0.464$). For Hispanic testers, sending the monitoring message decreases net discrimination in receiving a callback ($-0.036; p = 0.201$) and receiving an offer ($-0.017; p = 0.31$).

The results on punitive messaging point generally to reductions in discrimination against Hispanics but increases in discrimination against Blacks. Sending a punitive message increases net discrimination against Black testers (versus white testers) across the outcome measures when compared to control, but none of these effects are statistically significant at the 5% level. Sending the punitive message reduces the likelihood of receiving a post-visit follow-up callback from the landlord (0.02; $p = 0.675$) and reduces the likelihood of receiving a post-visit offer (0.018; $p = 0.697$). Punitive messaging instead decreases net discrimination against Hispanic testers in receiving a post-visit callback (-0.066; $p = 0.056$) and in receiving a post-visit offer (-0.021; $p = 0.268$). The effect of punitive messaging on net discrimination in receiving a post-visit callback is just shy of significance at 5% ($p = 0.056$). Notably, sending a punitive message worsens outcomes for Black compared to Hispanic testers across the outcome measures. Punitive messaging decreases net discrimination against Hispanic testers relative to Black testers in receiving callbacks (-0.085; $p = 0.037$), and in receiving offers (-0.039; $p = 0.242$). Note that of these two estimated treatment effects, only the effect on callbacks is statistically significant at the 5% level ($p = 0.037$).

We next examine the comparative effectiveness of punitive versus monitoring messaging. It is clear from Figure 1 that there is no statistically distinguishable difference in the effectiveness of the two treatments. We find mixed results when comparing outcomes between the two treatment groups—Black testers do slightly worse when interacting with landlords in the punitive condition

[29] We focus on the latter in our exposition of results, because both yield qualitatively similar results.

as compared to the monitoring condition, while Hispanic testers do slightly better. All p-values for these objective outcomes range from 0.268 to 0.962.

Hidden Discrimination: Estimates using Subjective Measures

Our primary analysis focuses on the causal effects of government messaging on two objective measures of racial discrimination in the housing market. Focusing on objective measures is important in part because discrimination in the housing market can be relatively invisible.[30] We also have access to a set of subjective measures of discrimination that capture, among other aspects of the housing search process, subjects' perceptions of steering. As such, they reflect both the behavior of the landlords, which may be a function of treatment conditions, and the dispositions of confederates to notice and report indicators of discrimination, which is not subject to treatment conditions but may be associated with racial group membership. The measures, while rich, are also flawed in that they suffer from substantial missingness (on the order of 25-30%).[31] While missingness is not correlated with racial groups or treatment arms, it is still plausibly related to potential outcomes: testers may be less likely to report if they felt either little or substantial discrimination.

With this caveat in mind, in Figure A3 in the SI, we report messaging effects on an index measure of testers' subjective perceptions of favorable treatment and a net discrimination measure using the subjective index.[32] Although on the objective measures there is strong evidence for discrimination against Hispanics, this does not appear at all in the subjective measures. Moreover, while there is some evidence for discrimination against Blacks from the objective measures, for the subjective measure, Blacks report if anything better treatment than whites. Our results are con-

[30] http://citylimits.org/2015/09/08/housing-discrimination-doesnt-begin-or-end-with-poor-doors; http://fairhousingjustice.org/resources/film

[31] We refer the reader to the SI for an analysis of missingness on subjective measures.

[32] We use an index measure to avoid a multiple comparisons problem. We conduct and report this analysis to maintain fidelity to our pre-analysis plan and to situate our main results in a broader policy context. Details about the procedures related to this analysis are shown in the SI.

sistent with the dominant viewpoint in the field, but they also have implications for the generality of our findings. The difficulty of identifying discrimination based on subjective perceptions may be one factor that makes the punitive threat of law relatively weak, even if there are normative benefits. Different effects may arise in areas where identifying violations is easier.

Discussion

Research on the political economy of race in the United States has shown that racial discrimination plays a central role in perpetuating racial inequalities, but has long overlooked questions about whether government efforts to reduce discrimination have been effective. We therefore conduct the first experimental test of two common government strategies to pre-emptively induce behavioral compliance with the law in the domain of fair housing. Partnering with the City of New York, we assess whether government campaigns that invoke fair housing law and make salient the costs of violating the law reduce racial discrimination levels in the New York City rental market. The design offers a strong test because it (1) individually targets appeals at potential discriminators using live calls, (2) measures outcomes soon after the treatment is delivered, and (3) occurs in a policy context where government's monitoring and enforcement signals are credible.

First, in the absence of government intervention, we find strong evidence of discrimination in the New York City rental market on outcomes that are consequential for racial housing inequality. In the control group, Hispanics are less likely than Blacks, who are themselves less likely than whites, to receive favorable treatment from landlords. In terms of the objective outcome measures, we find strong evidence of discrimination in the New York City rental market, particularly against Hispanics. They are 6.1 percentage points less likely to receive a callback and 5.7 percentage points less likely to receive an offer for an apartment than whites. These findings stand in contrast to estimates of discrimination levels from a comparable audit study of the New York metropolitan area rental market in 2012 which reports no discrimination against Blacks or Latinos in receiving a callback. We also find suggestive, but not strong, evidence that punitive government messaging campaigns that both invoke the law and make salient the costs of violating the law can reduce net discrimination in receiving a callback from a landlord following an appointment, but only for Hispanics. We find no statistically distinguishable difference between the effects of punitive and monitoring messaging relative to control.[33] We assess the robustness of our main results (from pre-registered analyses) to alternative estimation strategies that, at least in principle, improve efficiency; our main results are not materially affected by the choice of estimator.[34]

Despite the lack of statistical significance, the magnitudes of the estimated messaging effects we observe in this study appear to be substantively significant both in relative terms and in terms of the mean level of discrimination under treatment. In relative terms, the effects are very large, albeit from a small baseline.[35] Punitive messaging reduces discrimination against Hispanics relative to whites by 109% and relative to Blacks by 126%. Not only is there a large relative decrease in net discrimination against Hispanics, the sign of the predicted group mean changes from positive to negative suggesting more favorable treatment for Hispanics relative to whites and Blacks under the punitive messaging condition. Alternatively we can adduce the substantive meaning of effects by interpreting the mean discrimination level under treatment. Our estimates of discrimination levels on key measures drop to zero under the punitive condition. This creates an inferential puzzle for policymakers as well as a need for further experimental replication, which we discuss below as well. In the remainder of the article, we address concerns regarding the internal and external validity of our findings, implications for policymakers, and directions for future research.

[33] Employing a conservative Bonferroni approach to multiple comparisons across treatments, outcomes, and groups, we fail to reject the null that these treatments make no difference.

[34] First, using all the data we estimate models using binary treatment indicators, block fixed effects, and inverse probability weights, thus allowing for possible efficiency gains from tighter block effect estimation. Second, we use a lasso regression as a principled method for covariate selection and then estimate covariate-adjusted ITT effects. See the SI for full results.

[35] Interpreting treatment effects as relative differences offers an alternative substantive interpretation and is customary in experimental analyses. See the SI for full results.

Addressing Internal Validity Concerns about Spillovers

The main threat to internal validity arises if the non-interference assumption is violated, which could occur if landlords assigned to a messaging condition communicate with other landlords (who are assigned to other arms) about the treatment message they received. We use two main strategies to show that such interference is unlikely given the sheer size of the New York City rental market and the sampling procedures.[36] First, we estimate that the probability a landlord or broker enters the audit sample is between 3.2% and 15%, and that the probability a landlord or broker enters the experiment sample is between 0.77% and 3.6%. Second, we characterize the experimental sample as a very small random sample of landlords and brokers in the New York City rental market. Because the actual number of active landlords and brokers in the New York City rental market is unknown, we estimate a conservative upper bound to proxy the denominator. We infer that the experiment sample must be far less than 2% of the estimated population of landlords and brokers in the New York City rental market. Taken together, these estimates suggest a low probability of between-subject interaction and make a strong prima facie case for minimal interference.

Addressing External Validity Concerns

Next, we address concerns related to external validity deriving from the ways study samples are constructed. Although care was taken to sample ads randomly from the universe of Craigslist rental listings, there are several scope conditions that limit the generalizability of our findings. First, our results are limited to ads placed via public listings on Craigslist, which is a subset of advertised market-rate rental units, and excludes discrimination involving non-public tenant search processes and discrimination in the subsidized rental market. This would underestimate discrimination levels under the assumption that landlords who post rental ads on Craigslist are likely to be less discrim-

[36] Methodological details and additional analyses are in the SI.

inatory than those who select into non-public modes of advertising.[37]

Second, our findings are also limited to landlords and brokers who require housing seekers to reply to rental ads by phone.[38] This is a scope condition on our findings and that future research should explicitly assess whether findings differ by the mode of initial contact. Third, a possibly relevant factor is that Craigslist employs its own anti-discrimination messaging, which is shown to users before posting and acknowledges federal, state, and city prohibitions on discrimination.[39] If this messaging deters would-be discriminators from advertising on Craigslist, then this would bias our estimates of baseline discrimination levels downward. If it simply changes the ways in which landlords and brokers advertise, then this makes the evidence for discrimination all the more striking as the discrimination we uncover cannot be readily detected from simple text analysis.[40] It is also possible that Craiglist's messaging, by having effects of its own on discrimination and reducing the baseline level of discrimination, reduces the magnitude of the effect of government messaging. If this is the case, then our results should be interpreted as evidence of the effects of government messaging *in a context in which non-governmental messaging also exists*.

Fourth, our experiment applied treatment only after appointments were scheduled to avoid differential attrition. This might have the effect of limiting the generalizability of our analysis to encounters in which early-stage discrimination was absent. We can address this concern by assessing discrimination toward testers occurring over the phone *prior* to randomization—that is, across the full audit sample. Table A2 in the SI summarizes these findings and shows an overall

[37] This is consistent with research documenting the role of segregated informal social networks in employment and housing search processes (Loury 2001; Pager and Shepherd 2008).

[38] We estimate that 55.8% of New York City rental listings on Craigslist require contact by phone. See the SI for details.

[39] The message states, in part: "It is illegal to discriminate in the sale, rental or leasing of housing because of a person's race, color, creed, national origin, sexual orientation, marital status, familial status, or religion."

[40] However, other work suggests that some discriminatory text survives (Oliveri 2010).

lack of difference in early-stage discrimination. We find no statistically significant differences in the likelihood of making any contact when first replying to an advertisement by phone. And differences in success in setting up appointments were very small: the only significant difference from three comparisons is between white and Black testers, with Black testers somewhat *more* successful at scheduling appointments (36.1% versus 34.8%; $p = 0.035$). These results provide greater confidence that our results are not driven by sample selection resulting from all three testers successfully scheduling appointments.

Finally, given the potential rarity of both encountering a set of auditors and receiving a targeted phone call from the government, there may be concerns that landlords are "spooked" by the intervention. Being spooked by the matched auditors does not threaten the internal validity of the experimental results because the design maintains measurement symmetry across all arms.[41] Being spooked by the targeted call from the government is not a concern because it is part of the net effect of the bundled treatment that we are interested in understanding.[42]

Policy Implications

What are the policy implications of our findings? The results on discrimination levels are relatively clear for policymakers: discrimination is a greater concern in New York than previously believed, and this may be true elsewhere also. The evidence is especially strong for Hispanics. The implications for interventions are less clear but just as important. The challenge here is that there is suggestive evidence of strong effects but many of the estimated effects do not hit conventional levels of significance despite the large scale of this experiment. Power issues loom large when the behavior of interest, discrimination, is relatively rare, and is itself estimated with uncertainty. In such cases relying on conventional standards of significance may be overly conservative—an issue

[41] This could mean our estimates of discrimination levels in the control group are downwardly biased and could be interpreted as conservative estimates.

[42] There may also be an interaction effect between encountering auditors and getting a targeted call, but we remain agnostic about that effect and recommend future research to understand it.

that has been of concern also in the study of rare diseases and has led to calls to alter standards in such cases (Gobburu and Pastoor 2016). Rather than altering standards, we calculate appropriate beliefs that the intervention is effective using a Bayesian framework by estimating the probability of a hypothesis (that each messaging strategy is effective) given the data. This is a fundamentally distinct quantity than the probability of the data given a hypothesis, which is the quantity of interest for determining statistical significance (Gill 1999). Understanding this alternative question is plausibly more relevant to policymakers who care about the probability that a given intervention works, rather than the probability of rejecting a specified null hypothesis. Importantly, we clarify that this is not a method to fish for statistically significant results, but is instead a method to answer a substantively distinct question that can be used following future replications (given a cumulation of experimental data) to improve how policymakers form posterior beliefs about the effectiveness of different interventions.

[FIGURE 2 HERE]

Figure 2 plots the posterior densities along with the estimated probabilities of effectiveness (see the SI for details about our methodology). In the bottom row, we see that for each distribution, approximately half or less of the probability mass lies below zero: The data are not highly informative for posterior beliefs that either monitoring or punitive messages reduce discrimination against Blacks. In the case of punitive messages on callbacks, roughly three-quarters of probability mass lies *above* zero, providing some indication that the treatment may have even increased discrimination against Black testers. The top row shows a different story for Hispanics. For callbacks especially, the bulk of probability mass—at least 89%—lies beneath zero, implying posterior beliefs that the messages reduced discrimination against that group. We can be most confident in posterior beliefs about the effectiveness of the punitive treatment in reducing discrimination against Hispanics in callbacks: More than 99% of the mass lies below zero, centered around the posterior mean reduction in discrimination of 6.6 points. For reference, that estimate is roughly twice the corresponding posterior mean reduction due to the monitoring treatment.

Directions for Future Research

Lastly, we outline several directions for future research. First, an important limitation of the present field experiment is that despite the unusually large scale of our study (in comparison to prior in-person rental market audits), there is not enough statistical power in the design to detect treatment effects on the order of about 6 points. This is in part because the control group mean discrimination level is relatively low and because estimating treatment effects involves, in effect, estimating an interaction between tester race and treatment. As this is the first field experiment of its kind, it was not possible to conduct ex ante power analyses using previously reported effect sizes as benchmarks.[43] Additional experimental replications with larger sample sizes are needed to precisely estimate discrimination levels and messaging effects.

Second, additional research is needed to understand whether differential effects for Hispanic and Black housing seekers exist.[44] Existing work on the politics of policy enforcement and compliance does not explicitly develop explanations for why attempts to induce compliance with anti-discrimination law might vary by the group membership of the target of discrimination. Future work should develop and test hypotheses about the conditions under which there exist variation in baseline discrimination levels and heterogeneous effects of governmental appeals by tester and landlord group membership.[45]

Third, the treatments we tested in this study were, given practical considerations and constraints, bundled. They also did not test the full range of potential strategies government could deploy in pre-emptive messaging campaigns to reduce discrimination. Future studies could more

[43] Instead, the study was designed with the reasonable ex ante expectation that it was adequately powered to detect and reduce racial discrimination levels reported in rental market audit studies published prior to the start of the study. See the SI for details on studies informing our priors.

[44] Given space constraints, we speculate on possible explanations for mixed findings in the SI.

[45] See the SI for an exploratory analysis of heterogeneous messaging effects by the perceived race of landlords using the data from this experiment.

finely operationalize treatments to test specific parameters in the decision-theoretic model that build on existing theoretical frameworks from literatures on prejudice reduction (Paluck and Green 2009) and on behavioral policy compliance (Weaver 2014, 2015). Future studies could also test whether effects vary by the mode of treatment delivery or by the messenger of the appeal.

Finally, replication in other housing markets and political jurisdictions would be valuable to understand how messaging effects vary across political and economic contexts. There may be different baseline discrimination levels and reasons why market actors discriminate that moderate their behavioral response to governmental appeals. There may also be different expectations among market actors about the credibility of government efforts to enforce anti-discrimination laws. How citizens perceive the capacity of the state to enforce compliance (Weaver 2014) and the degree of government forbearance in the domain of civil rights (Holland 2016)—which are partly a function of the incentives and preferences of bureaucrats and their principals (White, Nathan and Faller 2015; Einstein and Glick 2016; Mummolo 2017)—may affect how they perceive the credibility of government signals to monitor market behavior and enforce the law, the expected intensity of government efforts to do so, and ultimately the costs of violating the law.

Acknowledgments

We thank Commissioner Patricia Gatling, Clifford Mulqueen, Lee Hudson, and the staff at the New York City Commission on Human Rights for their support. We thank Joan Russell, Clare Wiseman, and Ruby Hlivko for their excellent work as project managers. We also thank Stephen Bonnett, Pat Egan, Ingrid Gould Ellen, Ryan Enos, Alan Gerber, Don Green, Greg Huber, Winston Lin, Joe Soss, and participants at the 2013 MPSA and APSA meetings, WZB Berlin, and the Georgetown McCourt School of Public Policy for constructive feedback. See the SI for full acknowledgments.

References

Aigner, Dennis J and Glen G Cain. 1977. “Statistical theories of discrimination in labor markets.” *Industrial and Labor relations review* pp. 175–187.

Alexander, Michelle. 2012. *The New Jim Crow: Mass Incarceration in the Age of Color-blindness*. New York: New Press.

Arrow, Kenneth J. 1973. The Theory of Discrimination. In *Discrimination in Labor Markets*, ed. Orley Ashenfelter and Albert Rees. Princeton: Princeton University Press pp. 3–33.

Becker, Gary S. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bertrand, Marianne, Dolly Chugh and Sendhil Mullainathan. 2005. "Implicit Discrimination." *American Economic Review* 95(2):94–98.

Bertrand, Marianne and Esther Duflo. 2017. Field Experiments on Discrimination. In *Handbook of Field Experiments*, ed. Esther Duflo and Abhijit Banerjee. Vol. 1 Amsterdam: Elsevier.

Blumenthal, Marsha, Charles Christian, Joel Slemrod and Matthew G. Smith. 2001. "Do Normative Appeals Affect Tax Compliance? Evidence from Controlled Experiment in Minnesota." *National Tax Journal* 54(1):125–138.

Bobo, Lawrence, James R. Kluegel and Ryan A. Smith. 1997. Laissez-Faire Racism: The Crystallization of a Kinder, Gentler, Antiblack Ideology. In *Racial Attitudes in the 1990s: Continuity and Change*, ed. Steven A. Tuch and Jack K. Martin. Westport, CT: Praeger pp. 15–42.

Bowles, Samuel. 2016. *The Moral Economy: Why Good Incentives are No Substitute for Good Citizens*. New Haven: Yale University Press.

Butler, Daniel M. and Jonathan Homola. 2017. "An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments." *Political Analysis* 25(1):122–130.

Cardenas, Juan Camilo, John Stranlund and Cleve Willis. 2000. "Local Environmental Control and Institutional Crowding-Out." *World Development* 28(10):1719–1733.

Cover, Robert. 1995. The Origins of Judicial Activism in the Protection of Minorities. In *Narrative, Violence, and the Law*, ed. Martha Minow, Michael Ryan and Austin Sarat. University of Michigan Press.

Dawson, Michael C. 2016. "Hidden in Plain Sight: A Note on Legitimation Crises and the Racial Order." *Critical Historical Studies* 3(1):143–161.

Dawson, Michael C. and Megan Ming Francis. 2015. "Black Politics and the Neoliberal Racial Order." *Public Culture* 28(1):23–62.

Dixit, Avinash K. 2006. *Lawlessness and Economics: Alternative Modes of Governance*. New York: Oxford University Press.

Donohue, John J. and Peter Siegelman. 1991. "The Changing Nature of Employment Discrimination Litigation." *Stanford Law Review* 43(983-1033).

Dovidio, John F., Agata Glutszek, Melissa-Sue John, Ruth Ditlmann and Paul Lagunes. 2010. "Understanding Bias toward Latinos: Discrimination, Dimensions of Difference, and Experience of Exclusion." *Journal of Social Issues* 66(1):59–78.

Dunning, Thad, Felipe Monestier, Rafeal Pinero, Fernando Rosenblatt and Guadalupe Tunon. 2015. "Positive vs. Negative Incentives for Compliance: Evaluating a Randomized Tax Holiday in Uruguay." Working Paper. Available at `http://dx.doi.org/10.2139/ssrn.2650105`.

Edelman, Benjamin, Michael Luca and Dan Svirsky. 2017. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." *American Economic Journal: Applied Economics* 9(2):1–22.

Einstein, Katherine Levine and David M. Glick. 2016. "Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing." *American Journal of Political Science* 61(1):100–116.

Epp, Charles. 1998. *The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective*. Chicago: University of Chicago Press.

Feagin, Joseph R. 1991. "The Continuing Significance of Race: Antiblack Discrimination in Public Places." *American Sociological Review* 56(1):101–116.

Forman, James Jr. 2012. "Racial Critiques of Mass Incarceration: Beyond the New Jim Crow." *New York University Law Review* 87(1):101–146.

Fryer, Roland G and Steven D Levitt. 2004. "The causes and consequences of distinctively black names." *The Quarterly Journal of Economics* 119(3):767–805.

Frymer, Paul. 2003. "Acting When Elected Officials Won't: Federal Courts and Civil Rights Enforcement in U.S. Labor Unions, 1935-1985." *American Political Science Review* 97(3):483–499.

Galanter, Marc. 1974. "Why the Haves Come Out Ahead: Speculations on the Limits of Legal Change." *Law and Society Review* 9(1):95–160.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–674.

Gneezy, Uri and Aldo Rustichini. 2000. "A Fine is a Price." *Journal of Legal Studies* 29(1):1–18.

Gobburu, J and D Pastoor. 2016. "Drugs Against Rare Diseases: Are The Regulatory Standards Higher?" *Clinical Pharmacology & Therapeutics* 100(4):322–323.

Gurian, Craig. 2005. "A Return to Eyes on the Prize: Litigating Under the Restored New York City Human Rights Law." *Fordham Urban Law Journal* 33(2).

Hadfield, Gillian K. and Barry R. Weingast. 2014. "Microfoundations of the Rule of Law." *Annual Review of Political Science* 17:21–42.

Haynes, Laura C., Donald P. Green, Rory Gallagher, Peter John and David J. Torgerson. 2013. "Collection of Delinquent Fines: An Adaptive Randomized Trial to Assess the Effectiveness of Alternative Text Messages." *Journal of Policy Analysis and Management* 32(4):718–730.

Heckman, James J. 1998. "Detecting Discrimination." *The Journal of Economic Perspectives* 12(2):101–116.

Heckman, James J and Peter Siegelman. 1993. The Urban Institute Audit Studies: Their Methods and Findings. In *Clear and convincing Evidence: Measurement of Discrimination in America*, ed. M Fix and R Struyk. Urban Institute Press.

Holland, Alisha C. 2016. “Forbearance.” *American Political Science Review* 110(2):232–246.

Iyer, Govind S., Philip M. J. Reckers and Debra L. Sanders. 2010. “Increasing Tax Compliance in Washington State: A Field Experiment.” *National Tax Journal* 63(1):7–32.

King, Desmond S. and Rogers M. Smith. 2005. “Racial Orders in American Political Development.” *American Political Science Review* 99(1):75–92.

Loury, Glenn C. 2001. Politics, race, and poverty research. In *Understanding Poverty*, ed. Sheldon H. Danziger and Robert H. Haveman. Cambridge: Harvard University Press.

Major, Brenda and Tessa L. Dover. 2016. Attributions to Discrimination: Antecedents and Consequences. In *Handbook of Prejudice, Stereotyping, and Discrimination*, ed. Todd D. Nelson. 2nd ed. New York: Taylor and Francis.

Massey, Douglas S and Nancy A Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Cambridge: Harvard University Press.

McAdams, Richard H and Janice Nadler. 2005. “Testing the Focal Point Theory of Legal Compliance: The Effect of Third-Party Expression in an Experimental Hawk/Dove Game.” *Journal of Empirical Legal Studies* 2(1):87–123.

McCubbins, Mathew D. and Thomas Schwartz. 1984. “Congressional oversight overlooked: Police patrols versus fire alarms.” *American Journal of Political Science* pp. 165–179.

Mendelberg, Tali. 2001. *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press.

Mummolo, Jonathan. 2017. “Modern Police Tactics, Police-Citizen Interactions and the Prospects for Reform.” Forthcoming in *The Journal of Politics*.

National Fair Housing Alliance. 2014. “Fair Housing Trends Report 2014. Expanding Opportunity: Systemic Approaches to Fair Housing.” http://nationalfairhousing.org.

Oliveri, Rigel Christine. 2010. “Discriminatory Housing Advertisements On-Line: Lessons from Craigslist.” *Indiana Law Review* 43:1125.

Pager, Devah and Diana Karafin. 2009. “Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making.” *Annals of the American Academy of Political and Social Science* 621:70–93.

Pager, Devah and Hana Shepherd. 2008. “The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets.” *Annual Review of Sociology* 34:181–209.

Paluck, Elizabeth Levy and Donald P Green. 2009. “Prejudice reduction: What works? A review and assessment of research and practice.” *Annual Review of Psychology* 60:339–367.

Pehrson, Samuel, Vivian L. Vignoles and Rupert Brown. 2009. “National Identification and Anti-Immigrant Prejudice: Individual and Contextual Effects on National Definitions.” *Social Psychology Quarterly* 72(1):24–38.

Phelps, Edmund S. 1972. “The Statistical Theory of Racism and Sexism.” *American Economic Review* 62(4):659–661.

Rosenberg, Gerald N. 1991. *The Hollow Hope: Can Courts Bring About Social Change?* Chicago: University of Chicago Press.

Skrentny, John D. 2002. *The Minority Rights Revolution*. Cambridge: Belknap Press of Harvard University Press.

Slemrod, Joel, Marsha Blumenthal and Charles Christian. 2001. “Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota.” *Journal of Public Economics* 79:455–483.

Turner, Margery A., Claudia Aranda, Diane K. Levy, Rob Pitingolo, Rob Santos and Doug Wissoker. 2013. "Housing Discrimination Against Racial And Ethnic Minorities 2012." Washington, D.C.: U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

Tyler, Tom R. 2004. "Enhancing Police Legitimacy." *Annals of the American Academy of Political and Social Science* 693:84–99.

Tyler, Tom R. 2006. *Why people obey the law*. Princeton University Press.

Weaver, R. Kent. 2014. "Compliance regimes and barriers to behavioral change." *Governance* 27(2):243–265.

Weaver, R. Kent. 2015. "Getting People to Behave: Research Lessons for Policy Makers." *Public Administration Review* 75(6):806–816.

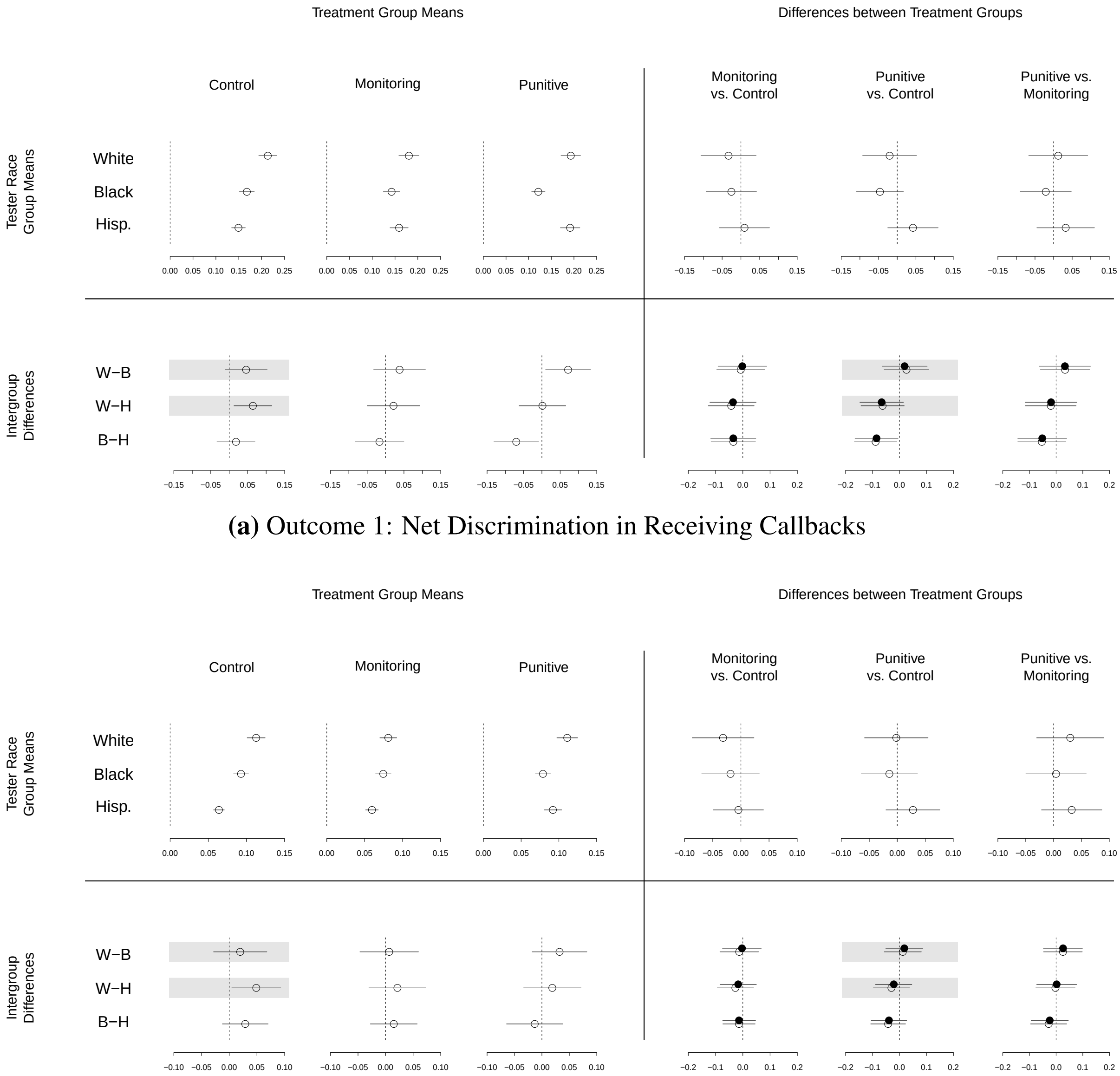
White, Ariel R., Noah L. Nathan and Julie K. Faller. 2015. "What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials." *American Political Science Review* 109(1):129–142.

Wilson, James Q. 1980. *The Politics of Regulation*. New York: Basic Books.

Albert H. Fang is Postdoctoral Associate in the Institution for Social and Policy Studies at Yale University, New Haven, CT 06520.

Andrew M. Guess is Assistant Professor of Politics and Public Affairs at Princeton University, Princeton, NJ 08544.

Macartan Humphreys is Professor of Political Science at Columbia University, New York, NY 10027.



(a) Outcome 1: Net Discrimination in Receiving Callbacks

(b) Outcome 2: Net Discrimination in Receiving Offers

Figure 1: Main results on discrimination levels and treatment effects. Each panel shows, for each of the two objective outcome measures, levels of favorable treatment for different racial groups (top left quadrant), differences in favorable treatment rates between groups (i.e., net discrimination levels) by treatment assignment (lower left quadrant), differences in favorable treatment rates across treatment conditions for the same group (top right quadrant), and the effects of treatment assignment on net discrimination levels relative to the control or monitoring comparison group (lower right quadrant) with weighted nonparametric estimates shown using open markers and regression estimates adjusted using block fixed effects and inverse probability weighting shown using filled markers. Lines indicate 95% confidence intervals. Our main quantities of interest are highlighted in light gray.

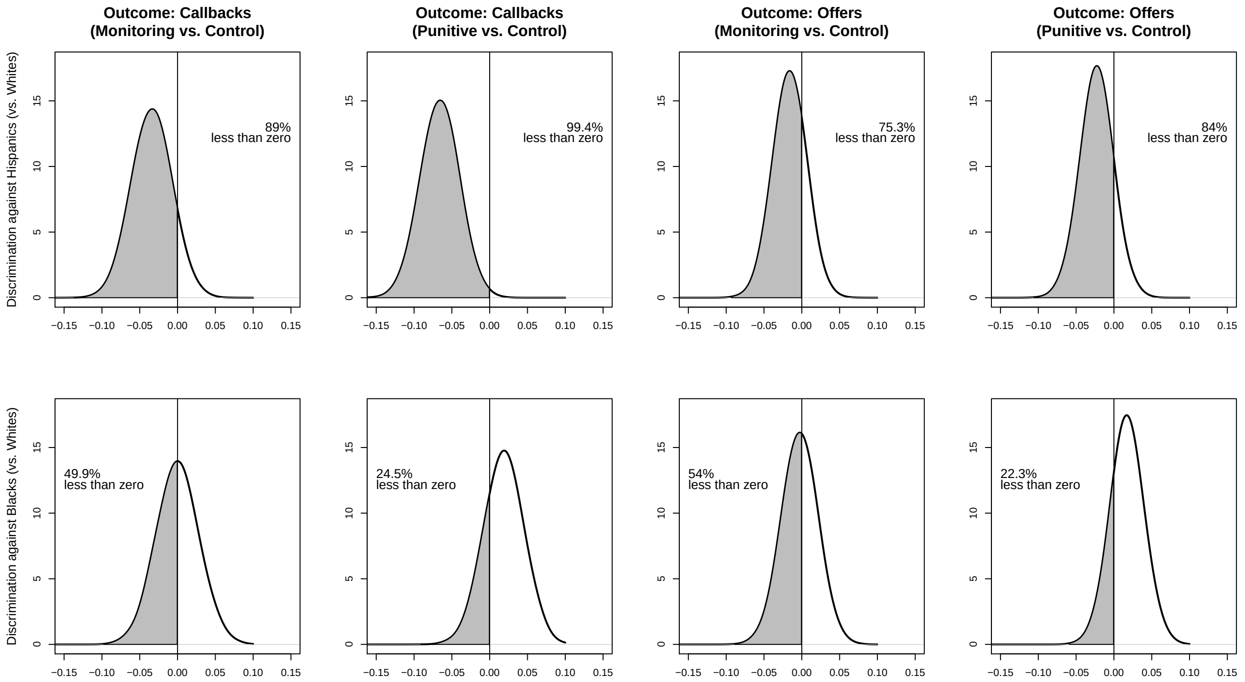


Figure 2: Posterior densities of treatment effects on discrimination against Hispanics (top) and African Americans (bottom) relative to whites. Columns correspond to combinations of outcome measure (callbacks, offers) and treatment message (punitive, monitoring) relative to control. The area under the curve below zero is shaded in each plot, and the mean of each posterior density is shown with a dotted line. Each posterior is estimated via 10,000 Monte Carlo draws from the marginal distribution of β_1 conditional on σ^2 and y given improper uniform priors, $\beta_1|\sigma^2, y \sim N(\hat{\beta}_1, V_{\beta_1}\sigma^2)$.