SUPPLEMENTAL APPENDIX FOR:

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

FOR ONLINE PUBLICATION ONLY

A SUPPLEMENTARY TABLES AND FIGURES REFERENCED IN THE MAIN ARTICLE TEXT

	Mean Level		Difference		
Measure	Majority	Minority	(Maj.-Min.)	p-value	[N]
A. White vs. Black Testers					
Any contact	0.512	0.524	-0.012	(0.184)	[2711]
Scheduling appointment	0.348	0.361	-0.013	(0.035)	[2711]
B. White vs. Hispanic Testers					
Any contact	0.512	0.512	0	(0.968)	[2711]
Scheduling appointment	0.348	0.354	-0.006	(0.283)	[2711]
C. Black vs. Hispanic Testers					
Any contact	0.524	0.512	0.012	(0.189)	[2711]
Scheduling appointment	0.361	0.354	0.007	(0.284)	[2711]

Table A1: Incidence of Early Stage Discrimination. Contact success rates and scheduling rates for white, Black, and Hispanic testers. We use OLS to estimate whether tester race predicts whether a tester makes any contact or successfully schedules an appointment, and then conduct F tests of the null hypothesis that the coefficients on tester race dummy variables equal zero. We find that tester race does not predict whether the tester makes any contact with the landlord ($F = 0.5203, p = 0.59$) and does not predict whether the tester successfully schedules an appointment ($F = 0.4936, p = 0.61$).

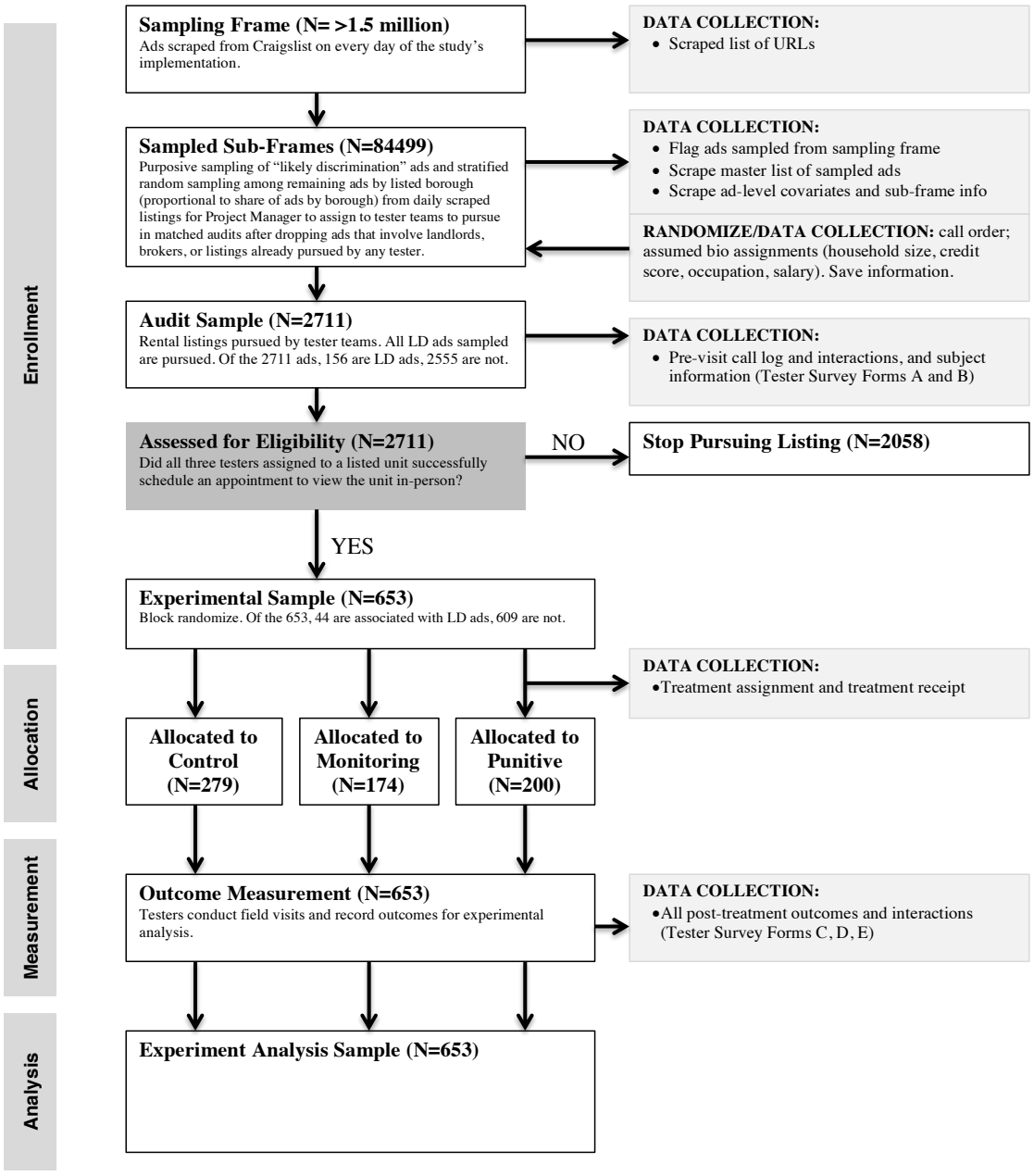


Figure A1: Flow Diagram of the Process Defining the Experimental Analysis Sample.

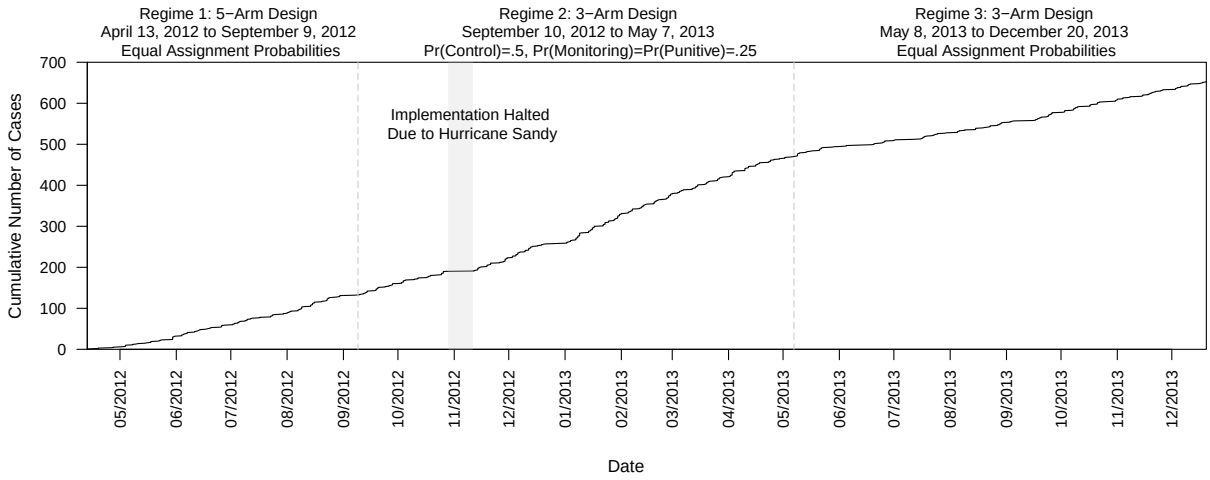


Figure A2: Cumulative Number of Cases Over Implementation Period

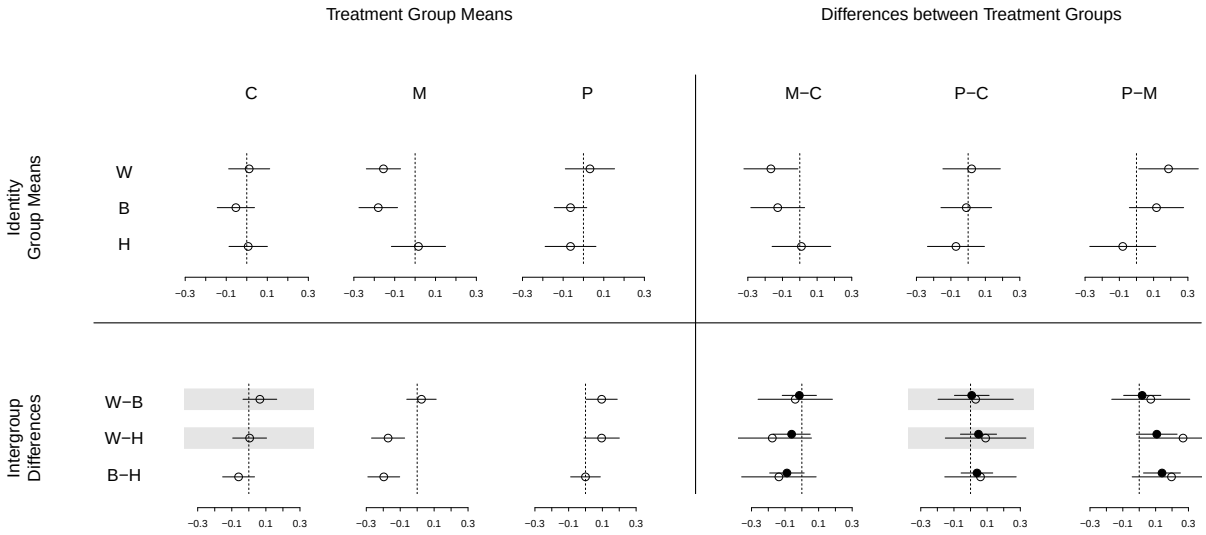


Figure A3: Levels of favorable treatment measured with the subjective index, for different racial groups (top left quadrant), differences in favorable treatment rates between groups (i.e., net discrimination levels) by treatment assignment (lower left quadrant), differences in favorable treatment rates 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.

B SUPPLEMENTARY INFORMATION ON STUDY IMPLEMENTATION AND FIELD PROCEDURES

B.1 Sampling

B.1.1 Additional Details about the Sampling Procedure

Vacant rental housing advertisements (which we call "cases") are sampled using a stratified random sampling procedure with proportional allocation by New York City borough from a widely used online classified listings website, Craigslist.[1] We restrict borough-specific listings to the "All Apartments" category available when browsing by borough-specific sub-sections of Craigslist. Stratified sampling by borough is important to increase homogeneity in potential outcomes across subjects to increase statistical power. In addition we construct a sampling frame of vacant rental housing ads that we designate as ads that contain language that suggest possible discrimination; this is done to create an oversample of "likely discrimination" ads to increase statistical power. Since explicitly discriminatory ads are flagged and removed, searches for words and phrases that are explicitly discriminatory generally yield no hits. We therefore limit our search to identify ads containing words and phrases that implicitly suggest markers of racial prejudices signaled through class preferences, given the strong relationship between race and class in the United States.[2] These ads are scraped from Craigslist and a random draw is sampled from this set prior to borough-stratified random sampling. These "likely discrimination" ads are excluded from the borough-stratified sampling frames so that they are not double-sampled.

Sampling occurs each day the study is implemented, which is limited to weekdays when the City is open for business. The sampling frame for each draw on a given day is the set of advertisements listed on Craigslist during that day up until the time of the draw, and any advertisements listed on previous business days during which the study was not conducted.[3] Sampling daily or near-daily ensures that vacant rental housing advertisements pursued by testers are recent ads that a real person looking for rental housing would likely pursue. Only ads containing landlord or broker telephone numbers are pursued by testers; the rest are discarded.[4] For sampled ads, copies of the original ads as they appear on Craigslist are saved.

B.1.2 "Scraping" Public Listings to Sample Available Units in the Housing Market

Both audit studies and field experiments in the housing market require sampling methods that are replicable and easy to implement. We were able to take advantage of two features of the study context: CCHR's interest in publicly listed units and the fact that the vast majority of such listings can be found online. In particular, by regularly sampling Craigslist, we were able to assemble a representative set of listings covering the range of units available in the New York City rental housing market.

[1] The proportions are 35% Manhattan, 30% Brooklyn, 20% Queens, 10% Bronx, and 5% Staten Island; these shares reflect the rough distribution of ads by borough on Craigslist as identified in the pilot study.

[2] Search terms used are: "hip," "up and coming," "yuppie," and "qualified."

[3] The study is implemented five business days every week; project staff do not work on City holidays.

[4] From the pilot study we learned that response rates, the probability of reaching a landlord or broker, and the probability of scheduling an appointment to view a rental unit are significantly higher among ads containing phone numbers in comparison to ads that request replies by email.

Before turning to best practices for future research and enforcement efforts, a short technical note. The Python programming environment proved to be well suited to the task of "scraping" Craigslist on a daily or almost-daily basis, sampling from the universe of listings, and saving the appropriate information to a secure location. Python can be installed on Windows-based PCs and comes included with Mac OS X, and our project managers were able to run the scripts for the most part without trouble. In particular, we highly recommend the BeautifulSoup screen-scraping library for its flexibility and straightforward implementation. Craigslist is famous for its plain, no-frills layout, and this was a major advantage when developing the script. However, there were several occasions when the layout subtly changed without warning, which caused numerous errors and hasty revisions. We would recommend building in robust error-catching routines in addition to notification systems (i.e. automated emails sent to the primary administrator) in order to minimize the risk of this kind of change. Having a backlog of available cases (perhaps one or two days old) also helped when there were technical issues impeding the usual sampling procedure.

Finally, while password-protected cloud-based storage services such as Dropbox are vital for data management in studies of this kind, we found that a system based on writing a large number of small files in embedded directory structures can greatly slow down the syncing process. One solution is to regularly move data files from completed cases to a secure location separate from the active operation of the scraping and sampling mechanism.

B.1.3 Best Practices for Future Approaches Using Online Listings

- *Random sampling.* Since the entire universe of relevant listings can be scraped (for example within a given time period, borough, or neighborhood), discretion at this stage can be eliminated. While hundreds of thousands of listings will be posted to Craigslist on a typical day in New York City, random samples can provide representative snapshots that are more manageable for a given purpose. However, it is important to note that scraping before the day is over may be necessary for studies or investigations requiring engagement with active listings. This introduces the possibility of bias due to the types of listings that may be posted at given times of the day. The issue can be minimized by scraping sufficiently backwards in time.

- *Search terms.* Researchers or enforcement officials may be interested in pursuing suspicious listings by using search terms rather than an open-ended scrape. We attempted to incorporate a version of this procedure into an earlier version of the study but found no systematic differences in the sample. Since Craigslist actively pulls listings containing certain words, explicitly discriminatory language may be difficult to find. Moreover, the possibility of false positives using this kind of directed search is real.

- *Handling duplicates.* A major difficulty with Craigslist (at least in the New York City rental housing market) is that some brokers post bulk listings for duplicate or even nonexistent units. Our solution was to keep a running list of phone numbers and broker names from completed cases which the scraping program used to automatically remove listings from the sample. However, even this procedure was far from perfect as names and even numbers seemed to change frequently. Project managers had to devote a significant amount of time to handling this problem.

B.2 Randomization Details

The randomization procedure varied over the course of the experiment. There were three randomization regimes. In the first regime, which lasted from the start of the study in April 2012 to September 9, 2012, a five-group design was employed where subjects could be randomly assigned with equal probability to one of five conditions: to one of the three experimental conditions of interest (control, monitoring, or punitive) or to one of two other treatment groups (receiving a value-based normative appeal or receiving both normative and punitive appeals).[5] In the second regime, which lasted from September 10, 2012, to May 7, 2013, the normative and combined normative-plus-punitive treatment arms were eliminated, resulting in a three-arm design where the probability of being assigned to control was 0.5 and the probability of being assigned to either the monitoring and punitive conditions was 0.25. In the third regime, which lasted from May 8 to December 20, 2013, the three-arm design was continued but equal assignment probabilities were used.

Thus, there are a total of 17 blocks across the three randomization regimes. Table A3 in Appendix C.1 summarizes the distribution of cases across blocks and treatment assignments.

The process of matching a subject to an experimental block via the listed advertisement is automated to minimize the possibility of human error. At the time of case scraping and sampling, the block membership of each case is automatically determined and saved to a master database. At the time of randomization, the Project Manager simply enters a unique case identifier into a Python-based user interface, which cross-references the block membership of that case and selects the next treatment assignment from stored treatment assignment vectors we generated by block given the design.

For each case that enters the study sample, information about the landlord or broker with whom the testers would meet is collected from testers' initial set of calls. This information is compiled by the Project Manager who then forwards it to a designated Treatment Administrator. This individual is a dedicated staff member working for the Commission on Human Rights. The Treatment Administrator delivers the assigned treatment message by phone before the first scheduled appointment time. The staffer at the Commission in charge of administering treatment records subject non-compliance with treatment message components conditional on assignment.

B.3 Treatment Scripts: Full Text

The treatment messages (with punitive components in bold) followed this script:

> Hello, could I speak with `[First Name of Landlord/Broker]` please?
> *If prompted for identifying information:*
> I'm calling with a message from the New York City Commission on Human Rights. This will take less than a minute.
>
> *Once the targeted recipient is on the line:*
> Good `[morning/afternoon/evening]`. I'm calling 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.

[5] These two additional arms (normative and normative-plus-punitive) were dropped due to project cost constraints and concerns about statistical power. We omit these arms from the analyses presented in this article as they are not of primary interest.

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.

Please take a moment to visit `nyc.gov/cchr` to learn how fair housing law protects individuals from discrimination. Thank you very much for your time.

B.4 Manipulating Markers of Racial Identity

There have been several debates on how to clearly signal racial identity in field experimental studies about discrimination. Audit studies studying discrimination employ matched pair (or triple) audit designs where the trait or marker of auditors' group membership, which is used by the landlord or broker to assess the auditor and affects discrimination, is manipulated. All other characteristics of the testers that affect potential outcomes are fixed. In this section we review how we contribute to three debates on how to manipulate markers of racial identity in field experimental research.

B.4.1 The Racial Soundingness of Names

Most field experimental work examining the effects of race on disparate treatment in employment and housing has manipulated the racial soundingness of names (most notably Bertrand and Mullainathan (2004)) since the racial identities signaled by testers' assumed names are an important signal of race (Fryer and Levitt 2004). Much of the use of names to signal racial identity has employed researcher discretion in choosing names to maximize the size of the expected effect of the racial signal on discrimination.

While this does not pose an internal validity problem, this is problematic with respect to external validity. When researchers employ discretion in choosing names to maximize the expected effect size of names on discrimination, the estimand of interest is a quantity that is not generalizable to the population since the distribution of names does not match the distribution of names in the population.

To address this issue, we turned to a rare publicly available data set of real names tagged with racial and gender information.[6] The data set consists of the names of children between the third and tenth grades tested in the Colorado state assessment system from 2007 to 2010. Approximately 400,000 students are tested each year. We randomly sampled (with replacement) from this database, separately drawing four first and last names for each of six race-gender groups (white male, white female, black male, black female, Hispanic male, Hispanic female). We then paired together sampled first and last names within each group. We then had a list of representative names from the given population such that more common names were more likely to be drawn.

The Colorado student database population is distinct from the population of interest in this study, but we argue that it may be employed under the assumption that, since New York City attracts many people from across the country, it is reasonable to assume that names associated with particular regions of the United States will be encountered in New York.

[6]The database of names is contained in a package for the R statistical programming language, **randomNames**, written by Damian W. Betebenner (2012).

In general, there was a final concern that a name signaling a particular racial identity also signals a particular ethnic identity. It would compromise the study if a tester was assigned a racial-sounding name that is incongruent with the tester's *actual* ethnic background, which landlords and brokers may be able to detect from the tester's physical attributes. Thus, if an obvious incongruence was detected, the name was discarded and another name from the list was chosen and assigned to the tester.

B.4.2 Linguistic and Class-Correlated Signals of Race

As Pager and Shepherd (2008) note in their review of the housing discrimination literature, "research using telephone audits further points to a gender and class dimension of racial discrimination in which black women and/or blacks who speak in a manner associated with a lower-class upbringing suffer greater discrimination than black men and/or those signaling a middle-class upbringing (Massey and Lundy 2001; Purnell et al. 1999)" (189).

To maximize the probability that minority testers are able to make an appointment to view a housing unit and to control for between-tester variation in class signaled through race and linguistic patterns, we account for linguistic and verbal markers of racial background they demonstrate during the phone conversation they have when replying to advertisements. All testers hired for the study are able to speak in a manner associated with a middle-class upbringing so as to not prime extreme class associations that drive racial perceptions.

B.5 Procedures for Screening and Hiring Testers to Pose as Interested Housing Seekers

Matched teams of three testers – one white, one black, and one Hispanic – are assigned vacant rental housing ads sampled from Craigslist to pursue. The effective composition and matching of testers to conduct in-person audits is therefore a major concern. The project originally aimed to compose a final team of 24 testers (or 24 FTE equivalents) with equal shares of testers for each race by sex combination. The city implements the following procedure to ensure the quality of testers employed in the study.[7]

Successful applicants are subject to two lengthy interviews. In the first round interview, conducted via a video chat client (e.g. Skype or Google Video Chat), applicants are required to articulate their interest in the study to assess overall fit; articulate concrete work experiences that demonstrate experience working in groups and working individually on detailed tasks; and demonstrate familiarity with multiple neighborhoods across New York City's five boroughs.

The second round interviews are conducted in-person. Applicants are required to participate in four simulated landlord/broker-tester interactions in which they take on the role of both the landlord/broker and tester given real ads pulled from Craigslist. Those playing the part of the tester are given an assumed biography and are evaluated on their ability to convincingly act out the part of an interested renter with that biography.

Testers are also asked complex questions for which they know little but that they are likely to encounter in the field, including: requests to elaborate reasons for moving to a particular neighborhood given one's current neighborhood of residence; elaborations on what one does at work; follow-up questions commonly asked by landlords and brokers about whether one is being truthful

[7] Methods used by Pager et al. (2009) serve as a benchmark.

about one's income and source of income; detailed questions about "what's going on" in one's assumed neighborhood of residence for which a tester may actually know little to nothing. This test is done to see how adeptly applicants can ad lib without falling "out of character" or compromising the audit.

Finally, applicants are required to recall interactions from a simulated landlord/broker-tester interaction and quickly produce a set of detailed field notes in 10 minutes. This exercise is used to evaluate testers' ability and capacity to conduct participant-observation research and record detailed observations about verbal interactions, non-verbal behavior, and contextual information about social interactions. Lastly the city assesses applicants' attention to detail and their ability to use online data entry interfaces by observing how successfully they follow nuanced application instructions and interview scheduling instructions.

Each tester hired for the study is required to complete a standard training and a training period. Ongoing spot checks for quality control by the Project Manager and quality control checks of the data collected were regularly conducted.

C DATA AND MEASUREMENT APPENDIX

C.1 Documenting Landlord-Tester Interactions

Testers document the following information about their interactions with landlords and brokers. In the pre-visit stage of the housing search process, testers documented information about how difficult it was to successfully schedule an appointment to view the unit of interest, including: whether they were able to schedule an appointment, the number of call attempts made before scheduling an appointment, the time when the appointment was made, the appointment date and time, who they interacted with, and if an appointment could not be made, the reasons why. Testers also documented the aspects of their assumed biography that came up during pre-visit interactions over the phone and how landlords and brokers with whom they interacted reacted to the information provided.[8]

During the appointment stage, testers collect detailed information about all the primary individuals with whom they interacted during the visit[9] and information about the units they were shown.[10] In addition to accounting for the people and housing units they encountered during

[8]The survey instrument prompts testers to indicate whether the following aspects of their biography arose in conversation: name, personal income, household income, occupation, employer, credit score, marital/partner status, children/dependents, reason for moving, location (neighborhood) of current residence, location (neighborhood) of workplace, gender, educational background/pedigree, race, ethnicity or national origin, sexual orientation, linguistic or speech-related traits, age, phone number, and employment stability or source of income. testers may also report additional attributes about their assumed biographies that are questioned.

[9]Testers recorded each individual's name, firm affiliation, job description, and whether each individual was the same person with whom they spoke to set up the appointment. Testers also recorded their perceptions of each individual's age range, race, and ethnicity. To verify this information, testers were also instructed to ask for and collect business cards.

[10]Relevant fields include whether a particular unit shown is the sampled unit; the unit's street address, borough, and neighborhood as described by the landlord or agent; the monthly rental price (quoted in person); the number of bedrooms and bathrooms; whether the building has a doorman and an elevator; whether the unit or building includes a washer/dryer; whether the landlord or broker claimed the unit would be renovated before move-in; the amenities included in the rent; the length of the lease; the security deposit required; any additional fees required to secure the apartment and their respective amounts; and whether an application is required (if yes, a copy of the application is

their visit, testers also provided open-ended responses about their interactions with landlords, brokers, and agents. Testers recorded the general demeanor (including but not limited to their body language; professionalism; and instances of expressed interest, lack of interest, skepticism, attentiveness, repulsion) of the landlords and brokers of interest toward them at the beginning, middle, and end of the visit. Testers recorded the sales efforts landlords and brokers make during the visit, which include rental incentives and extra amenities offered such as waived fees, discounted rent, discounts on local goods and services, gifts, or other "perks" meant to persuade testers to sign a contract soon; attempts to editorialize about the neighborhood, its residents, amenities, and/or character; attempts to editorialize about the building, its residents, amenities, and/or character; offers to follow up after the appointment; and attempts to editorialize about the housing search process or the housing market in general.

Testers also documented the other-regarding beliefs and group perceptions revealed by landlords and brokers during the appointment. Testers recorded whether landlords or brokers suggested, either explicitly or implicitly, that the presence of persons of any particular group in the area may result in an increase or decrease of property values, directly or indirectly; that the presence of persons of any particular group in the area may result in an increase or decrease of criminal or anti-social behavior in the neighborhood/area, either directly or indirectly; if landlords or brokers expressed judgment toward the tester based on their revealed perceptions of the tester; if landlords questioned their qualifications to rent; if landlords or brokers revealed prejudices or beliefs in stereotypes about any economic or social group, including the group to which the tester belongs, during the visit. Testers recorded their reactions in these interactions as well.

Post-visit stage interactions documented by testers include the callback date and time, if any; if the tester was offered the unit and if not, whether the unit was already rented out; whether the landlord or broker offered to show the tester other vacant rental units; and other interactions that occurred during post-visit correspondence.

Multiple data collection methods – specifically closed and open-ended survey questions and qualitative participant-observation field notes – allow us to unambiguously measure discrimination given ancillary information about the context of social interactions. This also allows to construct more stable composite measures of discrimination by utilizing information from both qualitative and quantitative data records of tester interactions with landlords and brokers. The added information also allows us to capture more nuanced forms of discrimination that may be implicitly indicative of discrimination against minorities, such as steering. Posing open-ended questions to testers provides an inductive mode of data collection where we are receptive to the many possible forms of discrimination that occur in the rental housing market today that may not be easily defined *a priori.*

Specifically, using detailed qualitative field notes provided by testers on their interactions with landlords during the appointment, we are able to collect numerous subjective measures of differential treatment occurring when testers interact with landlords in person during the appointment (and after randomization). As there are multiple types of qualitative interactions that occur, to avoid a multiple comparisons problem we construct an index measure of testers' subjective perceptions of favorable treatment and a net discrimination measure using the subjective index. Complete details on the procedures used to classify and code testers' qualitative field notes and to construct the subjective index measure are discussed below. Unfortunately, this index suffers from missing

requested). Testers also reported their subjective assessments of the unit interior and the building exterior.

data—most obvious for cases in which a landlord failed to show up for an appointment—which is not plausibly unrelated to potential outcomes. For this reason we exclude this measure from the main analysis but report results nevertheless to maintain fidelity to our pre-analysis plan and to situate our main results in a broader policy context.

C.2 Net versus Gross Measures of Discrimination

Table A2 illustrates the construction of the net measure and contrasts it with a common, alternative gross measure of discrimination. Suppose there are four cases for which we measure favorable treatment indicators for a majority and minority tester pair. For Case 1, both testers are treated favorably. For Case 2, the majority tester is treated favorably, but the minority tester is not. For Case 3, only the minority tester is treated favorably. For Case 4, neither tester is treated favorably. Across all four cases, we capture all possible combinations of treatment toward testers in the pair. These combinations are described in columns (A) and (B) in Table A2, where favorable treatment toward a tester is coded as "1", and unfavorable treatment toward a tester is coded as "0".

	(A)	(B)		
	Majority Tester	Minority Tester	Measure of Discrimination	
	Treated Favorably?	Treated Favorably?	Net	Gross
Case ID	(1=Yes, 0=No)	(1=Yes, 0=No)	(Equals A-B)	(1 only if A=1 & B=0)
1	Yes (1)	Yes (1)	0	0
2	Yes (1)	No (0)	1	1
3	No (0)	Yes (1)	-1	0
4	No (0)	No (0)	0	0
Average level of discrimination			0%	25%

Source: Authors' representation of net and gross measures of discrimination.

Table A2: Comparing Net versus Gross Measures of Discrimination

The net measure is constructed by subtracting column (B) from column (A). The gross measure is coded as a 1 only when column (A) equals 1 and column (B) equals 0. The main difference then is in the coding of Case 3. The net measure captures the average level of discrimination such that on the margins, a case where a minority tester is treated favorably but the majority tester is not treated favorably effectively "counteracts" a case where the majority tester is treated favorably but the minority tester is not treated favorably. The gross measure only counts up the share of cases where the majority tester is treated favorably but the minority tester unfavorably. The implication is most evident when describing the average level of discrimination across cases. In this example, the average level of net discrimination is 0%, whereas the average level of gross discrimination is 25%.

Both measures may contain bias in the measurement of differential treatment, but the net measure is preferable so long as it is interpreted as a lower bound on the level of discrimination that exists. This is because "the net measure is constructed under the assumption that adverse treatment against the white tester occurs only because the testers' visits differed, and so adverse treatment against the white tester provides an accurate measure of the number of instances of minority adverse treatment that arose because the testers' visits differed" (Ross 2002, 55).

C.3 Constructing Subjective Measures of Net Discrimination

Our measurement strategy also seeks to understand differences in subjective experiences in the housing search process. This focus is informed by well-known macro-level shifts in Americans' racial attitudes, as well as the impact of these shifts on behavior. First, the vast majority of Americans oppose discrimination against minorities in private transactions such as buying and selling a house, a finding that mirrors other gradual yet decisive changes in public opinion surrounding race relations in the United States. Second, social scientists have documented the emergence and persistence of "new racism," or discriminatory attitudes that manifest themselves not in overt behavior or socially desirable survey responses but in subtler attitudes about minorities' competence, abilities, or cultural tendencies.

These findings, suggesting some divergence between public norms and private behavior, have motivated continued research and debate on how best to measure individuals' so-called "implicit attitudes." Our contribution is to demonstrate one method of addressing this challenge in the context of a field experiment: marshaling testers' open-ended survey responses to shed light on subtler interactions that might not be captured via traditional quantitative outcome measures. One potential approach would be to hand-code testers' subjective observations in the field, a lengthy process that might uncover fresh insight but at the cost of introducing inconsistencies between coders. Another option would be to utilize unsupervised learning methods to reveal latent meaning, but this approach could overlook or misinterpret useful contextual information such as responses to specific neighborhoods. We opt for the middle path: using supervised learning algorithms trained with a set of hand-coded survey responses. This allows us to maintain focus on essential features of the text while ensuring uniformity during the classification stage.

Our procedure was as follows. First, we randomly selected 300 open-ended tester survey responses (roughly 15% of the total) to be manually coded. Based on a protocol we inductively developed, a pair of research assistants who were blind to treatment independently evaluated each case along 15 potential dimensions. The assistants then worked together to agree on a final set of codes where their individual assessments diverged. Once this adjudication process was complete, we took the subset from our list of 15 available codes with the highest intercoder reliability in the first step.[11] The resulting five codes are: sales efforts by landlords/brokers; praise about a tester's qualifications; positive response to a tester's background; positive editorializing about an apartment or neighborhood; and professionalism of landlords/brokers.

These codes cover a range of potential responses to testers' presence during the course of interactions with landlords and brokers in the field. Sales efforts can include inducements to rent an apartment. "Positive editorializing" captures instances in which landlords or brokers express their opinions about aspects of the neighborhood or apartment that a prospective tenant might find appealing. Such attempts to "talk up" a neighborhood present a favorable picture of the inhabitants, character, safety, and other features of an area. Each of the five responses was coded as either present or not present in the selected open-ended case-tester-level data.

Once the "training set" of documents was coded in this manner, we used natural language processing algorithms to classify the remainder of the cases. For each of the five codes, we took the classifications generated via *maximum entropy*[12] as our dichotomous measures of subjective

[11] Cohen's kappa for this subset ranged from 0.23-0.61, which reflects "fair" agreement on the low end to "substantial" agreement on the high end (Viera and Garrett 2005).

[12] Berger et al. (1996)

treatment due to its superior performance compared to *support vector machines*,[13] another well-known algorithm. As a validity check, we found that maximum entropy returned approximately the same proportion of codes as the original training set.

Nearly all of the classifications were made with greater than 90% confidence. The share of total case-tester-level responses classified as containing instances of positive editorializing was the highest at 79.4%. Just over half of the data, 50.4%, exhibited landlord/broker professionalism as indicated by testers' open-ended responses and the subsequent automated text analysis. Nearly half, 42.9%, of responses contained evidence of sales efforts. The other two codes do not indicate widespread behavior: 5.8% of responses showed evidence of praise for testers' rental qualifications, and 0.7% showed positive responses to an aspect of a tester's background.

We follow the index construction method used by Kling et al. (2007).[14] The resulting index measure is a continuous scale that ranges from -2 to 2, where 2 means 100% discrimination against the minority tester, and -2 means 100% discrimination against the majority tester.[15] A value of 0 means that both testers were treated equally.

C.4 Variance Decomposition of Discrimination Measures

We conduct a descriptive analysis to assess the sources of variation in the discrimination measures we use. Here, we decompose the variance of the outcome indicators of how landlords treated different testers. We focus in particular on decomposing the variance in outcome indicators by individual testers and by testers' racial group membership. By decomposing the variance in these indicators in this way, we develop a better understanding of our estimate of the baseline level of discrimination.[16]

We fit landlord-tester-level data to the following non-nested hierarchical model that models y_{ij}, an indicator for whether tester i receives certain types of treatment (in receiving a callback or an offer for a unit; or receiving praise about his or her qualifications to rent; sales efforts; positive comments about his or her background; positive editorial comments about the area; or general signals of professionalism) from landlord j as:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\mu_{ij} + \alpha_{i[ij]} + \beta_{k[ij]} + \gamma X_{ij}, \sigma_y^2) \forall (i \in I, j \in J) \\ \alpha_i &\sim \mathcal{N}(0, \sigma_\alpha^2) \text{ for } i = 1, ..., I, \\ \beta_k &\sim \mathcal{N}(0, \sigma_\beta^2) \text{ for } k = 1, ..., K. \end{aligned}$$

In this model, i indexes testers (I is the total number of testers), j indexes landlords (J is the total number of landlords), and k indexes the racial group membership of the tester ($K = 3$ is

[13] Joachims (1998)

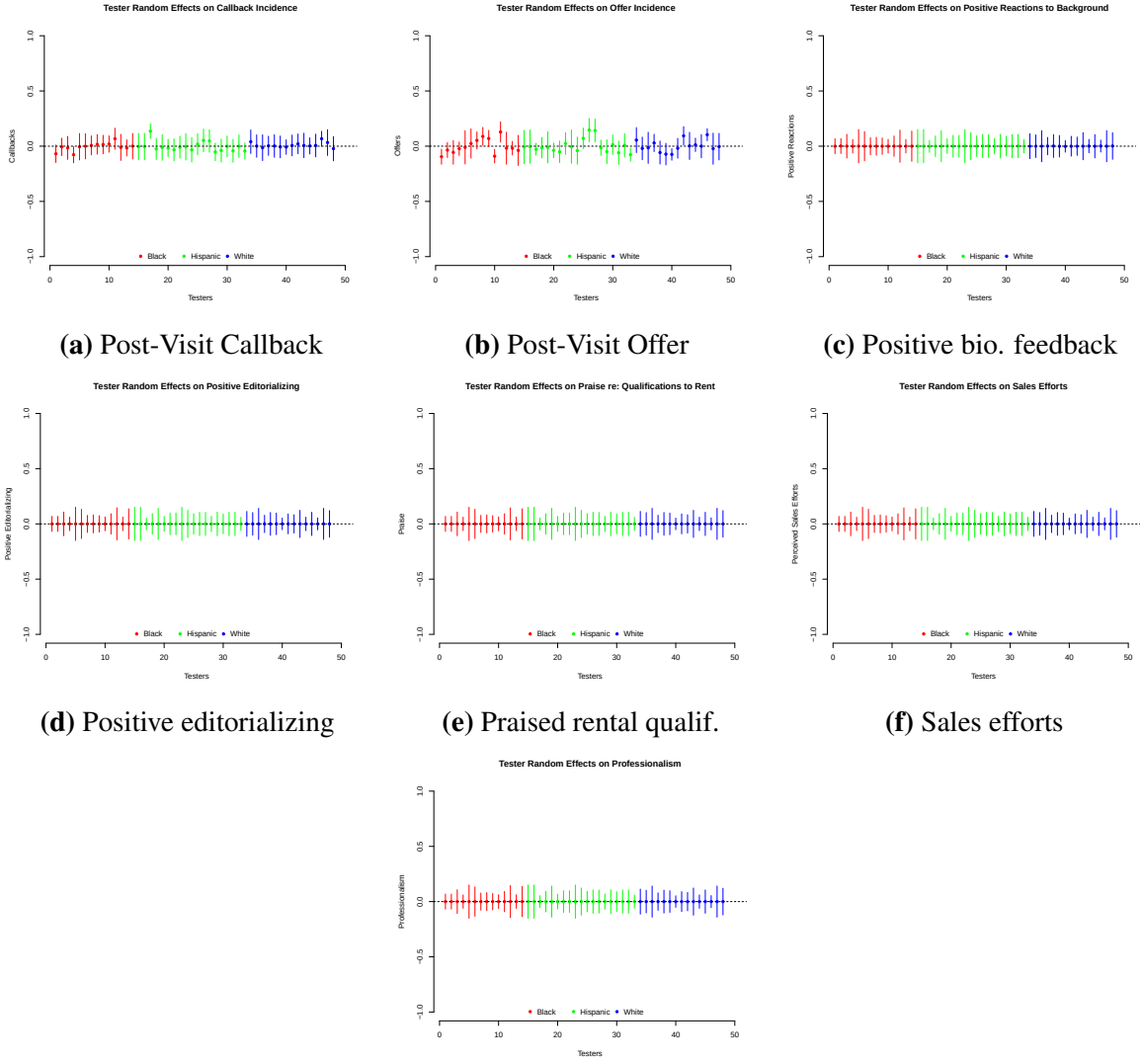
[14] For a set of outcome indicator variables $Y_1, ..., Y_k$ the value of each indicator variable for a given observation is differenced by the control group mean value of the corresponding variable; this difference is then divided by the standard deviation of the corresponding variable among the control group. The transformed indicator variables are then summed and divided by the total number of indicator variables to create a standardized summary index measure for that observation.

[15] Alternatively, these values may be interpreted as 100% favorable treatment toward the majority tester and 100% favorable treatment toward the minority tester, respectively.

[16] Because landlords' assignment to treatment is random and the ignorability of treatment assignment is procedure-driven, what we learn from this descriptive variance decomposition exercise does not have any bearing on what we learn about average causal effects.

the total number of racial groups); y_{ij} is a case-by-tester-level indicator of treatment; μ_{ij} is the mean level of y_{ij}; $\alpha_{i[ij]}$ is a tester random effect; $\beta_{k[ij]}$ is the race-level random effect; X_{ij} is a matrix of pretreatment covariates, including team gender, call order, day of the week (Monday through Friday), partnership status (partnered or single), the number of bedrooms of the listed unit, and whether a tester's qualifications elicited any negative or skeptical comments when making an appointment over the phone; and γ is a vector of coefficients on the covariates in X_{ij}.

Figure A4 plots tester-level random effects estimated from multilevel models of callback, offer, and coded subjective indicators of treatment during in-person visits. Models incorporate tester- and race-level random effects, and the aforementioned covariates.



(a) Post-Visit Callback

(b) Post-Visit Offer

(c) Positive bio. feedback

(d) Positive editorializing

(e) Praised rental qualif.

(f) Sales efforts

(g) Professionalism

Figure A4: Estimates of Tester Random Effects on Outcome Measures. This graph displays estimates and 95% confidence intervals for individual tester-level random effects $\alpha_{i[ij]}$ from the above non-nested hierarchical model.

Across all seven figures, within-group variation is clearly greater than between-group variation. Most of the random effects have confidence intervals that overlap with 0: they do not reach

statistical significance at the $p = 0.05$ level. Finally, to the extent that individual tester random effects differ significantly from 0, these figures do not seem to support any consistent pattern in incidence by group.

The main finding from these models is that the bulk of variation between racial groups (putting aside treatment assignment) is explained by individual-level random effects rather than race-level random effects, which could partially be a function of the large number of testers (48) relative to the number of racial/ethnic groups (3).

In addition, inspection of the variance of varying tester intercepts provides leverage to address concerns about measurement error, specifically whether estimates of discrimination levels are driven by the particular composition of testers in each racial group.

We estimate the above model for each of the objective outcome indicators (receiving a callback and receiving an offer) among the control group and among the experimental sample. We find that the estimated variance of the varying tester intercepts is negligible and infer that discrimination levels are not driven by the particular composition of testers in each racial group. In both samples, we find that the estimated variance of the varying tester intercepts is negligible (.003 for callbacks and .005 for offers). The bulk of the variance is in the residual: the estimated residual variance is .141 for callbacks and .076 for offers in the control group and .0135 for callbacks and .074 for offers in the experimental sample. We therefore infer that the discrimination levels are not driven by the composition of testers in each racial group.

D SAMPLE CHARACTERISTICS

D.1 Distribution of Subjects across Randomization Blocks

	Assigned to Control		Assigned to Monitoring		Assigned to Punitive	
Block	N	%	N	%	N	%
Regime 1: 13 Apr 2012 - 9 Sep 2012						
Brooklyn	14	0.021	11	0.017	17	0.026
Bronx	4	0.006	3	0.005	2	0.003
Manhattan	13	0.02	12	0.018	18	0.028
Queens	2	0.003	3	0.005	6	0.009
Staten Island	1	0.002	1	0.002	0	0
Likely Discrimination Frame	8	0.012	8	0.012	9	0.014
Regime 2: 10 Sep 2012 - 7 May 7, 2013						
Brooklyn	63	0.096	24	0.037	29	0.044
Bronx	21	0.032	10	0.015	11	0.017
Manhattan	50	0.077	32	0.049	23	0.035
Queens	28	0.043	10	0.015	14	0.021
Staten Island	8	0.012	2	0.003	3	0.005
Likely Discrimination Frame	11	0.017	4	0.006	4	0.006
Regime 3: 8 May 2013 to 20 Dec 2013						
Brooklyn	13	0.02	23	0.035	14	0.021
Bronx	6	0.009	3	0.005	8	0.012
Manhattan	25	0.038	18	0.028	25	0.038
Queens	8	0.012	7	0.011	14	0.021
Staten Island	4	0.006	3	0.005	3	0.005
Total	279	0.427	174	0.266	200	0.306

Table A3: Distribution of Experimental Subjects by Randomization Block. Cells contain counts of the number and share of experimental subjects (i.e., landlords and brokers) randomly assigned to the control group, the monitoring condition, and the punitive condition, by block. Each regime denotes a different randomization procedure used. The probability of assignment to each condition is described by regime. Percentages may not sum to 100 due to rounding.

D.2 Characteristics of Advertised Housing Stock

Table A4 summarizes the characteristics of the advertised housing units in the audit sample (Panel I), in the experimental sample (II), in the subset of cases in the experimental sample assigned to any treatment condition (III), and in the subset of cases in the experimental sample assigned to the control condition (IV). We briefly describe the characteristics of housing units in each of these samples.

- **Audit Sample** – For advertised housing units in the audit sample (N=2711), the mean advertised monthly asking rental price is $2,238 and the median advertised monthly asking rental price is $1,850. The advertised monthly asking rental price ranges from $400 per month to $15,000 per month. The mean advertised number of bedrooms is 0.94 and the average advertised square footage of a listed unit is 1,021.35 square feet. Of the listed units in the audit sample, 57.29% were listed by brokers.
- **Experimental Sample** – For advertised housing units in the experimental sample (N=653), the mean advertised monthly asking rental price is $2,435 and the median advertised monthly asking rental price is $2,200. The advertised monthly asking rental price ranges from $750 per month to $9,495 per month. The mean advertised number of bedrooms is 0.88 and the average advertised square footage of a listed unit is 1,017.49 square feet. Of the listed units in the audit sample, 84.53% were listed by brokers.
- **Cases Assigned to Any Treatment Group** – For advertised housing units corresponding to cases in the experimental sample assigned to any treatment condition (N=374), the mean advertised monthly asking rental price is $2,420 and the median advertised monthly asking rental price is $2,200. The advertised monthly asking rental price ranges from $750 per month to $9,495 per month. The mean advertised number of bedrooms is 0.85 and the average advertised square footage of a listed unit is 915 square feet. Of the listed units in the audit sample, 83.42% were listed by brokers.
- **Cases Assigned to the Control Group** – For advertised housing units corresponding to cases in the experimental sample assigned to the control condition (N=279), the mean advertised monthly asking rental price is $2,456and the median advertised monthly asking rental price is $2,025. The advertised monthly asking rental price ranges from $850 per month to $8,900 per month. The mean advertised number of bedrooms is 0.92 and the average advertised square footage of a listed unit is 1,116.68 square feet. Of the listed units in the audit sample, 86.02% were listed by brokers.

Variable	I. Audit Sample		II. Experimental Sample		III. Any Treatment Group		IV. Control Group	
		% of Audit		% of Audit		% of Exp.		% of Exp.
Panel A	N	Sample	N	Sample	N	Sample	N	Sample
Number of Units								
Total	2711	(100%)	653	(24.09%)	374	(57.27%)	279	(42.73%)
Bronx	337	(100%)	68	(20.18%)	37	(54.41%)	31	(45.59%)
Brooklyn	801	(100%)	208	(25.97%)	118	(56.73%)	90	(43.27%)
Manhattan	668	(100%)	216	(32.34%)	128	(59.26%)	88	(40.74%)
Queens	495	(100%)	92	(18.59%)	54	(58.7%)	38	(41.3%)
Staten Island	254	(100%)	25	(9.84%)	12	(48%)	13	(52%)
Likely Discrimination Frame	156	(100%)	44	(28.21%)	25	(56.82%)	19	(43.18%)
Panel B	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Monthly Asking Rental Price ($)								
Total	2238	(1295)	2435	(1204)	2420	(1157)	2456	(1272)
Bronx	1419	(512)	1578	(662)	1563	(628)	1597	(719)
Brooklyn	2194	(1047)	2319	(957)	2285	(855)	2370	(1096)
Manhattan	3252	(1600)	3163	(1447)	3134	(1404)	3206	(1520)
Queens	1718	(562)	1885	(558)	1847	(450)	1941	(693)
Staten Island	1383	(511)	1336	(596)	1369	(655)	1293	(559)
Likely Discrimination Frame	2479	(1452)	2321	(736)	2336	(778)	2302	(697)
Panel C	Median		Median		Median		Median	
Monthly Asking Rental Price ($)								
Total	1850		2200		2200		2025	
Bronx	1325		1400		1400		1400	
Brooklyn	1992		2200		2299		2050	
Manhattan	2950		2898		2900		2872	
Queens	1600		1850		1850		1850	
Staten Island	1300		1100		1195		1100	
Likely Discrimination Frame	2272		2250		2200		2300	
Panel D	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number of Bedrooms								
Total	0.94	(1.22)	0.88	(1.19)	0.85	(1.17)	0.92	(1.21)
Bronx	0.82	(1.2)	0.85	(1.16)	0.86	(1.23)	0.84	(1.1)
Brooklyn	1	(1.23)	1.01	(1.3)	0.97	(1.26)	1.07	(1.36)
Manhattan	0.85	(1.15)	0.69	(1.04)	0.63	(0.98)	0.76	(1.13)
Queens	0.81	(1.15)	0.59	(0.9)	0.56	(0.9)	0.63	(0.91)
Staten Island	0.89	(1.29)	0.8	(1.5)	1.08	(1.83)	0.54	(1.13)
Likely Discrimination Frame	1.88	(1.23)	1.95	(1.01)	1.92	(1)	2	(1.05)
Panel E	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Square Footage								
Total	1021.35	(618.52)	1017.49	(448.09)	915	(449.36)	1116.68	(430.93)
Bronx	999.59	(352.42)	1328.75	(453.84)	1100	(453.84)	1405	(523.52)
Brooklyn	1041.87	(462.63)	1125.94	(638.07)	1085.11	(661.49)	1171.88	(652.64)
Manhattan	1029.82	(495.06)	926.79	(350.32)	777.43	(334.36)	1135.9	(262.42)
Queens	914.92	(777.31)	917.6	(260.29)	866.67	(288.68)	994	(291.33)
Staten Island	1203.25	(981.19)	1233.33	(305.51)	1300	(305.51)	1200	(424.26)
Likely Discrimination Frame	1011.5	(568.51)	885	(272.45)	900	(141.42)	880	(315.91)
Panel F	N	%	N	%	N	%	N	%
Listed by Broker								
Total	1553	(57.29%)	552	(84.53%)	312	(83.42%)	240	(86.02%)
Bronx	172	(51.04%)	51	(75%)	27	(72.97%)	24	(77.42%)
Brooklyn	452	(56.43%)	178	(85.58%)	102	(86.44%)	76	(84.44%)
Manhattan	439	(65.72%)	192	(88.89%)	112	(87.5%)	80	(90.91%)
Queens	256	(51.72%)	75	(81.52%)	43	(79.63%)	32	(84.21%)
Staten Island	133	(52.36%)	14	(56%)	5	(41.67%)	9	(69.23%)
Likely Discrimination Frame	101	(64.74%)	42	(95.45%)	23	(92%)	19	(100%)

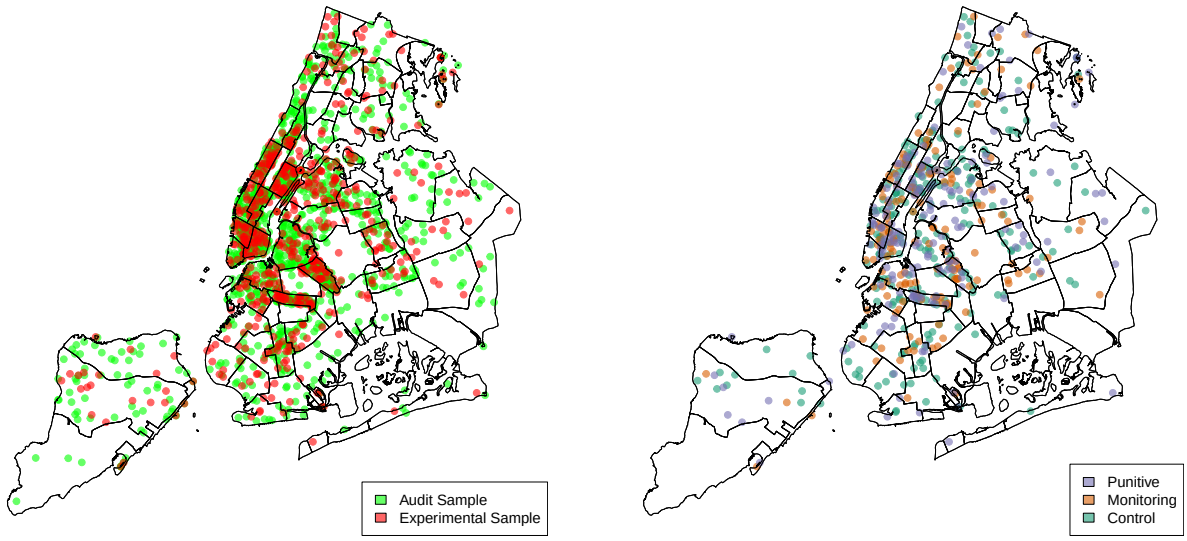
Table A4: Selected Characteristics of Housing Units in the Audit and Experimental Samples

D.3 Distribution of Cases Across Boroughs, by Sample

Table A5 and the maps in Figure A5 summarize the distribution of cases across boroughs by sample. Focusing on the sampling blocks corresponding to the five boroughs, the distribution of the audit sample (including all assigned cases, whether or not all testers in a group were able to schedule an appointment) is very close to the distribution of known rental units in New York City. Using the latest New York City Housing and Vacancy Survey (NYCHVS) data from 2011, we can see that the proportion of cases in each borough in the audit sample closely tracks the overall proportions, with the exception that Staten Island units appear to be overrepresented in our sample at the expense of apartments in the Bronx. In the final experimental sample, Manhattan and Brooklyn are overrepresented at the expense mainly of Queens. According to the NYCHVS, the overall net vacancy rate across all boroughs was 3.12%: 3.23% in the Bronx, 2.61% in Brooklyn, 2.80% in Manhattan, 3.79% in Queens, and 6.65% in Staten Island.

	Citywide 2011		Audit Sample		Experimental Sample		Control Group Only	
Borough	# Units	%	# Units	%	# Units	%	# Units	%
Bronx	388,022	17.86	337	13.19	68	11.17	31	11.11
Brooklyn	691,178	31.81	801	31.35	208	34.15	90	32.26
Manhattan	587,313	27.03	668	26.14	216	35.47	88	31.54
Queens	449,108	20.67	495	19.37	92	15.11	38	13.62
Staten Island	57,013	2.62	254	9.94	25	4.11	13	4.66
Total	2,172,634	100	2,555	100	609	100	279	100

Table A5: Distribution of Rental Units Across Boroughs, by Sample. The full audit and experimental samples are compared to the totals citywide (from the 2011 New York City Housing and Vacancy Survey). Cases from the likely-discrimination block not included.



(a) Audit versus experimental sample **(b)** By treatment group

Figure A5: Map of the Geographic Distribution of Housing Units Corresponding to Advertised Listings, Data Aggregated to Community Board Level. The exact locations are not reported to maintain the anonymity of study subjects. The geographic location of advertised units is aggregated to the Community Board level. The number of housing units within each Community Board is then randomly distributed within the boundaries of the Community Board.

E SUPPLEMENTARY ANALYSES

E.1 Discrimination Levels

	I. Cases Assigned to Control Group					II. Cases Assigned to Any Treatment Group					III. All Cases in Experimental Sample				
	Mean level of favorable treatment		Difference			Mean level of favorable treatment		Difference			Mean level of favorable treatment		Difference		
Net Measure of Discrimination	Majority	Minority	(Maj.-Min.)	*p*-value	[N]	Majority	Minority	(Maj.-Min.)	*p*-value	[N]	Majority	Minority	(Maj.-Min.)	*p*-value	[N]
A. White vs. Black Testers															
Making the Appointment															
Landlord/broker honored appointment	0.993	0.996	-0.004	(0.318)	[279]	0.995	0.987	0.008	(0.18)	[374]	0.994	0.991	0.003	(0.415)	[653]
Subjective Evaluations of Interaction Quality															
Perceived sales efforts	0.417	0.509	-0.091	(0.049)	[253]	0.436	0.421	0.015	(0.715)	[334]	0.428	0.46	-0.032	(0.294)	[587]
Received praise about rental qualifications	0.061	0.068	-0.008	(0.743)	[253]	0.052	0.045	0.008	(0.652)	[334]	0.056	0.055	0.001	(0.947)	[587]
Positive reactions to testers' background	0.017	0.004	0.013	(0.174)	[253]	0.007	0.003	0.003	(0.587)	[334]	0.011	0.004	0.007	(0.161)	[587]
Positive editorializing	0.817	0.765	0.052	(0.165)	[253]	0.797	0.743	0.054	(0.121)	[334]	0.806	0.753	0.053	(0.038)	[587]
Professionalism	0.522	0.483	0.039	(0.404)	[253]	0.498	0.507	-0.008	(0.836)	[334]	0.508	0.496	0.012	(0.691)	[587]
Post-Visit Follow-Up															
Received post-visit callback	0.215	0.168	0.047	(0.107)	[279]	0.187	0.131	0.056	(0.018)	[374]	0.199	0.147	0.052	(0.005)	[653]
Received post-visit offer for unit	0.118	0.09	0.029	(0.239)	[279]	0.094	0.08	0.013	(0.467)	[374]	0.104	0.084	0.02	(0.178)	[653]
B. White vs. Hispanic Testers															
Making the Appointment															
Landlord/broker honored appointment	0.993	0.996	-0.004	(0.318)	[279]	0.995	0.995	0	(1)	[374]	0.994	0.995	-0.002	(0.705)	[653]
Subjective Evaluations of Interaction Quality															
Perceived sales efforts	0.417	0.45	-0.033	(0.486)	[252]	0.436	0.394	0.042	(0.295)	[334]	0.428	0.418	0.01	(0.737)	[586]
Received praise about rental qualifications	0.061	0.045	0.015	(0.467)	[252]	0.052	0.081	-0.028	(0.164)	[334]	0.056	0.066	-0.01	(0.512)	[586]
Positive reactions to testers' background	0.017	0.009	0.008	(0.441)	[252]	0.007	0.007	0	(0.979)	[334]	0.011	0.008	0.003	(0.56)	[586]
Positive editorializing	0.817	0.786	0.031	(0.411)	[252]	0.797	0.838	-0.042	(0.186)	[334]	0.806	0.816	-0.011	(0.66)	[586]
Professionalism	0.522	0.591	-0.069	(0.14)	[252]	0.498	0.458	0.04	(0.321)	[334]	0.508	0.515	-0.006	(0.843)	[586]
Post-Visit Follow-Up															
Received post-visit callback	0.215	0.154	0.061	(0.019)	[279]	0.187	0.171	0.016	(0.503)	[374]	0.199	0.164	0.035	(0.099)	[653]
Received post-visit offer for unit	0.118	0.061	0.057	(0.011)	[279]	0.094	0.08	0.013	(0.476)	[374]	0.104	0.072	0.032	(0.04)	[653]
C. Black vs. Hispanic Testers															
Making the Appointment															
Landlord/broker honored appointment	0.996	0.996	0	(NaN)	[279]	0.987	0.995	-0.008	(0.18)	[374]	0.991	0.995	-0.005	(0.18)	[653]
Subjective Evaluations of Interaction Quality															
Perceived sales efforts	0.509	0.45	0.059	(0.213)	[251]	0.421	0.394	0.027	(0.501)	[339]	0.46	0.418	0.042	(0.169)	[590]
Received praise about rental qualifications	0.068	0.045	0.023	(0.292)	[251]	0.045	0.081	-0.036	(0.069)	[339]	0.055	0.066	-0.011	(0.472)	[590]
Positive reactions to testers' background	0.004	0.009	-0.005	(0.532)	[251]	0.003	0.007	-0.003	(0.572)	[339]	0.004	0.008	-0.004	(0.403)	[590]
Positive editorializing	0.765	0.786	-0.021	(0.586)	[251]	0.743	0.838	-0.095	(0.004)	[339]	0.753	0.816	-0.063	(0.013)	[590]
Professionalism	0.483	0.591	-0.108	(0.021)	[251]	0.507	0.458	0.049	(0.235)	[339]	0.496	0.515	-0.018	(0.555)	[590]
Post-Visit Follow-Up															
Received post-visit callback	0.168	0.154	0.014	(0.587)	[279]	0.131	0.171	-0.04	(0.079)	[374]	0.147	0.164	-0.017	(0.329)	[653]
Received post-visit offer for unit	0.09	0.061	0.029	(0.17)	[279]	0.08	0.08	0	(1)	[374]	0.084	0.072	0.012	(0.359)	[653]

Table A6: Baseline Incidence of Discrimination: In-Person and Post-Visit

E.2 Main ITT estimates

Outcome	Estimate	SE	t	p-value	95% CI
I. Monitoring vs. Control					
A. White vs. Black					
Index measure of favorable in-person interactions	-0.015	0.052	-0.285	(0.388)	[-0.117, 0.087]
Received post-visit callback	-0.002	0.045	-0.035	(0.486)	[-0.091, 0.088]
Received post-visit offer for unit	-0.003	0.036	-0.091	(0.464)	[-0.075, 0.068]
B. White vs. Hispanic					
Index measure of favorable in-person interactions	-0.061	0.057	-1.079	(0.141)	[-0.172, 0.05]
Received post-visit callback	-0.036	0.043	-0.837	(0.201)	[-0.121, 0.049]
Received post-visit offer for unit	-0.017	0.034	-0.496	(0.31)	[-0.084, 0.05]
C. Black vs. Hispanic					
Index measure of favorable in-person interactions	-0.089	0.052	-1.7	(0.09)	[-0.192, 0.014]
Received post-visit callback	-0.035	0.042	-0.822	(0.412)	[-0.118, 0.048]
Received post-visit offer for unit	-0.014	0.031	-0.444	(0.657)	[-0.074, 0.047]
II. Punitive vs. Control					
A. White vs. Black					
Index measure of favorable in-person interactions	0.007	0.053	0.139	(0.555)	[-0.097, 0.112]
Received post-visit callback	0.019	0.042	0.456	(0.676)	[-0.064, 0.103]
Received post-visit offer for unit	0.018	0.035	0.515	(0.697)	[-0.051, 0.087]
B. White vs. Hispanic					
Index measure of favorable in-person interactions	0.048	0.055	0.863	(0.806)	[-0.061, 0.157]
Received post-visit callback	-0.066	0.041	-1.596	(0.056)	[-0.147, 0.015]
Received post-visit offer for unit	-0.021	0.034	-0.618	(0.268)	[-0.089, 0.046]
C. Black vs. Hispanic					
Index measure of favorable in-person interactions	0.039	0.049	0.791	(0.429)	[-0.057, 0.134]
Received post-visit callback	-0.085	0.041	-2.097	(0.037)	[-0.165, -0.005]
Received post-visit offer for unit	-0.039	0.033	-1.172	(0.242)	[-0.105, 0.027]
III. Punitive vs. Monitoring					
A. White vs. Black					
Index measure of favorable in-person interactions	0.018	0.058	0.319	(0.75)	[-0.095, 0.132]
Received post-visit callback	0.033	0.049	0.665	(0.506)	[-0.064, 0.129]
Received post-visit offer for unit	0.026	0.037	0.685	(0.494)	[-0.048, 0.099]
B. White vs. Hispanic					
Index measure of favorable in-person interactions	0.107	0.063	1.701	(0.09)	[-0.017, 0.231]
Received post-visit callback	-0.019	0.049	-0.391	(0.696)	[-0.116, 0.078]
Received post-visit offer for unit	0.002	0.038	0.047	(0.962)	[-0.073, 0.077]
C. Black vs. Hispanic					
Index measure of favorable in-person interactions	0.139	0.057	2.431	(0.016)	[0.026, 0.251]
Received post-visit callback	-0.052	0.047	-1.11	(0.268)	[-0.144, 0.04]
Received post-visit offer for unit	-0.024	0.036	-0.667	(0.505)	[-0.094, 0.046]

Table A7: Estimated Effects of Messaging on Net Discrimination Levels. Cells contain ITT estimates from OLS models with inverse probability weights and block fixed effects. For each reference group versus comparison group pairing, outcomes are net discrimination measures against the comparison group relative to the reference group. Estimated effects that are positive (negative) are interpreted as increases (decreases) in net discrimination against the comparison group relative to the reference group. Estimated p-values are reported in parentheses; p-values correspond to a one-sided test of the null hypothesis of equality of means for the monitoring-control and punitive-control comparisons, and 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.

E.3 Unweighted ITT Estimates

This section presents ITT estimates of treatment assignment on objective net discrimination outcomes using raw unweighted data. We report mean levels of favorable treatment by tester race; differences in mean levels of favorable treatment across treatment groups, by tester race; differences in mean levels of favorable treatment between tester groups, by treatment (i.e., the mean net discrimination levels by treatment group); and the difference in net discrimination levels between treatment groups (i.e., estimates of the effects of treatment messaging on net discrimination levels). These estimates are presented only to provide a sense of the raw data, and should not be interpreted as the causal effects of sending different messages on net discrimination because they do not account for the randomization procedure.

	I. Mean Levels, by Group			**II. Differences in Means**		
	Control	Monitoring	Punitive	Monitoring vs. Control	Punitive vs. Control	Punitive vs. Monitoring
Panel A. Percent Favorable						
White Testers	0.215	0.184	0.19	-0.031	-0.025	0.006
				(0.418)	(0.5)	(0.881)
Black Testers	0.168	0.144	0.12	-0.025	-0.048	-0.024
				(0.478)	(0.133)	(0.502)
Hispanic Testers	0.154	0.155	0.185	0.001	0.031	0.03
				(0.976)	(0.378)	(0.444)
Panel B. Net Discrimination						
(% Majority Favorable - % Minority Favorable)						
White vs. Black Testers	0.047	0.04	0.07	-0.006	0.023	0.03
	(0.107)	(0.264)	(0.026)	(0.417)	(0.675)	(0.58)
White vs. Hispanic Testers	0.061	0.029	0.005	-0.032	-0.056	-0.024
	(0.019)	(0.425)	(0.876)	(0.271)	(0.086)	(0.575)
Black vs. Hispanic Testers	0.014	-0.011	-0.065	-0.026	-0.079	-0.054
	(0.587)	(0.733)	(0.037)	(0.346)	(0.031)	(0.246)
Sample Size	279	174	200	453	479	374

Table A8: Unweighted ITT Estimates of Messaging on Net Discrimination in Receiving a Post-Visit Callback. Cells in the upper-left quadrant (quadrants denoted by double lines) contain estimates of the levels of favorable treatment toward each population (white, black, Hispanic testers). Cells in the 'Control' column in the bottom-left quadrant contain estimates of baseline net discrimination rates, defined as the share of favorable majority treatment minus the share of favorable minority treatment. The remaining cells in the bottom-left quadrant contain estimates of net discrimination rates in non-control treatment groups. Cells in the upper-right quadrant contain estimates of the treatment effects on favorable treatment rates for specific populations. Cells in the bottom-right quadrant contain unweighted ITT estimates. We show estimates without weighting for different probabilities of assignment to treatment to provide a sense of the raw data. Estimated p-values (shown in parentheses), which are the probability of obtaining an effect at least as large (in absolute value) as the one observed in the actual experiment for the monitoring-control and punitive-control (punitive-monitoring) comparisons.

	I. Mean Levels, by Group			II. Differences in Means		
	Control	Monitoring	Punitive	Monitoring vs. Control	Punitive vs. Control	Punitive vs. Monitoring
Panel A. Percent Favorable						
White Testers	0.118	0.08	0.105	-0.038	-0.013	0.025
				(0.183)	(0.648)	(0.414)
Black Testers	0.09	0.08	0.08	-0.009	-0.01	0
				(0.734)	(0.709)	(0.987)
Hispanic Testers	0.061	0.063	0.095	0.002	0.034	0.032
				(0.922)	(0.178)	(0.254)
Panel B. Net Discrimination						
(% Majority Favorable - % Minority Favorable)						
White vs. Black Testers	0.029	0	0.025	-0.029	-0.004	0.025
	(0.239)	(1)	(0.319)	(0.084)	(0.239)	(0.51)
White vs. Hispanic Testers	0.057	0.017	0.01	-0.04	-0.047	-0.007
	(0.011)	(0.514)	(0.706)	(0.052)	(0.012)	(0.698)
Black vs. Hispanic Testers	0.029	0.017	-0.015	-0.011	-0.044	-0.032
	(0.17)	(0.44)	(0.565)	(0.413)	(0.048)	(0.251)
Sample Size	279	174	200	453	479	374

Table A9: Unweighted ITT Estimates of Messaging on Net Discrimination in Receiving a Post-Visit Offer for the Unit. Cells in the upper-left quadrant (quadrants denoted by double lines) contain estimates of the levels of favorable treatment toward each population (white, black, Hispanic testers). Cells in the 'Control' column in the bottom-left quadrant contain estimates of baseline net discrimination rates, defined as the share of favorable majority treatment minus the share of favorable minority treatment. The remaining cells in the bottom-left quadrant contain estimates of net discrimination rates in non-control treatment groups. Cells in the upper-right quadrant contain estimates of the treatment effects on favorable treatment rates for specific populations. Cells in the bottom-right quadrant contain unweighted ITT estimates. We show estimates without weighting for different probabilities of assignment to treatment to provide a sense of the raw data. Estimated p-values (shown in parentheses), which are the probability of obtaining an effect at least as large (in absolute value) as the one observed in the actual experiment for the monitoring-control and punitive-control (punitive-monitoring) comparisons.

E.4 ITT Estimates from a Three-Group Parametric Estimator

As a sensitivity check on our main estimation results, we estimate the ITTs using a three-group parametric estimator with block fixed effects and inverse probability weights.

Outcome: Net Discrimination (Reference vs. Comparison Group)	Estimate	SE	t	*p*-value
I. Monitoring vs. Control				
A. White vs. Black				
Index measure of favorable in-person interactions	-0.012	0.054	-0.225	(0.411)
Received post-visit callback	-0.009	0.046	-0.189	(0.425)
Received post-visit offer for unit	-0.007	0.036	-0.194	(0.423)
B. White vs. Hispanic				
Index measure of favorable in-person interactions	-0.059	0.059	-0.999	(0.159)
Received post-visit callback	-0.044	0.045	-0.987	(0.162)
Received post-visit offer for unit	-0.023	0.036	-0.654	(0.257)
C. Black vs. Hispanic				
Index measure of favorable in-person interactions	-0.092	0.053	-1.745	(0.082)
Received post-visit callback	-0.036	0.043	-0.822	(0.411)
Received post-visit offer for unit	-0.016	0.033	-0.489	(0.625)
II. Punitive vs. Control				
A. White vs. Black				
Index measure of favorable in-person interactions	0.01	0.052	0.198	(0.579)
Received post-visit callback	0.02	0.044	0.454	(0.675)
Received post-visit offer for unit	0.016	0.035	0.465	(0.679)
B. White vs. Hispanic				
Index measure of favorable in-person interactions	0.05	0.056	0.893	(0.814)
Received post-visit callback	-0.068	0.043	-1.568	(0.059)
Received post-visit offer for unit	-0.025	0.035	-0.72	(0.236)
C. Black vs. Hispanic				
Index measure of favorable in-person interactions	0.039	0.05	0.781	(0.435)
Received post-visit callback	-0.088	0.042	-2.1	(0.036)
Received post-visit offer for unit	-0.041	0.032	-1.278	(0.202)
III. Punitive vs. Monitoring				
A. White vs. Black				
Index measure of favorable in-person interactions	0.022	0.055	0.412	(0.68)
Received post-visit callback	0.029	0.046	0.631	(0.528)
Received post-visit offer for unit	0.023	0.036	0.647	(0.518)
B. White vs. Hispanic				
Index measure of favorable in-person interactions	0.109	0.058	1.879	(0.061)
Received post-visit callback	-0.024	0.045	-0.532	(0.595)
Received post-visit offer for unit	-0.001	0.036	-0.042	(0.966)
C. Black vs. Hispanic				
Index measure of favorable in-person interactions	0.132	0.053	2.48	(0.014)
Received post-visit callback	-0.053	0.043	-1.216	(0.224)
Received post-visit offer for unit	-0.025	0.033	-0.752	(0.452)

Table A10: Sensitivity Analysis: Estimated Effects of Messaging on Net Discrimination Levels. Cells contain ITT estimates from OLS models with inverse probability weights and block fixed effects. The estimator is a three-group parametric estimator where the data are not subset prior to estimation. For each reference group versus comparison group pairing, outcomes are net discrimination measures against the comparison group relative to the reference group. Estimated effects that are positive (negative) are interpreted as increases (decreases) in net discrimination against the comparison group relative to the reference group. Estimated *p*-values are reported in parentheses; *p*-values correspond to a one-sided test of the null hypothesis of equality of means for the monitoring-control and punitive-control comparisons, and 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.

E.5 Predicted Treatment and Control Means and Estimated Percent Differences

Outcome: Net Discrimination (Reference vs. Comparison Group)	Predicted Treatment Mean	Predicted Comparison Mean	Estimated Impact ($\widehat{ITT}$)	Percent Difference
I. Monitoring vs. Control				
A. White vs. Black				
Index measure of favorable in-person interactions	0.023	0.038	-0.015	-39.204%
Received post-visit callback	0.006	0.007	-0.002	-21.372%
Received post-visit offer for unit	0.005	0.008	-0.003	-40.863%
B. White vs. Hispanic				
Index measure of favorable in-person interactions	-0.095	-0.034	-0.061	-178.801%
Received post-visit callback	0.001	0.038	-0.036	-96.164%
Received post-visit offer for unit	0.023	0.04	-0.017	-42.45%
C. Black vs. Hispanic				
Index measure of favorable in-person interactions	-0.162	-0.073	-0.089	-121.727%
Received post-visit callback	-0.004	0.03	-0.035	-114.689%
Received post-visit offer for unit	0.018	0.032	-0.014	-42.853%
II. Punitive vs. Control				
A. White vs. Black				
Index measure of favorable in-person interactions	0.007	0	0.007	7904.738%
Received post-visit callback	0.012	-0.007	0.019	281.195%
Received post-visit offer for unit	0.032	0.014	0.018	127.36%
B. White vs. Hispanic				
Index measure of favorable in-person interactions	0.009	-0.039	0.048	122.284%
Received post-visit callback	-0.005	0.061	-0.066	-108.818%
Received post-visit offer for unit	0.013	0.034	-0.021	-62.201%
C. Black vs. Hispanic				
Index measure of favorable in-person interactions	-0.012	-0.051	0.039	76.051%
Received post-visit callback	-0.018	0.068	-0.085	-126.385%
Received post-visit offer for unit	-0.019	0.02	-0.039	-196.921%
III. Punitive vs. Monitoring				
A. White vs. Black				
Index measure of favorable in-person interactions	0.019	0	0.018	6621.1%
Received post-visit callback	0.055	0.022	0.033	147.447%
Received post-visit offer for unit	0.018	-0.008	0.026	322.614%
B. White vs. Hispanic				
Index measure of favorable in-person interactions	0.014	-0.092	0.107	115.589%
Received post-visit callback	-0.02	0	-0.019	-4046.942%
Received post-visit offer for unit	-0.002	-0.004	0.002	49.995%
C. Black vs. Hispanic				
Index measure of favorable in-person interactions	-0.008	-0.147	0.139	94.696%
Received post-visit callback	-0.074	-0.023	-0.052	-229.774%
Received post-visit offer for unit	-0.019	0.004	-0.024	-551.477%

Table A11: Predicted Means, Difference, and Percent Difference from Two-Group Parametric Estimators. Predicted values calculated using estimates from OLS models with inverse probability weights and block fixed effects.

E.6 Subjective Indicators of Early Stage Discrimination

In addition to the objective measures of early stage discrimination we gathered a set of subjective measures drawn from reporter field notes. These are all pre-treatment measures that reflect the propensity of testers to interpret various responses by landlords in a positive or negative light.

We see a number of differences that are statistically significant at the 0.1% ($p < 0.001$) level or below. Broadly these show blacks reporting less skepticism or negative responses than either Hispanics or Whites. First, white testers reported encountering more skepticism about their attributes or qualifications to rent a given apartment than black testers—about twice as much (white testers in 3.2% of pursued cases experienced skeptical mentions of an attribute by landlords or brokers over the phone compared to 1.6% of those cases for black testers). The same was true of negative comments about their attributes or qualifications (3.5% vs. 1.6%). Other measures intended to capture the same difference in treatment (such as the percentage of attributes or the number of attributes mentioned) showed broadly the same pattern. This was counter to expectations and could reflect real differences in treatment, differential perceptions in treatment, or differences in reporting. The second set of statistically significant differences showed that Hispanic testers reported substantially more negative and skeptical comments about their attributes and qualifications to rent than African American testers. For example, in 3.6% of pursued cases, Hispanic testers encountered skeptical responses as compared to 1.6% of cases for African American testers. The corresponding percentages for outright negative reactions were 3% and 1.6%, respectively. As seen in the left column of Table A12, the average levels of unfavorable treatment as captured by these measures were fairly low across all groups. These findings are broadly consistent with the result reported in the main text that black confederates

	I. All Pursued Cases in Audit Sample					II. All Cases in Experimental Sample					III. Cases Assigned to Control Group				
	Mean Level		Difference			Mean Level		Difference			Mean Level		Difference		
Measure	Majority	Minority	(Maj.-Min.)	*p*-value	[N]	Majority	Minority	(Maj.-Min.)	*p*-value	[N]	Majority	Minority	(Maj.-Min.)	*p*-value	[N]
A. White vs. Black Testers															
No. of attributes brought up by landlord/broker	1.101	1.041	0.06	(0.136)	[2711]	2.023	1.936	0.087	(0.334)	[653]	2.151	1.961	0.19	(0.145)	[279]
No. attributes - skeptical response	0.053	0.027	0.026	(0.001)	[2711]	0.077	0.041	0.035	(0.073)	[653]	0.097	0.039	0.057	(0.106)	[279]
No. attributes - positive response	0.132	0.125	0.007	(0.637)	[2711]	0.262	0.256	0.006	(0.876)	[653]	0.269	0.258	0.011	(0.852)	[279]
No. attributes - neutral response	0.918	0.894	0.024	(0.501)	[2711]	1.703	1.668	0.035	(0.675)	[653]	1.821	1.685	0.136	(0.259)	[279]
No. attributes - negative response	0.051	0.022	0.029	(0)	[2711]	0.058	0.012	0.046	(0.001)	[653]	0.061	0.018	0.043	(0.064)	[279]
Pct. of attributes - skeptical response	0.012	0.008	0.005	(0.021)	[2711]	0.018	0.015	0.003	(0.602)	[653]	0.022	0.011	0.012	(0.176)	[279]
Pct. of attributes - positive response	0.033	0.033	0	(0.933)	[2711]	0.07	0.07	0.001	(0.925)	[653]	0.071	0.066	0.005	(0.743)	[279]
Pct. of attributes - neutral response	0.35	0.328	0.022	(0.026)	[2711]	0.706	0.646	0.06	(0.011)	[653]	0.745	0.666	0.078	(0.028)	[279]
Pct. of attributes - negative response	0.013	0.006	0.007	(0)	[2711]	0.014	0.003	0.011	(0.002)	[653]	0.016	0.003	0.013	(0.027)	[279]
Responded skeptically for any attribute	0.032	0.016	0.017	(0)	[2711]	0.049	0.025	0.025	(0.018)	[653]	0.054	0.018	0.036	(0.025)	[279]
Responded negatively for any attribute	0.035	0.016	0.019	(0)	[2711]	0.04	0.009	0.031	(0)	[653]	0.039	0.011	0.029	(0.032)	[279]
B. White vs. Hispanic Testers															
No. of attributes brought up by landlord/broker	1.101	1.07	0.031	(0.434)	[2711]	2.023	2.072	-0.049	(0.549)	[653]	2.151	1.989	0.161	(0.19)	[279]
No. attributes - skeptical response	0.053	0.065	-0.012	(0.241)	[2711]	0.077	0.081	-0.005	(0.831)	[653]	0.097	0.039	0.057	(0.074)	[279]
No. attributes - positive response	0.132	0.14	-0.008	(0.532)	[2711]	0.262	0.27	-0.008	(0.832)	[653]	0.269	0.24	0.029	(0.533)	[279]
No. attributes - neutral response	0.918	0.884	0.035	(0.32)	[2711]	1.703	1.75	-0.047	(0.524)	[653]	1.821	1.728	0.093	(0.383)	[279]
No. attributes - negative response	0.051	0.046	0.005	(0.543)	[2711]	0.058	0.052	0.006	(0.706)	[653]	0.061	0.022	0.039	(0.07)	[279]
Pct. of attributes - skeptical response	0.012	0.016	-0.004	(0.114)	[2711]	0.018	0.023	-0.005	(0.39)	[653]	0.022	0.012	0.01	(0.218)	[279]
Pct. of attributes - positive response	0.033	0.037	-0.004	(0.27)	[2711]	0.07	0.077	-0.007	(0.514)	[653]	0.071	0.069	0.002	(0.898)	[279]
Pct. of attributes - neutral response	0.35	0.34	0.01	(0.305)	[2711]	0.706	0.714	-0.009	(0.656)	[653]	0.745	0.733	0.011	(0.699)	[279]
Pct. of attributes - negative response	0.013	0.013	0	(0.929)	[2711]	0.014	0.015	-0.002	(0.742)	[653]	0.016	0.007	0.009	(0.204)	[279]
Responded skeptically for any attribute	0.032	0.036	-0.004	(0.435)	[2711]	0.049	0.051	-0.002	(0.898)	[653]	0.054	0.029	0.025	(0.145)	[279]
Responded negatively for any attribute	0.035	0.03	0.004	(0.324)	[2711]	0.04	0.038	0.002	(0.876)	[653]	0.039	0.018	0.022	(0.109)	[279]
C. Black vs. Hispanic Testers															
No. of attributes brought up by landlord/broker	1.041	1.07	-0.029	(0.47)	[2711]	1.936	2.072	-0.136	(0.107)	[653]	1.961	1.989	-0.029	(0.813)	[279]
No. attributes - skeptical response	0.027	0.065	-0.038	(0)	[2711]	0.041	0.081	-0.04	(0.045)	[653]	0.039	0.039	0	(1)	[279]
No. attributes - positive response	0.125	0.14	-0.015	(0.327)	[2711]	0.256	0.27	-0.014	(0.749)	[653]	0.258	0.24	0.018	(0.772)	[279]
No. attributes - neutral response	0.894	0.884	0.01	(0.775)	[2711]	1.668	1.75	-0.083	(0.298)	[653]	1.685	1.728	-0.043	(0.703)	[279]
No. attributes - negative response	0.022	0.046	-0.024	(0)	[2711]	0.012	0.052	-0.04	(0.002)	[653]	0.018	0.022	-0.004	(0.782)	[279]
Pct. of attributes - skeptical response	0.008	0.016	-0.008	(0)	[2711]	0.015	0.023	-0.008	(0.196)	[653]	0.011	0.012	-0.001	(0.849)	[279]
Pct. of attributes - positive response	0.033	0.037	-0.004	(0.278)	[2711]	0.07	0.077	-0.008	(0.498)	[653]	0.066	0.069	-0.003	(0.831)	[279]
Pct. of attributes - neutral response	0.328	0.34	-0.013	(0.202)	[2711]	0.646	0.714	-0.069	(0.003)	[653]	0.666	0.733	-0.067	(0.036)	[279]
Pct. of attributes - negative response	0.006	0.013	-0.007	(0.001)	[2711]	0.003	0.015	-0.012	(0.001)	[653]	0.003	0.007	-0.005	(0.302)	[279]
Responded skeptically for any attribute	0.016	0.036	-0.02	(0)	[2711]	0.025	0.051	-0.026	(0.011)	[653]	0.018	0.029	-0.011	(0.367)	[279]
Responded negatively for any attribute	0.016	0.03	-0.014	(0)	[2711]	0.009	0.038	-0.029	(0)	[653]	0.011	0.018	-0.007	(0.415)	[279]

Table A12: Incidence of Early Stage Discrimination: Subjective Measures

E.7 Complier Average Causal Effects

	Subjects, by Treatment Received						
	Control		Monitoring		Punitive		
Assigned Arm	N	Percent	N	Percent	N	Percent	Row Totals
Control	279	1	0	0	0	0	279
Monitoring	31	0.18	143	0.82	0	0	174
Punitive	38	0.19	17	0.08	145	0.72	200

Table A13: Treatment Noncompliance Incidence. Cells contain the number of subjects by treatment assignment (row) and by treatment received (column). Row percents are displayed next to counts to show the extent of noncompliance by treatment assignment.

The CACE is defined as the ITT scaled by the proportion of Compliers[17] ITT_D, or $CACE = \frac{ITT}{ITT_D} = \frac{\mathbb{E}[Y_i(Z=T)]-\mathbb{E}[Y_i(Z=C)]}{\mathbb{E}[D_i(Z=T)]-\mathbb{E}[D_i(Z=C)]}$, where Y is the outcome, Z is the treatment assigned, and D is the treatment received (Gerber and Green 2012). We use an instrumental variables (IV) regression to estimate the CACEs, where treatment receipt is endogenous to treatment assignment Z. To do so, we first subset the data to include only those subjects assigned to the two treatment arms relevant for a given pairwise treatment-comparison difference. It is necessary to subset the data and use two-group estimators because CACEs are not identified in a principal stratification framework in trials with more than two arms and partial compliance (Imbens and Rubin 1997; Long et al. 2010). For each treatment-comparison difference of interest, we use the following system of equations to estimate the IV regression with inverse probability weights:

$$\begin{aligned} Y_{ib} &= \alpha + \tau D_{ib} + \gamma_b + \varepsilon_{ib} \\ D_{ib} &= \omega + \delta Z_{ib} + \gamma_b + \eta_{ib} \end{aligned} \tag{1}$$

where i indexes landlords, b indexes experimental blocks, Y is the outcome, D is the treatment received, Z is the treatment assigned, γ is a set of block fixed effects, τ is the CACE, and δ is the ITT_D.

By measuring which parts of treatment messages were successfully delivered to each landlord (see Table A13), we are able to observe treatment receipt for each subject. For the monitoring-control comparison, we employ a straightforward application of this estimation strategy for CACEs. In contrast, for the punitive-control comparison, we encounter one-sided noncompliance where some subjects assigned to the punitive condition instead receive the monitoring message. We estimate CACEs in two ways for the punitive-control comparison. First, we code subjects who were assigned to the punitive message but received the monitoring message as Non-compliers (i.e., $D_i = 0$). Second, we code subjects who were assigned to these same set of "partial compliers" as effectively receiving a punitive message from the city (despite not receiving the punitive appeal) and code their received treatment as a punitive message (i.e., $D_i = 2$). We interpret the two CACE estimates we calculate as an upper and lower bound (in magnitude) around the true CACE because each of these quantities scales the estimated ITT with a smaller and larger estimated proportion of Compliers, respectively.

[17] Compliers are defined as subjects who take up a treatment condition if and only if they are assigned to that condition.

Full tables summarizing CACE estimates are shown below. Across analyses, the estimated proportion of Compliers is high at 81% for the monitoring-control comparison and between 71.8% and 80% for the punitive-control comparison (for all first stage estimates, $p < 0.001$). As a result, the CACE estimates are generally consistent and qualitatively similar to the ITT estimates; the CACE point estimates are slightly larger than the estimated ITTs since the latter are divided by a proportion. We find that among Compliers, receiving the full punitive message when compared to the pure control decreases net discrimination against Hispanics (relative to whites) in receiving a callback by between 8.3 and 9.2 percentage points ($p = 0.056$) and decreases net discrimination against Hispanics (relative to blacks) in receiving a callback between 10.7 and 11.9 percentage points ($p = 0.02$).

Table A14 presents CACE estimates, or the average treatment effect among Compliers for the monitoring-control and punitive-control comparisons. These estimates tell us the effect of government messages on net discrimination among subjects who would comply with their assignment treatment for all possible treatment assignments. The CACE estimates are generally consistent with the ITT estimates since the estimated share of Compliers is relatively high across analyses.[18]

[18] Without a high proportion of Compliers, the CACE estimates would be less credible because treatment assignment would be a weak instrument for treatment receipt, which exacerbates bias.

Outcome	ITT	ITT_D	$CACE$	p-value
I. Monitoring vs. Control				
A. White vs. Black				
Index measure of favorable in-person interactions	-0.015	0.81	-0.018	0.388
Received post-visit callback	-0.002	0.81	-0.002	0.486
Received post-visit offer for unit	-0.003	0.81	-0.004	0.464
B. White vs. Hispanic				
Index measure of favorable in-person interactions	-0.061	0.81	-0.075	0.14
Received post-visit callback	-0.036	0.81	-0.045	0.202
Received post-visit offer for unit	-0.017	0.81	-0.021	0.31
C. Black vs. Hispanic				
Index measure of favorable in-person interactions	-0.089	0.81	-0.106	0.045
Received post-visit callback	-0.035	0.81	-0.043	0.206
Received post-visit offer for unit	-0.014	0.81	-0.017	0.328
II. Punitive vs. Control (Upper and Lower Bounds)				
A. White vs. Black				
Index measure of favorable in-person interactions				
Upper bound	0.007	0.718	0.01	0.555
Lower bound	0.007	0.8	0.009	0.555
Received post-visit callback				
Upper bound	0.019	0.718	0.027	0.675
Lower bound	0.019	0.8	0.024	0.676
Received post-visit offer for unit				
Upper bound	0.018	0.718	0.025	0.696
Lower bound	0.018	0.8	0.023	0.696
B. White vs. Hispanic				
Index measure of favorable in-person interactions				
Upper bound	0.048	0.718	0.067	0.806
Lower bound	0.048	0.8	0.061	0.806
Received post-visit callback				
Upper bound	-0.066	0.718	-0.092	0.056
Lower bound	-0.066	0.8	-0.083	0.056
Received post-visit offer for unit				
Upper bound	-0.021	0.718	-0.029	0.269
Lower bound	-0.021	0.8	-0.026	0.269
C. Black vs. Hispanic				
Index measure of favorable in-person interactions				
Upper bound	0.039	0.718	0.053	0.785
Lower bound	0.039	0.8	0.047	0.785
Received post-visit callback				
Upper bound	-0.085	0.718	-0.119	0.02
Lower bound	-0.085	0.8	-0.107	0.019
Received post-visit offer for unit				
Upper bound	-0.039	0.718	-0.055	0.123
Lower bound	-0.039	0.8	-0.049	0.122

Table A14: Estimated Complier Average Causal Effects of Messages on Net Discrimination Levels for Monitoring-Control and Punitive-Control Comparisons. Cells contain estimates of the ITT, ITT_D (proportion of Compliers), $CACE$, and p-values. ITT and ITT_D estimates are from OLS models with inverse probability weights and block fixed effects. $CACE$ estimates are from IV regression models with inverse probability weights and block fixed effects. Estimated p-values are reported in parentheses; p-values correspond to a one-sided test of the null hypothesis of equality of means for the monitoring-control and punitive-control comparisons, and to a two-sided test of the null hypothesis of equality of means for all analyses involving net discrimination against Hispanic (vs. black) testers. For the punitive-control comparison, the upper bound on the $CACE$ treats all subjects assigned to the punitive condition but who received the monitoring condition as subjects not complying with their treatment assignment. The lower bound on the CACE interprets the receipt of a monitoring message when assigned to punitive as effectively receiving a punitive message.

E.8 Details on Lasso Procedure to Select Covariates

We sought to improve the precision of our treatment effect estimates using covariate adjustment. To avoid biases that can arise from *ex post* covariate selection, we employed a lasso model selection procedure, a principled machine learning approach to model building[19] that is applied in this setting to select pre-treatment covariates highly prognostic of each outcome variable by treatment arm. Then following Yuan et al. (2012), we use the selected covariates to estimate the semiparametric covariate-adjusted ITT by predicting the mean regression-adjusted response by arm and calculating the difference in predicted means as the covariate-adjusted estimate of the ITT.

For each outcome variable and treatment arm, we estimate a 5-fold lasso regression with inverse probability weights and designate the predictors with non-zero coefficients for the model minimizing the tuning parameter λ as the covariates selected.[20] The set of variables from which we selected covariates include characteristics of the sampled housing unit, including the listed rent, the number of bedrooms, the listed square footage of the unit, and the borough of the unit; net discrimination in pre-treatment interactions between testers and subjects (i.e., landlords or brokers) occurring over the phone; the order in which testers were randomly assigned to respond to the housing advertisement; testers' biographical attributes, including their gender, whether they were partnered, and the relative ranking of their assumed incomes by race; subject characteristics including the modal tester perception of the subject's race, gender, and age, and whether the subject is a broker; tester fixed effects; block fixed effects; and the sampling frame associated with the unit.

Since the results of the lasso are sensitive to the specification of folds, we randomly shuffle the folds 1,000 times and re-estimate the lasso across fold specifications and treat the "selected" set of predictors as candidate vectors of covariates. Then, to select covariates for a given outcome and treatment arm, we regress the outcome variable on each candidate vector of covariates and select the vector that yields the highest adjusted R^2 with an F-test p-value $< .05$.[21] Following Yuan et al. (2012), we then use the selected covariates to estimate the semiparametric covariate-adjusted ITT by predicting the mean regression-adjusted response by arm and calculating the difference in predicted means as the covariate-adjusted estimate of the ITT. Variance estimation is conducted using the empirical sandwich variance estimator following Yuan et al. (2012, Equation 13), as well as using 95% Studentized, basic, and percentile bootstrap confidence intervals.

E.9 Covariate Adjusted Analyses

[19] See Bloniarz et al. (2016).

[20] We use the `cv.glmnet` function in the `glmnet` R library (Friedman et al. 2010) to perform the lasso.

[21] By doing so, this effectively undoes the shrinkage but this is not substantively concerning since the purpose of this exercise is variable selection for model building.

	Unadjusted			Covariate adjusted			Covariate adjusted - bootstrap				
Outcome Measure	Estimate	SE	95% CI	Estimate	SE	95% CI	Mean	SE	95% Studentized CI	95% Basic CI	95% Percentile CI
I. Monitoring vs. Control											
A. White vs. Black											
Index measure of favorable in-person interactions	-0.015	0.052	[-0.117, 0.087]	-0.008	(0.045)	[-0.096,0.079]	-0.008	(0.055)	[-0.143,0.127]	[-0.116,0.098]	[-0.114,0.099]
Received post-visit callback	-0.002	0.045	[-0.091, 0.088]	-0.018	(0.042)	[-0.1,0.065]	-0.021	(0.049)	[-0.128,0.101]	[-0.109,0.083]	[-0.118,0.073]
Received post-visit offer	-0.003	0.036	[-0.075, 0.068]	-0.015	(0.033)	[-0.079,0.049]	-0.015	(0.035)	[-0.089,0.058]	[-0.083,0.052]	[-0.082,0.053]
B. White vs. Hispanic											
Index measure of favorable in-person interactions	-0.061	0.057	[-0.172, 0.05]	-0.098	(0.051)	[-0.198,0.003]	-0.089	(0.059)	[-0.254,0.026]	[-0.221,0.011]	[-0.207,0.026]
Received post-visit callback	-0.036	0.043	[-0.121, 0.049]	-0.034	(0.042)	[-0.116,0.048]	-0.032	(0.046)	[-0.136,0.068]	[-0.124,0.058]	[-0.125,0.057]
Received post-visit offer	-0.017	0.034	[-0.084, 0.05]	-0.021	(0.032)	[-0.084,0.042]	-0.021	(0.035)	[-0.102,0.058]	[-0.092,0.049]	[-0.09,0.05]
C. Black vs. Hispanic											
Index measure of favorable in-person interactions	-0.089	0.052	[-0.192, 0.014]	-0.095	(0.048)	[-0.189,0]	-0.087	(0.056)	[-0.237,0.022]	[-0.216,0.006]	[-0.196,0.026]
Received post-visit callback	-0.035	0.042	[-0.118, 0.048]	0.002	(0.04)	[-0.076,0.08]	0.005	(0.046)	[-0.11,0.106]	[-0.094,0.089]	[-0.085,0.098]
Received post-visit offer	-0.014	0.031	[-0.074, 0.047]	0.004	(0.029)	[-0.052,0.061]	0.005	(0.031)	[-0.063,0.071]	[-0.057,0.064]	[-0.055,0.066]
II. Punitive vs. Control											
A. White vs. Black											
Index measure of favorable in-person interactions	0.007	0.053	[-0.097, 0.112]	-0.046	(0.047)	[-0.139,0.047]	-0.045	(0.059)	[-0.2,0.098]	[-0.163,0.07]	[-0.162,0.071]
Received post-visit callback	0.019	0.042	[-0.064, 0.103]	0.030	(0.039)	[-0.046,0.107]	0.027	(0.042)	[-0.055,0.129]	[-0.047,0.119]	[-0.058,0.108]
Received post-visit offer	0.018	0.035	[-0.051, 0.087]	0.023	(0.03)	[-0.036,0.083]	0.025	(0.034)	[-0.054,0.101]	[-0.044,0.089]	[-0.043,0.091]
B. White vs. Hispanic											
Index measure of favorable in-person interactions	0.048	0.055	[-0.061, 0.157]	0.023	(0.049)	[-0.073,0.12]	0.022	(0.056)	[-0.107,0.158]	[-0.086,0.137]	[-0.09,0.132]
Received post-visit callback	-0.066	0.041	[-0.147, 0.015]	-0.051	(0.037)	[-0.123,0.021]	-0.055	(0.043)	[-0.145,0.059]	[-0.128,0.04]	[-0.142,0.027]
Received post-visit offer	-0.021	0.034	[-0.089, 0.046]	0.000	(0.031)	[-0.061,0.061]	0.001	(0.035)	[-0.078,0.077]	[-0.069,0.068]	[-0.068,0.069]
C. Black vs. Hispanic											
Index measure of favorable in-person interactions	0.039	0.049	[-0.057, 0.134]	0.037	(0.045)	[-0.052,0.125]	0.036	(0.049)	[-0.072,0.147]	[-0.058,0.137]	[-0.064,0.131]
Received post-visit callback	-0.085	0.041	[-0.165, -0.005]	-0.098	(0.038)	[-0.173,-0.023]	-0.096	(0.043)	[-0.198,-0.002]	[-0.184,-0.015]	[-0.182,-0.013]
Received post-visit offer	-0.039	0.033	[-0.105, 0.027]	-0.032	(0.031)	[-0.093,0.03]	-0.033	(0.035)	[-0.109,0.046]	[-0.098,0.036]	[-0.1,0.035]
III. Punitive vs. Monitoring											
A. White vs. Black											
Index measure of favorable in-person interactions	0.018	0.058	[-0.095, 0.132]	-0.023	(0.057)	[-0.136,0.091]	-0.019	(0.057)	[-0.14,0.087]	[-0.138,0.087]	[-0.133,0.093]
Received post-visit callback	0.033	0.049	[-0.064, 0.129]	0.038	(0.045)	[-0.051,0.127]	0.035	(0.049)	[-0.07,0.148]	[-0.059,0.137]	[-0.061,0.136]
Received post-visit offer	0.026	0.037	[-0.048, 0.099]	0.029	(0.034)	[-0.037,0.095]	0.028	(0.035)	[-0.043,0.105]	[-0.038,0.099]	[-0.04,0.097]
B. White vs. Hispanic											
Index measure of favorable in-person interactions	0.107	0.063	[-0.017, 0.231]	0.082	(0.061)	[-0.038,0.201]	0.082	(0.06)	[-0.039,0.203]	[-0.04,0.2]	[-0.037,0.203]
Received post-visit callback	-0.019	0.049	[-0.116, 0.078]	-0.061	(0.049)	[-0.157,0.036]	-0.060	(0.05)	[-0.166,0.042]	[-0.158,0.037]	[-0.158,0.037]
Received post-visit offer	0.002	0.038	[-0.073, 0.077]	-0.014	(0.035)	[-0.084,0.055]	-0.013	(0.037)	[-0.096,0.064]	[-0.087,0.058]	[-0.086,0.059]
C. Black vs. Hispanic											
Index measure of favorable in-person interactions	0.139	0.057	[0.026, 0.251]	0.141	(0.06)	[0.023,0.26]	0.139	(0.056)	[0.037,0.251]	[0.03,0.253]	[0.03,0.253]
Received post-visit callback	-0.052	0.047	[-0.144, 0.04]	-0.080	(0.043)	[-0.164,0.005]	-0.077	(0.046)	[-0.182,0.022]	[-0.171,0.011]	[-0.17,0.012]
Received post-visit offer	-0.024	0.036	[-0.094, 0.046]	-0.039	(0.034)	[-0.105,0.027]	-0.038	(0.036)	[-0.118,0.035]	[-0.109,0.032]	[-0.11,0.031]

Table A15: Covariate adjusted ITT estimates. The left panel presents the main estimates with block fixed effects and inverse probability weights. The middle panel presents covariate adjusted estimates with inverse probability weights; the uncertainty estimates are based on the empirical sandwich variance estimator by Yuan et al. (2012). The right panel presents bootstrapped covariate adjusted estimates with 95% Studentized, basic, and percentile confidence intervals.

E.10 Missingness Analyses

Comparison	Estimate	SE	t	p-value	F-statistic	F-test p-value
A. White vs. Black						
Monitoring vs. Control	0.073	0.042	1.729	0.085	1.481	0.097
Punitive vs. Control	0.036	0.04	0.897	0.37	1.263	0.212
Punitive vs. Monitoring	-0.021	0.048	-0.436	0.663	0.936	0.532
B. White vs. Hispanic						
Monitoring vs. Control	0.019	0.044	0.428	0.669	1.081	0.369
Punitive vs. Control	-0.023	0.04	-0.573	0.567	2.187	0.004
Punitive vs. Monitoring	-0.045	0.047	-0.944	0.346	1.293	0.194
C. Black vs. Hispanic						
Monitoring vs. Control	0.083	0.043	1.913	0.056	1.651	0.049
Punitive vs. Control	0.038	0.042	0.903	0.367	1.247	0.224
Punitive vs. Monitoring	-0.042	0.05	-0.854	0.394	1.091	0.36

Table A16: The estimated correlation between treatment assignment and missingness on the subjective index measure of net discrimination in interactions during appointments, estimated from OLS models regressing missingness on treatment assignment and block fixed effects with inverse probability weighting. The F-statistic and F-test p-value tests the null hypothesis that all coefficients equal zero.

	White vs. Black		White vs. Hispanic		Black vs. Hispanic	
	Callbacks	Offers	Callbacks	Offers	Callbacks	Offers
Missing Index Measure, W-B	-0.05	-0.06	-0.08	-0.11	-0.02	-0.05
Missing Index Measure, W-H	-0.11	-0.05	-0.02	0	0.09	0.06
Missing Index Measure, B-H	0	0	0.01	0	0.02	0

Table A17: Pairwise correlations between missing subjective net discrimination index measures and objective net discrimination measures.

Variable	Estimate	SE	t	p-value
(Intercept)	0.143	0.014	10.267	0
Hispanic tester	-0.006	0.02	-0.324	0.746
White tester	-0.008	0.02	-0.387	0.699
F-statistic:	0.086			
F test p-value:	0.917			

Table A18: The table presents OLS estimates from a regression predicting missingness on subjective indicators of favorable treatment during appointments as a function of tester race. The reference racial group is black. The F-test results reported at the bottom of the table are for a test of the null hypothesis that all coefficients equal zero.

E.11 Balance Tables

	Control						Monitoring						Punitive					
	Unweighted			Weighted			Unweighted			Weighted			Unweighted			Weighted		
	Pct	SE	N	Pct	SE	N	Pct	SE	N	Pct	SE	N	Pct	SE	N	Pct	SE	N
Frame																		
Likely discrimination	0.068	0.015	19	0.07	0.015	46	0.069	0.019	12	0.066	0.019	40	0.065	0.017	13	0.063	0.017	43
Representative	0.932	0.015	260	0.93	0.015	610	0.931	0.019	162	0.934	0.019	564	0.935	0.017	187	0.937	0.017	641
Household size																		
1	0.774	0.025	216	0.771	0.025	506	0.816	0.029	142	0.823	0.029	497	0.795	0.029	159	0.794	0.029	543
2	0.226	0.025	63	0.229	0.025	150	0.184	0.029	32	0.177	0.029	107	0.205	0.029	41	0.206	0.029	141
Tester team gender																		
Female	0.505	0.03	141	0.514	0.03	337	0.54	0.038	94	0.545	0.038	329	0.525	0.035	105	0.525	0.035	359
Male	0.495	0.03	138	0.486	0.03	319	0.46	0.038	80	0.455	0.038	275	0.475	0.035	95	0.475	0.035	325

Table A19: Balance Table for Categorical Covariates. Cells contain proportions, standard errors, and counts.

Table A20: Balance Table for Continuous Covariates. Cells contain means and standard deviations in parentheses.

	Control		Monitoring		Punitive	
Variable	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
	I. Apartment Characteristics					
Advertised number of bedrooms	0.925	0.928	0.764	0.755	0.93	0.928
	(1.214)	(1.217)	(1.126)	(1.108)	(1.201)	(1.216)
Advertised monthly rental price	2442.696	2439.609	2409.632	2378.686	2395.57	2392.987
	(1281.253)	(1235.87)	(1239.418)	(1223.697)	(1115.931)	(1113.922)
Advertised square footage	1116.677	1124.844	836.364	856	960.526	992.508
	(430.927)	(444.738)	(331.52)	(324.802)	(508.235)	(516.829)
Borough: Bronx	0.115	0.114	0.103	0.108	0.105	0.108
	(0.319)	(0.318)	(0.305)	(0.31)	(0.307)	(0.311)
Borough: Brooklyn	0.358	0.351	0.379	0.373	0.34	0.345
	(0.48)	(0.478)	(0.487)	(0.484)	(0.475)	(0.476)
Borough: Manhattan	0.341	0.352	0.368	0.371	0.345	0.338
	(0.475)	(0.478)	(0.484)	(0.483)	(0.477)	(0.473)
Borough: Queens	0.14	0.136	0.115	0.116	0.18	0.178
	(0.347)	(0.343)	(0.32)	(0.32)	(0.385)	(0.383)
Borough: Staten Island	0.047	0.047	0.034	0.033	0.03	0.031
	(0.211)	(0.212)	(0.183)	(0.179)	(0.171)	(0.173)
	II. Randomization Regime					
Regime 1	0.151	0.192	0.218	0.189	0.26	0.228
	(0.358)	(0.394)	(0.414)	(0.392)	(0.44)	(0.42)

(continued)

	Control		Monitoring		Punitive	
Variable	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Regime 2	0.649	0.552	0.471	0.543	0.42	0.491
	(0.478)	(0.498)	(0.501)	(0.499)	(0.495)	(0.5)
Regime 3	0.201	0.256	0.31	0.268	0.32	0.281
	(0.401)	(0.437)	(0.464)	(0.443)	(0.468)	(0.45)
III. Early Stage Discrimination						
Net Diff. in Num. Attributes Inquired About over Phone: Black-Hispanic	-0.029	-0.07	-0.305	-0.272	-0.14	-0.099
	(2.018)	(2.089)	(2.218)	(2.218)	(2.299)	(2.266)
Net Diff. in Num. Attributes Inquired About over Phone: White-Hispanic	0.161	0.102	-0.149	-0.124	-0.255	-0.235
	(2.051)	(2.074)	(2.204)	(2.175)	(2.018)	(2.008)
Net Diff. in Num. Attributes Inquired About over Phone: White-Black	0.19	0.172	0.155	0.147	-0.115	-0.136
	(2.17)	(2.17)	(2.362)	(2.322)	(2.446)	(2.414)
Net Diff. in Num. Attributes Eliciting Skeptical Reaction over Phone: Black-Hispanic	0	0.008	-0.063	-0.063	-0.075	-0.067
	(0.416)	(0.447)	(0.506)	(0.518)	(0.609)	(0.579)
Net Diff. in Num. Attributes Eliciting Skeptical Reaction over Phone: White-Hispanic	0.057	0.047	-0.04	-0.036	-0.06	-0.044
	(0.533)	(0.515)	(0.448)	(0.452)	(0.639)	(0.63)
Net Diff. in Num. Attributes Eliciting Skeptical Reaction over Phone: White-Black	0.057	0.04	0.023	0.026	0.015	0.023
	(0.591)	(0.604)	(0.416)	(0.438)	(0.431)	(0.442)
Net Diff. in Pct. of Attributes Raised Eliciting Skeptical Reaction over Phone: Black-Hispanic	-0.001	-0.001	0.001	0.003	-0.025	-0.025
	(0.119)	(0.121)	(0.202)	(0.21)	(0.161)	(0.161)
Net Diff. in Pct. of Attributes Raised Eliciting Skeptical Reaction over Phone: White-Hispanic	0.01	0.007	-0.015	-0.014	-0.018	-0.016
	(0.139)	(0.135)	(0.128)	(0.127)	(0.179)	(0.179)
Net Diff. in Pct. of Attributes Raised Eliciting Skeptical Reaction over Phone: White-Black	0.012	0.008	-0.016	-0.017	0.007	0.008
	(0.143)	(0.142)	(0.177)	(0.186)	(0.111)	(0.11)
Net Diff. in Num. Attributes Eliciting Positive Reaction over Phone: Black-Hispanic	0.018	0.04	-0.006	-0.018	-0.065	-0.063
	(1.03)	(1.103)	(1.023)	(0.978)	(1.252)	(1.186)
Net Diff. in Num. Attributes Eliciting Positive Reaction over Phone: White-Hispanic	0.029	0.02	0.017	0.035	-0.08	-0.06
	(0.768)	(0.774)	(0.909)	(0.907)	(1.118)	(1.073)
Net Diff. in Num. Attributes Eliciting Positive Reaction over Phone: White-Black	0.011	-0.02	0.023	0.053	-0.015	0.003
	(0.961)	(1.03)	(0.918)	(0.905)	(1.123)	(1.086)
Net Diff. in Num. Attributes Eliciting Neutral Reaction over Phone: Black-Hispanic	-0.043	-0.11	-0.241	-0.192	0	0.037
	(1.883)	(1.936)	(2.126)	(2.123)	(2.136)	(2.11)
Net Diff. in Num. Attributes Eliciting Neutral Reaction over Phone: White-Hispanic	0.093	0.044	-0.132	-0.124	-0.17	-0.175
	(1.783)	(1.791)	(2.131)	(2.094)	(1.854)	(1.821)
Net Diff. in Num. Attributes Eliciting Neutral Reaction over Phone: White-Black	0.136	0.154	0.109	0.068	-0.17	-0.212
	(2.013)	(2.013)	(2.231)	(2.212)	(2.233)	(2.2)
Net Diff. in Num. Attributes Eliciting Negative Reaction over Phone: Black-Hispanic	-0.004	0	-0.057	-0.061	-0.075	-0.073
	(0.216)	(0.228)	(0.368)	(0.37)	(0.387)	(0.372)
Net Diff. in Num. Attributes Eliciting Negative Reaction over Phone: White-Hispanic	0.039	0.038	-0.034	-0.035	-0.005	0
	(0.363)	(0.362)	(0.386)	(0.391)	(0.496)	(0.499)
Net Diff. in Num. Attributes Eliciting Negative Reaction over Phone: White-Black	0.043	0.038	0.023	0.026	0.07	0.073
	(0.386)	(0.402)	(0.284)	(0.298)	(0.382)	(0.395)
Net Diff. in Pct. of Attributes Raised Eliciting Positive Reaction over Phone: Black-Hispanic	-0.003	0.005	0.02	0.014	-0.038	-0.034
	(0.268)	(0.283)	(0.325)	(0.311)	(0.279)	(0.272)

(continued)

	Control		Monitoring		Punitive	
Variable	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Net Diff. in Pct. of Attributes Raised Eliciting Positive Reaction over Phone: White-Hispanic	0.002	0.001	0.009	0.011	-0.032	-0.027
	(0.226)	(0.226)	(0.266)	(0.263)	(0.296)	(0.284)
Net Diff. in Pct. of Attributes Raised Eliciting Positive Reaction over Phone: White-Black	0.005	-0.003	-0.012	-0.004	0.006	0.007
	(0.263)	(0.282)	(0.292)	(0.281)	(0.251)	(0.247)
Net Diff. in Pct. of Attributes Raised Eliciting Neutral Reaction over Phone: Black-Hispanic	-0.067	-0.077	-0.099	-0.077	-0.044	-0.046
	(0.533)	(0.546)	(0.654)	(0.641)	(0.611)	(0.596)
Net Diff. in Pct. of Attributes Raised Eliciting Neutral Reaction over Phone: White-Hispanic	0.011	0.011	-0.048	-0.045	-0.002	-0.015
	(0.483)	(0.487)	(0.546)	(0.547)	(0.466)	(0.457)
Net Diff. in Pct. of Attributes Raised Eliciting Neutral Reaction over Phone: White-Black	0.078	0.089	0.05	0.033	0.043	0.031
	(0.592)	(0.603)	(0.616)	(0.61)	(0.615)	(0.611)
Net Diff. in Pct. of Attributes Raised Eliciting Negative Reaction over Phone: Black-Hispanic	-0.005	-0.004	-0.013	-0.015	-0.023	-0.024
	(0.073)	(0.071)	(0.106)	(0.107)	(0.123)	(0.128)
Net Diff. in Pct. of Attributes Raised Eliciting Negative Reaction over Phone: White-Hispanic	0.009	0.009	-0.012	-0.013	-0.007	-0.008
	(0.113)	(0.109)	(0.093)	(0.094)	(0.151)	(0.154)
Net Diff. in Pct. of Attributes Raised Eliciting Negative Reaction over Phone: White-Black	0.013	0.012	0.001	0.002	0.016	0.016
	(0.098)	(0.099)	(0.062)	(0.062)	(0.097)	(0.094)
Net Diff. in Receiving Any Skeptical Reaction to Attributes over Phone: Black-Hispanic	-0.011	-0.011	-0.034	-0.033	-0.04	-0.038
	(0.199)	(0.21)	(0.338)	(0.344)	(0.262)	(0.257)
Net Diff. in Receiving Any Skeptical Reaction to Attributes over Phone: White-Hispanic	0.025	0.017	-0.023	-0.02	-0.02	-0.013
	(0.287)	(0.284)	(0.322)	(0.325)	(0.316)	(0.322)
Net Diff. in Receiving Any Skeptical Reaction to Attributes over Phone: White-Black	0.036	0.027	0.011	0.013	0.02	0.025
	(0.266)	(0.264)	(0.304)	(0.315)	(0.223)	(0.231)
Net Diff. in Receiving Any Negative Reaction to Attributes over Phone: Black-Hispanic	-0.007	-0.006	-0.034	-0.038	-0.055	-0.056
	(0.147)	(0.156)	(0.238)	(0.245)	(0.25)	(0.248)
Net Diff. in Receiving Any Negative Reaction to Attributes over Phone: White-Hispanic	0.022	0.02	-0.023	-0.025	-0.005	-0.004
	(0.223)	(0.224)	(0.214)	(0.218)	(0.309)	(0.308)
Net Diff. in Receiving Any Negative Reaction to Attributes over Phone: White-Black	0.029	0.026	0.011	0.013	0.05	0.051
	(0.223)	(0.23)	(0.186)	(0.191)	(0.24)	(0.24)
IV. Subject (Landlord/Broker) Characteristics						
Subject is a Broker	0.86	0.861	0.816	0.82	0.85	0.849
	(0.347)	(0.346)	(0.389)	(0.385)	(0.358)	(0.358)
Modal Perception of Landlord Race among Testers: Asian	0.111	0.111	0.08	0.079	0.085	0.086
	(0.315)	(0.315)	(0.273)	(0.271)	(0.28)	(0.281)
Modal Perception of Landlord Race among Testers: Black	0.122	0.114	0.109	0.109	0.135	0.135
	(0.328)	(0.318)	(0.313)	(0.312)	(0.343)	(0.341)
Modal Perception of Landlord Race among Testers: Hispanic/Latino	0.158	0.159	0.149	0.152	0.12	0.115
	(0.365)	(0.366)	(0.358)	(0.36)	(0.326)	(0.32)
Modal Perception of Landlord Race among Testers: White	0.53	0.537	0.534	0.53	0.56	0.56
	(0.5)	(0.499)	(0.5)	(0.5)	(0.498)	(0.497)
Modal Perception of Landlord Age among Testers: 18 to 34	0.437	0.441	0.5	0.498	0.515	0.512
	(0.497)	(0.497)	(0.501)	(0.5)	(0.501)	(0.5)
Modal Perception of Landlord Age among Testers: 35 to 44	0.262	0.265	0.241	0.243	0.28	0.285
	(0.44)	(0.442)	(0.429)	(0.429)	(0.45)	(0.452)

(continued)

	Control		Monitoring		Punitive	
Variable	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Modal Perception of Landlord Age among Testers: 45 to 64	0.211	0.207	0.144	0.137	0.12	0.117
	(0.409)	(0.406)	(0.352)	(0.345)	(0.326)	(0.322)
Modal Perception of Landlord Age among Testers: 65 and up	0.011	0.011	0.017	0.017	0.015	0.015
	(0.103)	(0.103)	(0.131)	(0.128)	(0.122)	(0.12)
Modal Perception of Landlord Age among Testers: Unknown/No Consensus	0.079	0.076	0.098	0.104	0.07	0.072
	(0.27)	(0.266)	(0.298)	(0.306)	(0.256)	(0.258)
	V. Tester Call Order					
Randomized Tester Call Order: White before Black	0.47	0.48	0.506	0.52	0.53	0.537
	(0.5)	(0.5)	(0.501)	(0.5)	(0.5)	(0.499)
Randomized Tester Call Order: White before Hispanic	0.434	0.434	0.489	0.493	0.555	0.554
	(0.496)	(0.496)	(0.501)	(0.5)	(0.498)	(0.497)
Randomized Tester Call Order: Black before Hispanic	0.541	0.544	0.477	0.487	0.48	0.48
	(0.499)	(0.498)	(0.501)	(0.5)	(0.501)	(0.5)
	VI. Testers' Assumed Income					
Assumed Tester Incomes: White > Black	0.466	0.468	0.431	0.442	0.45	0.455
	(0.5)	(0.499)	(0.497)	(0.497)	(0.499)	(0.498)
Assumed Tester Incomes: White > Hispanic	0.455	0.463	0.506	0.508	0.46	0.461
	(0.499)	(0.499)	(0.501)	(0.5)	(0.5)	(0.499)
Assumed Tester Incomes: Black > Hispanic	0.441	0.445	0.523	0.525	0.47	0.472
	(0.497)	(0.497)	(0.501)	(0.5)	(0.5)	(0.5)
Assumed Tester Incomes: White = Black	0.079	0.081	0.098	0.101	0.075	0.073
	(0.27)	(0.273)	(0.298)	(0.302)	(0.264)	(0.26)
Assumed Tester Incomes: White = Hispanic	0.082	0.085	0.069	0.071	0.07	0.073
	(0.276)	(0.28)	(0.254)	(0.257)	(0.256)	(0.26)
Assumed Tester Incomes: Black = Hispanic	0.154	0.157	0.075	0.073	0.08	0.08
	(0.362)	(0.364)	(0.264)	(0.26)	(0.272)	(0.272)
Assumed Tester Incomes: White < Black	0.455	0.451	0.471	0.457	0.475	0.472
	(0.499)	(0.498)	(0.501)	(0.499)	(0.501)	(0.5)
Assumed Tester Incomes: White < Hispanic	0.462	0.451	0.425	0.421	0.47	0.466
	(0.499)	(0.498)	(0.496)	(0.494)	(0.5)	(0.499)
Assumed Tester Incomes: Black < Hispanic	0.405	0.398	0.402	0.402	0.45	0.447
	(0.492)	(0.49)	(0.492)	(0.491)	(0.499)	(0.498)
Assumed Tester Incomes: White Highest	0.384	0.393	0.385	0.396	0.355	0.354
	(0.487)	(0.489)	(0.488)	(0.489)	(0.48)	(0.478)
Assumed Tester Incomes: Black Highest	0.394	0.39	0.408	0.401	0.39	0.392
	(0.49)	(0.488)	(0.493)	(0.49)	(0.489)	(0.489)
Assumed Tester Incomes: Hispanic Highest	0.38	0.373	0.333	0.329	0.37	0.367
	(0.486)	(0.484)	(0.473)	(0.47)	(0.484)	(0.482)
	VII. Tester Fixed Effects					
Tester ID A01	0.108	0.11	0.121	0.126	0.11	0.111
	(0.31)	(0.313)	(0.327)	(0.332)	(0.314)	(0.314)
Tester ID A10	0.136	0.122	0.08	0.084	0.11	0.111
	(0.344)	(0.327)	(0.273)	(0.278)	(0.314)	(0.314)
Tester ID A11	0.029	0.024	0.023	0.026	0.035	0.039

(continued)

	Control		Monitoring		Punitive	
Variable	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
	(0.167)	(0.154)	(0.15)	(0.161)	(0.184)	(0.195)
Tester ID A13	0.154	0.168	0.144	0.141	0.17	0.156
	(0.362)	(0.374)	(0.352)	(0.348)	(0.377)	(0.364)
Tester ID A02	0	0	0	0	0.01	0.009
	(0)	(0)	(0)	(0)	(0.1)	(0.093)
Tester ID A21	0.007	0.009	0.011	0.01	0.005	0.004
	(0.085)	(0.095)	(0.107)	(0.099)	(0.071)	(0.066)
Tester ID A22	0.043	0.055	0.121	0.104	0.055	0.048
	(0.203)	(0.228)	(0.327)	(0.306)	(0.229)	(0.214)
Tester ID A03	0.057	0.061	0.069	0.063	0.085	0.085
	(0.233)	(0.239)	(0.254)	(0.243)	(0.28)	(0.279)
Tester ID A04	0.086	0.081	0.069	0.073	0.1	0.111
	(0.281)	(0.273)	(0.254)	(0.26)	(0.301)	(0.314)
Tester ID A05	0.14	0.128	0.155	0.164	0.11	0.12
	(0.347)	(0.334)	(0.363)	(0.37)	(0.314)	(0.325)
Tester ID A06	0.061	0.059	0.052	0.051	0.06	0.058
	(0.24)	(0.237)	(0.222)	(0.221)	(0.238)	(0.235)
Tester ID A07	0.004	0.005	0.011	0.01	0.005	0.004
	(0.06)	(0.068)	(0.107)	(0.099)	(0.071)	(0.066)
Tester ID A08	0.168	0.169	0.138	0.142	0.14	0.137
	(0.375)	(0.375)	(0.346)	(0.35)	(0.348)	(0.345)
Tester ID A09	0.007	0.009	0.006	0.005	0.005	0.004
	(0.085)	(0.095)	(0.076)	(0.07)	(0.071)	(0.066)
Tester ID B01	0.204	0.189	0.27	0.291	0.19	0.203
	(0.404)	(0.392)	(0.445)	(0.455)	(0.393)	(0.403)
Tester ID B11	0.061	0.052	0.029	0.031	0.055	0.063
	(0.24)	(0.222)	(0.168)	(0.175)	(0.229)	(0.243)
Tester ID B12	0.011	0.009	0.006	0.007	0	0
	(0.103)	(0.095)	(0.076)	(0.081)	(0)	(0)
Tester ID B14	0.108	0.116	0.115	0.106	0.115	0.115
	(0.31)	(0.32)	(0.32)	(0.308)	(0.32)	(0.32)
Tester ID B16	0.032	0.027	0.034	0.04	0.035	0.035
	(0.177)	(0.163)	(0.183)	(0.195)	(0.184)	(0.184)
Tester ID B02	0.025	0.032	0.04	0.035	0.04	0.035
	(0.157)	(0.176)	(0.197)	(0.183)	(0.196)	(0.184)
Tester ID B20	0	0	0	0	0.005	0.004
	(0)	(0)	(0)	(0)	(0.071)	(0.066)
Tester ID B23	0.014	0.018	0.011	0.01	0.03	0.026
	(0.119)	(0.134)	(0.107)	(0.099)	(0.171)	(0.16)
Tester ID B24	0.022	0.027	0.063	0.055	0.04	0.035
	(0.145)	(0.163)	(0.244)	(0.227)	(0.196)	(0.184)
Tester ID B25	0.039	0.05	0.057	0.05	0.08	0.07
	(0.195)	(0.219)	(0.233)	(0.217)	(0.272)	(0.256)

(continued)

	Control		Monitoring		Punitive	
Variable	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Tester ID B27	0.022	0.027	0.046	0.04	0.02	0.018
	(0.145)	(0.163)	(0.21)	(0.195)	(0.14)	(0.131)
Tester ID B03	0.104	0.117	0.086	0.083	0.095	0.091
	(0.306)	(0.322)	(0.281)	(0.276)	(0.294)	(0.287)
Tester ID B04	0.022	0.027	0.034	0.03	0.045	0.039
	(0.145)	(0.163)	(0.183)	(0.17)	(0.208)	(0.195)
Tester ID B06	0.043	0.044	0.04	0.04	0.065	0.064
	(0.203)	(0.206)	(0.197)	(0.195)	(0.247)	(0.246)
Tester ID B07	0.018	0.023	0.023	0.02	0.035	0.031
	(0.133)	(0.15)	(0.15)	(0.14)	(0.184)	(0.173)
Tester ID B08	0.233	0.203	0.126	0.144	0.13	0.146
	(0.423)	(0.402)	(0.333)	(0.351)	(0.337)	(0.354)
Tester ID B09	0.043	0.037	0.017	0.02	0.02	0.023
	(0.203)	(0.188)	(0.131)	(0.14)	(0.14)	(0.151)
Tester ID C01	0.025	0.032	0.029	0.025	0.015	0.013
	(0.157)	(0.176)	(0.168)	(0.156)	(0.122)	(0.114)
Tester ID C10	0.022	0.023	0.046	0.041	0.035	0.034
	(0.145)	(0.15)	(0.21)	(0.199)	(0.184)	(0.18)
Tester ID C12	0.004	0.003	0.011	0.013	0.015	0.015
	(0.06)	(0.055)	(0.107)	(0.114)	(0.122)	(0.12)
Tester ID C13	0.097	0.082	0.063	0.073	0.045	0.053
	(0.296)	(0.275)	(0.244)	(0.26)	(0.208)	(0.223)
Tester ID C14	0.007	0.009	0.057	0.05	0.015	0.013
	(0.085)	(0.095)	(0.233)	(0.217)	(0.122)	(0.114)
Tester ID C15	0.018	0.023	0.052	0.045	0.045	0.039
	(0.133)	(0.15)	(0.222)	(0.207)	(0.208)	(0.195)
Tester ID C02	0.28	0.276	0.207	0.21	0.245	0.249
	(0.45)	(0.447)	(0.406)	(0.408)	(0.431)	(0.432)
Tester ID C27	0.039	0.05	0.023	0.02	0.07	0.061
	(0.195)	(0.219)	(0.15)	(0.14)	(0.256)	(0.24)
Tester ID C29	0.05	0.064	0.069	0.06	0.05	0.044
	(0.219)	(0.245)	(0.254)	(0.237)	(0.218)	(0.205)
Tester ID C03	0.004	0.005	0	0	0.03	0.026
	(0.06)	(0.068)	(0)	(0)	(0.171)	(0.16)
Tester ID C31	0.004	0.005	0	0	0	0
	(0.06)	(0.068)	(0)	(0)	(0)	(0)
Tester ID C33	0.011	0.009	0.006	0.007	0.005	0.006
	(0.103)	(0.095)	(0.076)	(0.081)	(0.071)	(0.076)
Tester ID C04	0.179	0.171	0.126	0.131	0.16	0.161
	(0.384)	(0.377)	(0.333)	(0.337)	(0.368)	(0.368)
Tester ID C05	0.011	0.014	0.029	0.025	0.04	0.035
	(0.103)	(0.116)	(0.168)	(0.156)	(0.196)	(0.184)

(continued)

	Control		Monitoring		Punitive	
Variable	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Tester ID C06	0.004	0.005	0.006	0.005	0.005	0.004
	(0.06)	(0.068)	(0.076)	(0.07)	(0.071)	(0.066)
Tester ID C07	0.244	0.226	0.253	0.276	0.2	0.225
	(0.43)	(0.418)	(0.436)	(0.448)	(0.401)	(0.418)
Tester ID C08	0	0	0.006	0.005	0	0
	(0)	(0)	(0.076)	(0.07)	(0)	(0)
Tester ID C09	0.004	0.005	0.017	0.015	0.025	0.022
	(0.06)	(0.068)	(0.131)	(0.121)	(0.157)	(0.147)

E.12 ITT Estimates among Subsample Excluding Likely Discrimination Cases

Outcome	Estimate	SE	t	p-value	95% CI
I. Monitoring vs. Control					
A. White vs. Black					
Index measure of favorable in-person interactions	-0.011	0.054	-0.201	(0.42)	[-0.117, 0.095]
Received post-visit callback	0.015	0.047	0.315	(0.624)	[-0.077, 0.107]
Received post-visit offer for unit	0.008	0.038	0.2	(0.579)	[-0.066, 0.082]
B. White vs. Hispanic					
Index measure of favorable in-person interactions	-0.047	0.059	-0.794	(0.214)	[-0.163, 0.069]
Received post-visit callback	-0.022	0.045	-0.501	(0.308)	[-0.11, 0.065]
Received post-visit offer for unit	-0.009	0.035	-0.252	(0.401)	[-0.079, 0.061]
C. Black vs. Hispanic					
Index measure of favorable in-person interactions	-0.075	0.054	-1.389	(0.166)	[-0.182, 0.032]
Received post-visit callback	-0.037	0.043	-0.856	(0.392)	[-0.123, 0.048]
Received post-visit offer for unit	-0.016	0.032	-0.509	(0.611)	[-0.08, 0.047]
II. Punitive vs. Control					
A. White vs. Black					
Index measure of favorable in-person interactions	0.015	0.056	0.267	(0.605)	[-0.095, 0.125]
Received post-visit callback	0.017	0.043	0.389	(0.651)	[-0.068, 0.102]
Received post-visit offer for unit	0.027	0.035	0.773	(0.78)	[-0.042, 0.095]
B. White vs. Hispanic					
Index measure of favorable in-person interactions	0.061	0.059	1.033	(0.849)	[-0.055, 0.176]
Received post-visit callback	-0.067	0.043	-1.58	(0.057)	[-0.151, 0.016]
Received post-visit offer for unit	-0.013	0.035	-0.363	(0.358)	[-0.082, 0.057]
C. Black vs. Hispanic					
Index measure of favorable in-person interactions	0.042	0.052	0.813	(0.417)	[-0.06, 0.144]
Received post-visit callback	-0.084	0.042	-2.018	(0.044)	[-0.166, -0.002]
Received post-visit offer for unit	-0.04	0.034	-1.17	(0.243)	[-0.107, 0.027]
III. Punitive vs. Monitoring					
A. White vs. Black					
Index measure of favorable in-person interactions	0.018	0.06	0.302	(0.763)	[-0.1, 0.136]
Received post-visit callback	0.023	0.048	0.469	(0.639)	[-0.072, 0.118]
Received post-visit offer for unit	0.032	0.036	0.895	(0.371)	[-0.039, 0.104]
B. White vs. Hispanic					
Index measure of favorable in-person interactions	0.105	0.066	1.58	(0.116)	[-0.026, 0.236]
Received post-visit callback	-0.032	0.051	-0.636	(0.525)	[-0.132, 0.068]
Received post-visit offer for unit	0.008	0.04	0.2	(0.841)	[-0.07, 0.086]
C. Black vs. Hispanic					
Index measure of favorable in-person interactions	0.127	0.059	2.135	(0.034)	[0.01, 0.244]
Received post-visit callback	-0.055	0.047	-1.18	(0.239)	[-0.147, 0.037]
Received post-visit offer for unit	-0.024	0.036	-0.688	(0.492)	[-0.094, 0.046]

Table A21: Estimated Effects of Messaging on Net Discrimination Levels Among Subsample Excluding Likely Discrimination Cases. Cells contain ITT estimates from OLS models with inverse probability weights and block fixed effects. For each reference group versus comparison group pairing, outcomes are net discrimination measures against the comparison group relative to the reference group. Estimated effects that are positive (negative) are interpreted as increases (decreases) in net discrimination against the comparison group relative to the reference group. Estimated p-values are reported in parentheses; p-values correspond to a one-sided test of the null hypothesis of equality of means for the monitoring-control and punitive-control comparisons, and 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.

E.13 Heterogeneous Messaging Effects by the Perceived Race of the Landlord

As an exploratory analysis conducted post hoc, we explore heterogenous messaging effects by the perceived race of the landlord. We code a landlord's perceived race as known only if at least two of the testers independently perceive the landlord's racial group membership in the same way, and other/unknown otherwise. Table A22 presents the distribution of subjects by their perceived race based on this coding procedure.

Table A22: Distribution of Subjects by their Perceived Race. A subject is classified as Black, Hispanic, or White if at least two testers in a matched trio perceive them to belong to that racial group. All other subjects are classified in the Other category.

Subject's Perceived Race	Number of Subjects	Proportion
Black	75	0.11
Hispanic	83	0.13
Other	157	0.24
White	338	0.52

We then partition the experimental sample by the perceived racial category of the landlord, and re-estimate the main specification for each subgroup. Figures A6 and A7 present coefficient plots of the estimated effect among each subgroup with 90% and 95% confidence intervals.

We find that among white landlords and brokers, both the monitoring and punitive conditions have no effect on discrimination against Blacks and Hispanics that is distinguishable from zero. We find suggestive evidence that among Black landlords and brokers, the monitoring condition reduces discrimination against both Blacks and Hispanics (vs. whites) in making callbacks and offers and the punitive condition reduces discrimination against Hispanics (and has no effect on discrimination against Blacks) in making callbacks and offers. Among Hispanic landlords and brokers, we find suggestive evidence that the monitoring condition increases discrimination against Blacks and Hispanics in receiving callbacks and offers, and that the punitive condition reduces discrimination against Hispanics in receiving a callback. Among landlords and brokers for whom their perceived race is coded as unknown, we find suggestive evidence that the monitoring condition reduces discrimination against Hispanics in receiving callbacks and against Blacks in receiving offers, but that the punitive condition has no effect on discrimination against Blacks or Hispanics.

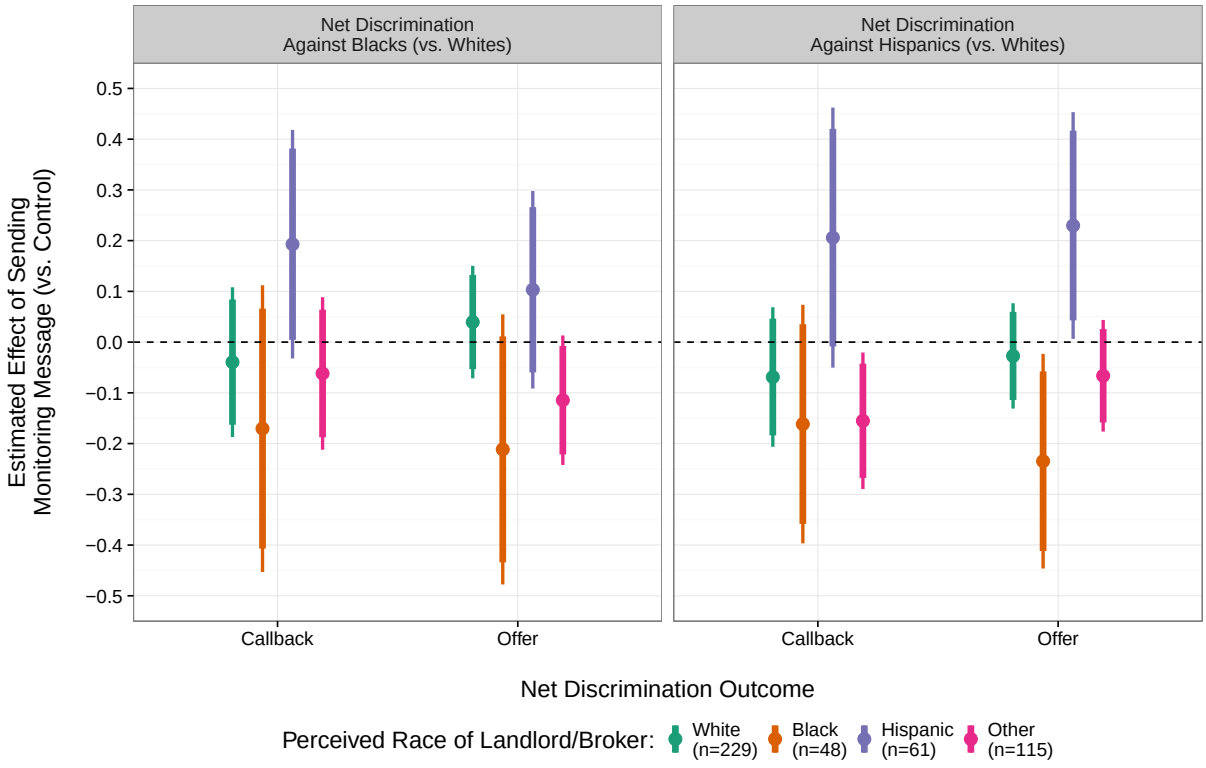


Figure A6: Estimated Effects of Monitoring Messaging on Net Discrimination Levels Relative to Control, by the Perceived Race of the Landlord/Broker

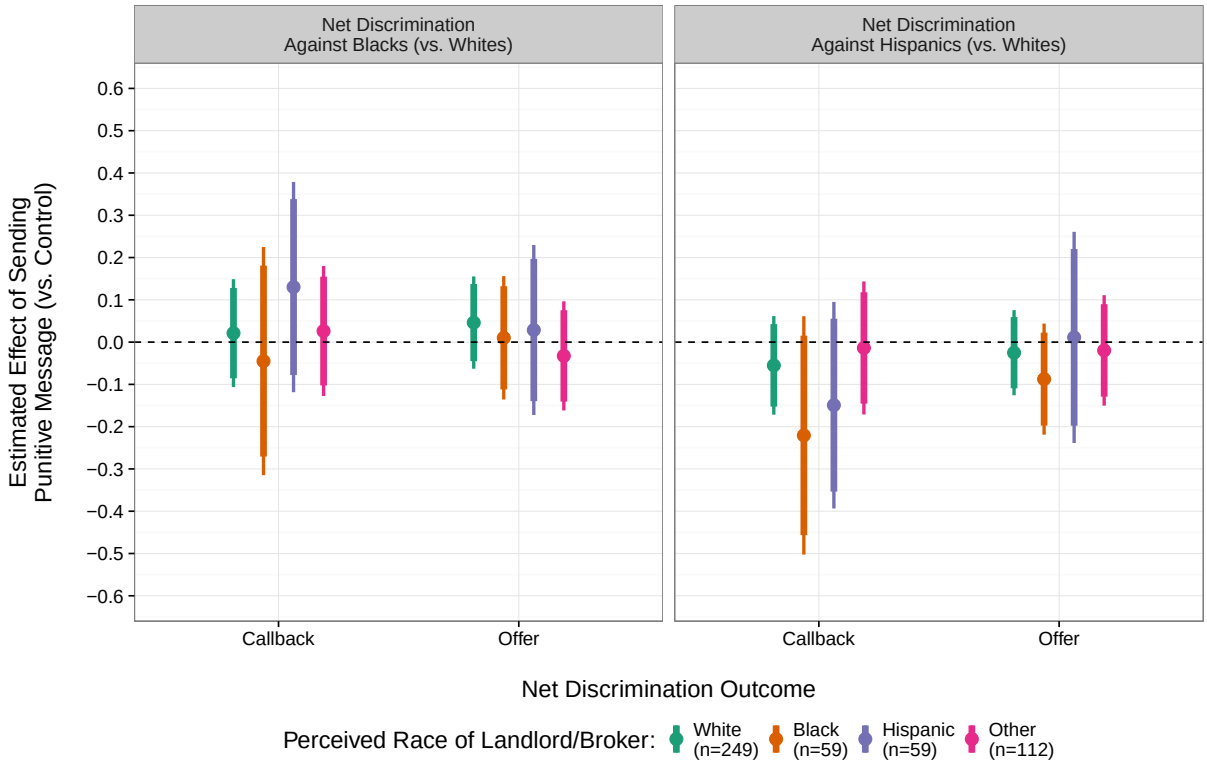


Figure A7: Estimated Effects of Punitive Messaging on Net Discrimination Levels Relative to Control, by the Perceived Race of the Landlord/Broker

E.14 Details of Bayesian Analysis to Assess Policy Implications

Implementing an analogue of our core specifications in a Bayesian framework, we calculate a posterior distribution over the effects of the treatment messages on net discrimination (callbacks and offers) against Black and Hispanic testers. To estimate posterior distributions of the treatment effects, we assume a non-informative, uniform (improper) prior on all parameters. For the data likelihood, we use the model from Equation 1, weighting all observations by the inverse of the probability of assignment to the relevant treatment condition. We then sample from the posterior marginal distribution, $\beta_1|\sigma^2, y \sim N(\hat{\beta}_1, V_{\beta_1}\sigma^2)$, with 10,000 draws from a simple Monte Carlo algorithm. For each posterior, we can then calculate the implied probability that the treatment is effective in reducing discrimination—i.e., the mass under the curve below zero—as well as summary statistics, such as the posterior mean.

E.15 Addressing Spillover Concerns

We acknowledge that spillovers are a potential concern, but we argue that spillovers are unlikely to occur. Specifically, given the context that is the New York City rental housing market and the sampling procedures used, we argue that there is a very low probability that subjects in the experiment interact with each other. To adduce this, we employ a multi-pronged empirical strategy.

First, we estimate bounds on the probability that a landlord or broker (who posts rental ads on Craigslist) enters the audit and experiment samples and show that these probabilities are low. 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 experimental sample is between 0.77% and 3.6%.

To estimate lower bounds, we divide the number of subjects in the audit and experiment samples by the total number of sampled listings (which were randomly sampled from the universe of listings every day the study was implemented). The probability a subject enters the audit sample is 2,711 / 84,499 = 3.2% and the probability a subject enters the experiment sample is 653 / 84,499 = 0.77%. The reason why we treat these estimates as lower bounds is as follows. These estimates are unbiased only under the strong and unlikely assumption that there are no duplicate listings on Craigslist and that each listing uniquely corresponds to a landlord or broker. Because we ensure that there are no duplicate subjects or listings pursued, relaxing these assumptions would deflate the denominator and thereby increase the estimated probability.

To estimate upper bounds, we (1) estimate the percentage of duplicate-subject listings that exist on Craigslist (i.e., listings with duplicate landlord or brokers); (2) multiply that estimate by the number of sampled listings from the study to estimate the number of duplicate-subject listings among all of the sampled listings from the study (i.e., the number of duplicate-subject listings among the 84,499 sampled ads); (3) subtract off the estimated number of duplicate-subject listings from the total number of sampled ads; and (4) use the resulting estimate of the "de-duplicated" number of listings as the denominator to estimate the probability a landlord or broker enters the audit sample or the experiment sample.

As a first step toward estimating upper bounds, we estimated the proportion of rental listings requiring contact by phone (55.8%) by searching the text of all Craigslist listings that we scraped during the study for matches against a regular expression for U.S. 7- or 10-digit phone numbers.

Corresponding to the steps above, we estimate the percentage of duplicate-subject listings by

phone number, which at 79% is very high. When we use this estimate to calculate a new denominator, we find that the upper bounds are 2,711 / 18,091 = 15% for the audit sample and 653 / 18,091 = 3.6% for the experimental sample. For the experimental sample especially, we think this makes a strong prima facie case for minimal interference.

Finally, we characterize the experiment sample as a very small random sample of landlords and brokers in the New York City rental market. The actual number of active landlords and brokers in the New York City rental market is unknown. We therefore estimate the denominator to adduce a very conservative upper bound on this quantity. 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 which, when interpreted as a very small random sample of the population, suggests that interactions among subjects in the experiment are highly unlikely. The 2% estimate is calculated by dividing the number of subjects by the estimated number of licensed real estate brokers in salespeople in Manhattan alone, or 27,000 (see, e.g., `https://cooperator.com/article/new-york-citys-real-estate-brokers`). Since this denominator does not include brokers in the other 4 boroughs or landlords in any of the 5 boroughs of New York City, we infer that the true percentage must be much smaller than 2%.

Alternatively, we estimate the total number of landlords and brokers in the New York City rental market using capture-recapture sampling and the Lincoln-Petersen population size estimator. Capture-recapture provides an estimate of a population from two sampling steps: First, a sample (from, e.g., a population of animals) is taken and marked. In the second period, another sample is taken and the proportion of this second sample that has been marked is used as an estimate of the ratio of the size of the first sample to the whole population. We analogize this procedure to our data as follows: We take a random sample of 1,000 listings with phone numbers from the entire study sample (preprocessed to remove duplicate-landlord listings) to be our sample of "marked" observations. We then compute the proportion "marked" in a second random sample. By dividing the number of listings with phone numbers in the first set by the proportion of listings in the second set with matching numbers, we generate an estimate of the number of landlords and brokers in New York City who can be contacted by phone, 10,526. This number can then be divided by the estimated proportion of landlords and brokers contactable by phone (0.558) to generate a final estimate of 18,864. Again, this suggests a large enough population that spillover effects are unlikely to have occurred.

E.16 Joint Distribution of the Number of Testers in Matched Trios Who Receive a Callback and an Offer

We examine the joint distribution of the number of testers in a matched trio who received a callback and the number of tester in a matched trio who received an offer.

We find that multiple testers in a matched trios receive offers and that receiving an offer does not drive receiving a callback. Among matched trios where only one tester received a callback (n=143), only in 75 of those trios (52.45%) did the tester who received the callback also receive an offer, and in the other 68 cases (47.6%) the callback did not include an offer. Among matched trios where 2 testers received a callback (n=59), both callbacks included an offer in 20 trios (33.9%), only one of the two callbacks included an offer in 20 trios (33.9%) and none of the callbacks included an offer in the remaining 19 trios (33.2%). Among matched trios where all 3 testers received a callback (n=24), all three callbacks included an offer for 5 trios (20.8%), two of the

three callbacks included an offer in 6 trios (25%), one of the three callbacks included an offer in 8 trios (33.3%), and none of the three callbacks included an offer in 5 trios (20.8%).

This is a peculiar pattern, but one that is (at least anecdotally) known to occur in the competitive New York City rental market where a verbally communicated rental offer is non-binding and is a signal from the landlord/broker to move forward in the process to execute a lease. Because it is non-binding and therefore costless, multiple housing applicants may receive this verbal signal as a mixed strategy, and the person who responds first will be the one who in fact gets to rent the unit.

We rule out the possibility that we observe one tester receiving an offer because another tester (who previously received an offer) turned it down. This is because *all* testers were instructed per the experiment's field protocol to not make a decision on offers (e.g., to say they still had some units to view before making a decision) – and importantly to not decline an offer – until 48 hours after the appointment, when they would tell the subject they were no longer interested in the unit.

F OTHER SUPPLEMENTARY MATERIAL

F.1 Additional Potential Interpretations for Mixed Findings for Blacks and Hispanics

Two additional interpretations presuppose that landlords have prejudices against both Black and Hispanic renters with more entrenched prejudices against Blacks than Hispanics. Given this assumption, one possible interpretation is that landlords in the punitive condition who are potentially spooked by the audit and the governmental message may strategically decide to act favorably toward the minority renter they dislike the least—that is, they treat Hispanic renters more favorably than Black and white renters to appear non-discriminatory, but do so in part to avoid having to rent to a Black tester. Another possible interpretation is that under the punitive condition, landlords do not act favorably toward the group for which they have the strongest prejudices in order to avoid repeated interactions with renters belonging to that group in the future.

F.2 Deviations from the Pre-Analysis Plan

We registered our pre-analysis plan at Experiments in Governance and Politics on April 3, 2013 (Link to Pre-Analysis Plan: [LINK]). Table A23 documents deviations from the plan as well as a set of clarifications.

Table A23: Pre-Analysis Plan. This table notes any deviation from the planned analyses.

	Analysis Plan	**Inconsistency / Clarification**
Experimental analyses for encouragement design	Main analysis	1. Deviation: For parametric estimates of the ITT and CACE, we weight all observations by the inverse of the probability of assignment to the relevant treatment condition as noted in the Pre-Analysis Plan. In our analysis, however, we additionally include block dummy variables which is more faithful to the randomization strategy. 2. Deviation: Table A14 in Appendix E.7 reports estimates using IV regression models with inverse probability weights and block fixed effects. We do not include nonparametric estimates of the CACE and instead employ parametric estimators of the CACE that are comparable to our parametric ITT estimators, which are specified in order to be more faithful with the randomization strategy. We also do not estimate covariate-adjusted CACEs given the minimal efficiency gains shown in Table A15 for the ITT. 3. Clarification: We use a principled covariate selection and covariate adjustment estimation strategy for the ITT. While not specified in the PAP, the procedure we employ minimizes researcher discretion and fishing.
	Hypothesis testing & inference	4. Deviation: To compute p-values, we use the cumulative density of the t distribution rather than randomization inference, because the primary null hypotheses of interest concern whether average effects equal zero (for which randomization inference would not make sense), rather than sharp null hypotheses that requires an additional assumption of constant zero treatment effects for all subjects.
	Sensitivity analysis for tester heterogeneity	5. Clarification: The model described in Section 4.6 of the Pre-Analysis Plan is equivalent to the one estimated in section C.4 and displayed graphically. 6. Deviation: The "cross-validation procedure" described in Section 4.6 of the PAP is best thought of as an additional sensitivity check. We did not follow this procedure because it effectively induces attrition (and potential bias) at each step. Rather than rely on this analysis, we address the substantive concern in the ANOVA reported in Section C.4 (see Figure A4). Finally, if sensitivity to particular testers were an issue, the lasso regression procedure outlined in Section E.8 would have chosen tester fixed effects as covariates to be included in the models.
Employment stability signal manipulation	Treatment by treatment interaction	1. Deviation: We do not include these analyses in the paper, which focuses on primary results only.

F.3 Acknowledgments

We extend special thanks to Joan Russell, Clare Wiseman, and Ruby Hlivko for their excellent work as project managers over the course of the study. We thank the following individuals for their dedicated work as testers for the project: Tania Aparicio, Evette Addai, Hannah Blume, Destinee Bowrin, John Cales, Jillian Cantwell, Courtney Cauthan, Xilonem Clarke, Reese Crispen, Dana DeBari, Cristina De La Rosa, Lucas Denton, Matthew Diaz, Kenneth Edusei, Inemesit Essien, Talisa Feliciano, Ellen Gagne, Bianca Garcia, Elizabeth Garcia, Emmanuel Garcia, Gabriel Garcia, Lais Gomes Duarte, Ruby Hlivko, Jevaun Joseph, Rumando Kelley, Keenan Lambert, Douglas Land, Vanesa Lauradin, John Long, Julia Mesler, Linden Miller, Richard Minaya, Ryan Mitchell, Shannon Murray, Luis Ortiz, Joe Palmisano, Christiaan Perez, Shanta Pamphile, Dale Reyes, Haley Rice, Julieta Salgado, Ayinde Samuel, Camillia Shofani, Rebecca Suldan, Fadumo Tahlil, Matthew Taylor, Joshua Thompson, Kirya Traber, Steven Velez, Hannah-Sophie Wahle, and Jordan Woodsmall. We thank Yasmine Ergas from the Columbia University Institute for the Study of Human Rights for her support. We thank Caroline Peters for her valuable contributions to the project in its initial stages. We thank Glenn Martin for sharing best practices in implementing audit studies. We thank Sarah Khan for providing expert editorial assistance, and Alex Coppock, Shaynah Jones, Alexa Pazniokas, and Jean Pierre Salendres for providing excellent research assistance. The authors serve a pro bono advisory role for the present study, by providing input into the experimental design and by conducting the analysis of the study data. The study and data are wholly owned by the New York City Commission of Human Rights.

REFERENCES

Berger, Adam L, Vincent J Della Pietra and Stephen A Della Pietra. 1996. "A maximum entropy approach to natural language processing." *Journal of Computational Linguistics* 22(1):39–71.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94(4):991–1013.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet Sekhon and Bin Yu. 2016. "Lasso adjustments of treatment effect estimates in randomized experiments." *Proceedings of the National Academy of Sciences of the United States of America* 113(27):7383–7390.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1):1–22.

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.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. New York: WW Norton.

Imbens, Guido W. and Donald B. Rubin. 1997. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25(1):305–327.

Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*. Springer Berlin Heidelberg pp. 137–142.

Kling, Jeffrey R, Jeffrey B Liebman and Lawrence F Katz. 2007. "Experimental analysis of neighborhood effects." *Econometrica* 75(1):83–119.

Long, Qi, Roderick J. A. Little and Xihong Lin. 2010. "Estimating Causal Effects in Trials Involving Multi-Treatment Arms Subject to Non-compliance: A Bayesian Framework." *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 59(3):513–531.

Massey, Douglas S and Garvey Lundy. 2001. "Use of Black English and Racial Discrimination in Urban Housing Markets New Methods and Findings." *Urban Affairs Review* 36(4):452–469.

Pager, Devah, Bruce Western and Bart Bonikowski. 2009. "Discrimination in a Low-Wage Labor Market A Field Experiment." *American Sociological Review* 74(5):777–799.

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.

Purnell, Thomas, William Idsardi and John Baugh. 1999. "Perceptual and phonetic experiments on American English dialect identification." *Journal of Language and Social Psychology* 18(1):10–30.

Ross, Stephen. 2002. Paired Testing and the 2000 Housing Discrimination Study. In *Measuring Housing Discrimination in a National Study: Report of a Workshop*, ed. National Research Council. Washington, D.C.: National Academies Press.

Viera, A. J. and J.M. Garrett. 2005. "Understanding interobserver agreement: the kappa statistic." *Family Medicine* 37(5):360–363.

Yuan, Shuai, Hao Helen Zhang and Marie Davidian. 2012. "Variable selection for covariate-adjusted semiparametric inference in randomized clinical trials." *Statistics in Medicine* 31(29):3789–3804.