

Research Article

Do Public Housing Agencies Discriminate Against Hispanics? A Large-Scale Replication Study to Assess Discrimination

Jacob Authement*, Michelangelo Landgrave*, Nicholas Weller**

Abstract: Previous studies have demonstrated discrimination by political elites. However, research about discrimination by street-level bureaucrats has mixed findings. We conducted a study that followed up on a prior audit of public housing agencies by Einstein and Glick (2017), and we failed to replicate their main finding—differences in how white (non-Hispanic) and Hispanic aliases are treated. We made two significant improvements in our study: we pre-test the names used to ensure that they primarily manipulate perceptions of ethnicity, and we include approximately three times more public housing agencies. Our larger sample size produces more precise estimates and allows us to better explore potential moderating factors such as racial resentment, professionalization, and population size. Like Einstein and Glick (2017), we do not find discrimination in either overall reply rates or reply rates within 24 hours. Unlike Einstein and Glick (2017), we fail to find statistically significant differences in the friendliness of replies to white and Hispanic aliases. We also do not find any meaningful variation across racial resentment or professionalization but find suggestive evidence of moderation by city population size.

Keywords: Discrimination, Latino Politics, Housing, Audit Study

One of the key methods to study the presence and consequences of discrimination is audit studies (Gaddis, 2018; Grimmekhuijsen et al., 2019). Audit studies involve contacting a sample of individuals or organizations via email and studying if the reply rate or characteristics of the replies varies based on the name in the email. Most audit studies in public administration and political science have focused on the behavior of political elites, and sufficient studies have been done to allow for aggregation and meta-analyses of these studies (Costa, 2017). Researchers have also studied the behavior of non-elite political actors, such as street-level bureaucrats (Einstein & Glick, 2017), but relatively few of these studies have been published and the more recent studies have not found evidence of discrimination (Lowande & Proctor, 2020). Since the public is more likely to interact with street-level bureaucrats on a regular basis than other political actors and because ensuring equal access to housing resources is an important policy goal, we replicated a major prior study (Einstein and Glick 2017) that examined how bureaucrats in public housing agencies respond to emails from putatively Hispanic and white (e.g., individuals of predominantly non-Hispanic European descent) residents. Einstein and Glick (2017)'s initial study found that, while public housing agencies do not discriminate between white and Hispanic contacts in their reply rates, public housing agencies are less friendly in their responses to Hispanic constituents. Einstein and Glick (2017) argue that this means Hispanics receive lower-quality service than their white counterparts. Given the substantive importance of this finding we replicated their study with two important improvements: we pre-tested the names used in the experiment to ensure they manipulated the intended characteristics, and our sample size is approximately three times that used in the original study. By using pretested names and removing the gender conditions (which found null results in the original study), our study focuses on racial differences between white and Hispanic aliases.¹ Additionally, our much larger sample size provides us with both more statistical power to detect small effects and more confidence that our estimated

*University of Missouri, **University of California, Riverside

Address correspondence to Michelangelo Landgrave at michelangelolandgrave@missouri.edu.

Copyright: © 2021. The authors license this article under the terms of the Creative Commons Attribution 4.0 International License.

effects are not due to some aspects of the sample.² Consistent with the original study, we do not find a significant difference in reply rates between Hispanic and white names. We also do not find a difference in the friendliness of replies between Hispanic and white names. The latter null result is important, because Einstein and Glick (2017)'s major finding was evidence of discrimination against Hispanics in terms of friendliness, which we do not find in our replication.

In addition to estimating the average treatment effect, we examine if estimated discrimination differs between the type of city (e.g., core vs non-core city, principal vs non-principal city). We do not find that estimated discrimination differs between core and non-core cities, but we found that the estimated treatment effects in micropolitan principal (population area less than 50,000) and non-principal cities (all other cities in an urban area) differ from metropolitan principal cities (population area more than 50,000). This led to us to explore city population size as a moderator, and we find weak evidence that public housing authorities in bigger cities are more likely to reply and more likely to be friendly to Hispanic aliases than those in smaller cities. However, we caution against interpreting this result as a causal effect of population size alone since population size is bundled with many unobserved factors that could confound the results.

We also examine if the treatment effect is moderated by characteristics such as state-level racial resentment or the performance of public housing agencies. We do not find any meaningful moderation by either attribute. The null moderation results are surprising to us given that we fielded our experiment in summer 2020 when inter-racial relations in the United States were especially salient due to the George Floyd protests and related incidents. The time period of our study could plausibly affect discrimination either by making it more likely in certain states and/or make it less likely in other states as public officials become more aware of their own biases and acted to reduce them. These countervailing factors may have also led to no aggregate difference, but we cannot parse that out from our data.

There is a continuing discussion about the ethics of audit studies (Bischof et al., 2021; Landgrave, 2020; Zittel et al., 2021), and while it is beyond the scope of the present paper to discuss research ethics in great length we believe that replication studies like our own can be done ethically and should be done to replicate prior results rather than canonizing the results of a particular study just because of when it was conducted.³ Our study was pre-registered on Open Science Framework, and an anonymized version of the pre-registration can be found in **Appendix C**.

Name Selection

An important part of a well-executed audit study is to choose names that represent the relevant traits of the hypothetical individuals in a study, and the use of ethnically distinctive names is common in correspondence audit studies (Butler & Homola, 2017). The previous study by Einstein and Glick (2017) used six aliases; one characteristically male and one characteristically female name for each racial/ethnic category used (white, Black and Hispanic). They used Brett Smith (white male), Emily Smith (white female), Tyrone Johnson (Black male), Shanice Johnson (Black female), Santiago Martinez (Hispanic male) and Gabriela Martinez (Hispanic female).

We did not include female or black names in our study because the original study did not find any statistically significant differences between men and women or white and black aliases. See Butler (2014) for a review on the (mostly) null differences between male and female constituents in political science/public administration audit studies. By focusing on Hispanic and white aliases we enhance our statistical power to test the original study's main finding that Hispanic aliases were less likely to receive a friendly reply than their white counterparts.

When selecting names it is critical to ensure that names only differ on the characteristic(s) of interest otherwise the information equivalence assumption is violated (Dafoe et al., 2018; Landgrave & Weller, 2022), and we cannot interpret whether results are due to discrimination against Hispanics or due to discrimination against some other attribute(s) associated with Hispanicity but separate from ethnicity (e.g. socio-economic status, nativity, etc.).⁴ Hispanic surnames were paired with English first names in order to signal that our Hispanic aliases were native-born Americans, as opposed to first generation migrants (Abramitzky et al., 2020), because we wanted to increase our confidence that we were comparing white citizens with Hispanic citizens as opposed to white citizens with Hispanic non-citizens.

To ensure that the names we used as aliases were ethnically distinct, but as similar as possible in other ways, we conducted a name perception study using subjects from Amazon Mechanical Turk (MTurk) in late

2018. Ideally we would have wanted to test name perceptions using public housing agency bureaucrats, but due to the difficulty of doing so without our field experiment being detected, we used MTurk workers as a second-best option (Gaddis, 2017; Lahey & Mosquera, 2022). As part of a larger name perception study, each name's perceived race/ethnicity, gender, nativity, and socio-economic status was rated by approximately 60 to 65 unique MTurk subjects. See **Appendix D** for further discussion of name perceptions by socio-economic status and gender.

The names used in the audit were Nicholas Smith (white), John Ryan (white), Joe Garcia (Hispanic), and Michael Hernandez (Hispanic). To signal Hispanicity we relied on surnames primarily held by Hispanics according to the United States Census. In **Table 1** we report MTurk subjects' perception of a name's race and ethnicity. A supermajority of MTurk subjects respectively perceived Nicholas Smith and John Ryan to be white. Similarly, a supermajority perceived Michael Hernandez and Joe Garcia to be Hispanic.

Table 1 – Perceived Race and Ethnicity by Name

Race/Ethnicity	Name			
	Nicholas Smith	John Ryan	Joe Garcia	Michael Hernandez
White	86.15%	96.88	12.31	10.45
Black	6.15	0.00	0.00	1.49
Hispanic	0.00	0.00	81.54	83.58
Asian	1.54	0.00	0.00	0.00
Native American	1.54	1.56	0.00	1.49
Middle Eastern	1.54	0.00	0.00	0.00
Mixed	1.54	0.00	6.15	1.49
Other	1.54	1.56	0.00	1.49

Note: Purportedly Hispanic names are 82.57 percentage points more likely to be perceived as Hispanic than purportedly white names. Two-sided t-test, p-value < 0.01.

Previous studies on the perceptions of Hispanic names (Gaddis, 2017b; Gaddis & Ghoshal, 2020) caution against using English first names with Hispanic surnames without pretesting and critically thinking of the signals they will send. The names we use attempt to manipulate perceptions of race while minimizing differences in perceived nativity. We use English first names with Hispanic surnames to signal that our Hispanic aliases are U.S. citizens. Our efforts appear to be successful, because a majority of MTurk subjects perceive all four aliases, including the Hispanic aliases, to be citizens as shown in **Table 2**. Although all aliases are perceived to be citizens, and despite our efforts to signal Joe Garcia and Michael Hernandez as being native-born by pairing Hispanic surnames with English first names, a substantial portion of subjects perceive the Hispanic aliases as being foreign-born (non-citizens or naturalized citizens). This is not unique to these two specific names, but a feature common to Hispanic names. Therefore, a practical limitation is that if we find evidence of discrimination against Hispanic aliases, we cannot infer that the effect is due to ethnicity alone rather than a combination of ethnicity and migrant status (Garcia & Sadhwani, 2022). Nonetheless, we believe our design gets closer to testing for ethnic/racial based discrimination (Landgrave, 2021) than prior studies that used names which were even more likely to be perceived as non-citizens.

Table 2 – Perceived Nativity by Name

Nativity	Name			
	Nicholas Smith	John Ryan	Joe Garcia	Michael Hernandez
Natural-born Citizen	78.46%	85.94	60	64.18
Naturalized Citizen	20	12.5	30.77	29.85
Non-citizen Migrant	1.54	1.56	9.23	5.97
Total	100	100	100	100

Audit Research Design

We fielded our study in late August 2020. We collected email addresses from all public housing agencies that had contact information available through the United States Department of Housing and Urban Development (HUD) website. We deviate from Einstein and Glick (2017) here in our choice of sample. Their sample is limited to public housing authorities defined clearly within a core city. We included all available public housing authorities, because we did not have a theoretical reason to eliminate non-core cities from our sample, and we desired as large a sample size as possible. Each public housing agency received a short informational request via email asking for help with the application process. We used the same script as Einstein and Glick (2017). See **Appendix A** for further information about the script. Within each state, we randomly assigned public housing agencies to receive an email from either a purportedly white or Hispanic alias using Blair et al. (2019)'s block randomization tool. Block randomization ensured that we have a similar number of housing agencies assigned to the Hispanic and white names within each state, so our inferences are less likely to be affected by cross-state differences in public housing agency behavior. Our dependent variables are (a) is a reply received, (b) is a reply received within 24 hours, and (c) is a reply friendly. A friendly reply is defined as a reply including the sender's name – which is the same definition used in the original study. Differences in these outcomes between white and Hispanic aliases can be interpreted as evidence of discrimination.

Results

We restrict our analysis to public housing authorities in sub-county localities (cities, towns, etc.) that we can match with census data, as in Einstein and Glick (2017). After removing autoreplies and other invalid observations, we are left with a sample size of 2,099 observations. In comparison, Einstein and Glick (2017)'s white/Hispanic sample was 665 public housing agencies.⁵

We use a linear probability model (Gomila, 2021) to estimate ethnic differences in (a) whether any reply is received, (b) does the reply occur within 24 hours, and (c) reply friendliness. All three outcome variables are binary. It is possible that non-responses are themselves indicative of discrimination, and if so analyses (b) and (c) above would be biased. It is recommended to adjust for this bias (Coppock, 2018), which we do by treating unobserved replies as either occurring after 24 hours or being unfriendly (depending on the analysis). We also conduct our analysis by simply ignoring the unobserved replies and thereby conditioning the analyses on a reply being received, and we re-estimated our analyses assuming that unobserved replies would have been friendly. Results remain substantially unchanged if we alter our post-treatment bias correction strategy – see **Appendix B**.

Results are summarized in **Figure 1**. For comparison, the results with the same estimation technique, including adjustment for bias, using Einstein and Glick (2017)'s original data are also included in **Figure 1**. Tabular results, adjustment for multiple comparisons, logit estimation and other additional analyses are included in **Appendix B**. Across specifications and estimation methods our substantive results are unchanged.

We find no statistically significant evidence that Hispanics are less likely to receive any reply (p -value = 0.317), less likely to receive a reply within 24 hours (p -value = 0.063), or less likely to receive a friendly reply (p -value = 0.667) using a sample size that is over three times the size of the initial study's white/Hispanic sample.

We formally test if the results differ between the two studies in **Table 3**. We find that both the overall reply rate and the reply rate within 24 hours are lower in our replication study than the original study. At conventional rates of statistical significance, we do not find a difference between white and Hispanic names for either the overall reply rate or the reply within 24 hours outcomes. In our study, the effect of Hispanic names on replies within 24 hours is just barely insignificant and is opposite signed from the original study, but the two estimates are not statistically different from each other. Prior research has found that constituent-initiated contacts from minorities tend to be more urgent (Thomas, 1986, Hamel & Derek, forthcoming), and perhaps bureaucrats respond more quickly to Hispanic aliases due to the perceived effect of the pandemic on minority access to housing.

In our study there is no difference in the rate of friendly replies between white and Hispanic names (-0.008 , p -value = 0.667) whereas in the original there was a significant difference (-0.132 , p -value < 0.001), and these two estimates are significantly different from each other (p -value = 0.002).

Figure 1 - Average Treatment Effect of a Hispanic Alias on Reply Rate, Reply within 24 Hours, and Reply Friendliness

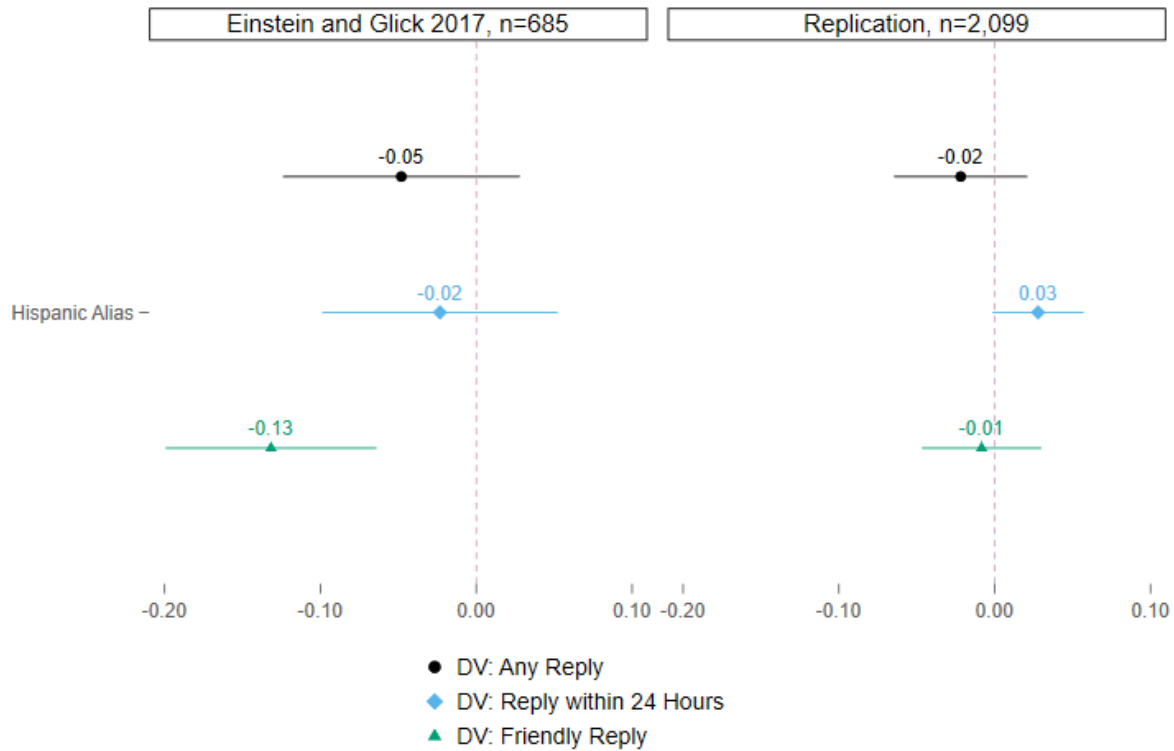


Table 3 – Testing for Different Treatment Effects Across Studies

VARIABLES	Any Reply		Reply within 24 Hours		Friendly Reply	
	(1)	(2)	(3)	(4)	(5)	(6)
Hispanic Alias	-0.028 (0.019)	-0.048 (0.039)	0.016 (0.015)	-0.023 (0.030)	-0.038** (0.017)	-0.132*** (0.035)
Replication Study	-0.030 (0.022)	-0.044 (0.032)	-0.287*** (0.017)	-0.314*** (0.025)	-0.000 (0.020)	-0.065** (0.029)
Hispanic Alias * Replication Study		0.026 (0.045)		0.051 (0.034)		0.123*** (0.040)
Constant	0.564*** (0.022)	0.575*** (0.028)	0.414*** (0.017)	0.435*** (0.022)	0.297*** (0.020)	0.347*** (0.025)
Observations	2,764	2,764	2,764	2,764	2,764	2,764
R-squared	0.001	0.002	0.093	0.094	0.002	0.005

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

The findings reported in this section suggest that in our study public housing agencies do not appear to discriminate in their email replies based on whether a contact has a white or Hispanic name.

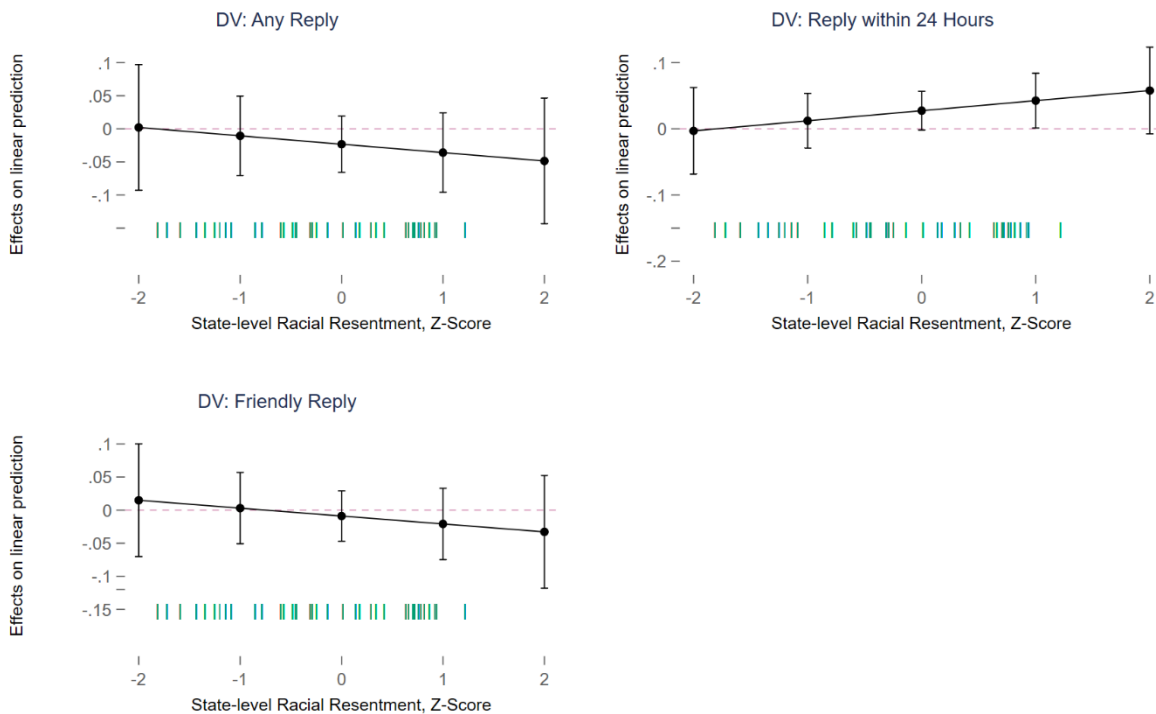
Exploratory Moderation Analyses

We turn in this section to analyses focused on moderating variables that might be related to heterogeneity in treatment effects. These results should be interpreted as exploratory because we did not pre-register them. We look at two primary moderators that we have good prior reason to believe might affect discrimination against Hispanic names – state-level racial resentment and public housing agency professionalism. We expect

discrimination against Hispanic constituents should be greater in states where racial resentment is highest. We also examine whether professionalization moderates discrimination, because a number of prior studies find that discrimination is reduced by institutional professionalization (Landgrave & Weller, 2020; Lowande & Proctor, 2020).

We first turn to analyzing whether state-level racial resentment affects discrimination against Hispanic aliases. To measure state-level racial resentment we use data from Smith, Kreitzer, and Suo (2020). For ease of interpretation, we calculate the standardized score (i.e., Z-score) of state-level racial resentment. As controls we use Trump’s 2016 state level vote share and the 2019 Hispanic share of the state population.⁶ We show results graphically, as advised by Brambor, Clark, and Golder (2005) when presenting interactive effects, in **Figure 2** (n=2,099). The figure includes a rug plot showing the distribution of observations by racial resentment. Across all three outcomes, we do not find evidence that discrimination varies with racial resentment.

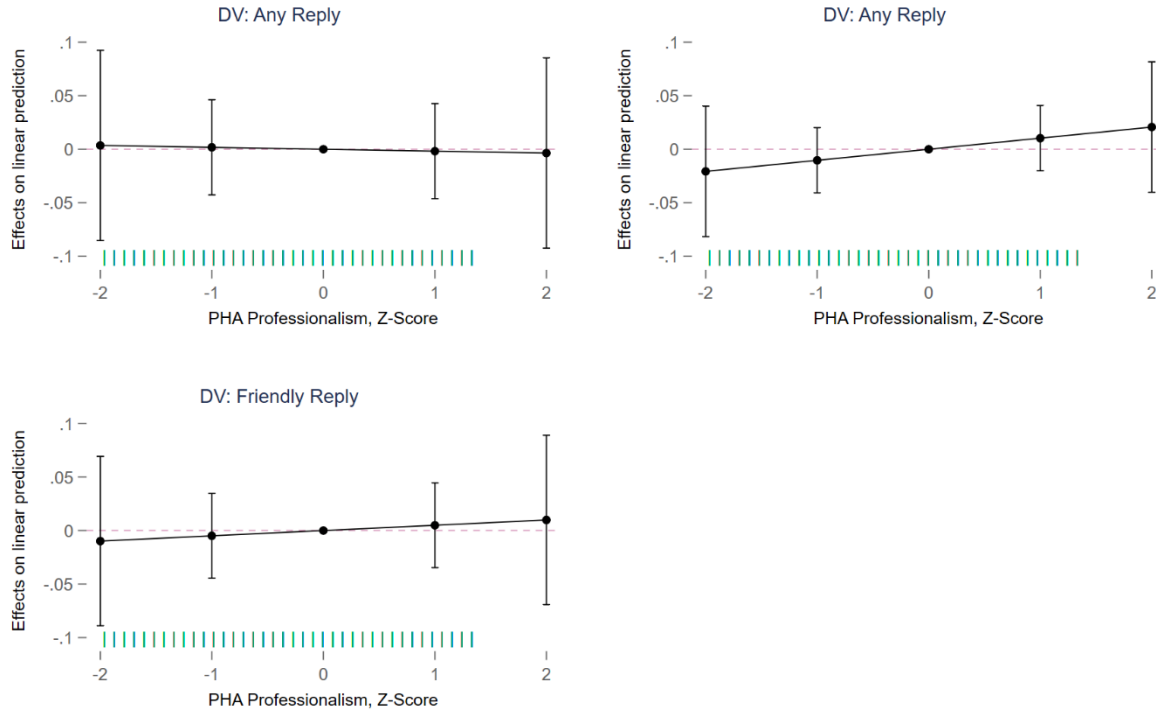
Figure 2 – Relationship between Discrimination and State-level Racial Resentment



Note: Controls include 2016 Trump Vote Share, and 2019 State Hispanic Share. n=2,099.

We also test moderation by the professionalism of the public housing agency. Past studies have argued that more professionalized political institutions are less likely to discriminate (Landgrave & Weller, 2020; Lowande & Proctor, 2020). For this analysis we use HUD’s assessment scores for PHAs. For ease of interpretation, we calculate the Z-score for PHA professionalism. As controls we include the Hispanic share of the population and city population size.

As shown in **Figure 3** (n=1,946), we fail to find any evidence that our outcomes are moderated by professionalism. This moderation analysis was also conducted by Einstein and Glick (2017), and they similarly failed to find evidence of moderation by professionalism. One plausible explanation is that, despite variation in professionalization, public housing agencies on a whole are sufficiently professionalized to reduce inequities in service delivery. Future research should examine the moderating effect of professionalism in government agencies with greater variation.

Figure 3 – Relationship between Discrimination and Professionalism

Note: Controls include 2018 City Hispanic Share, and 2018 City Population. n=1,946.

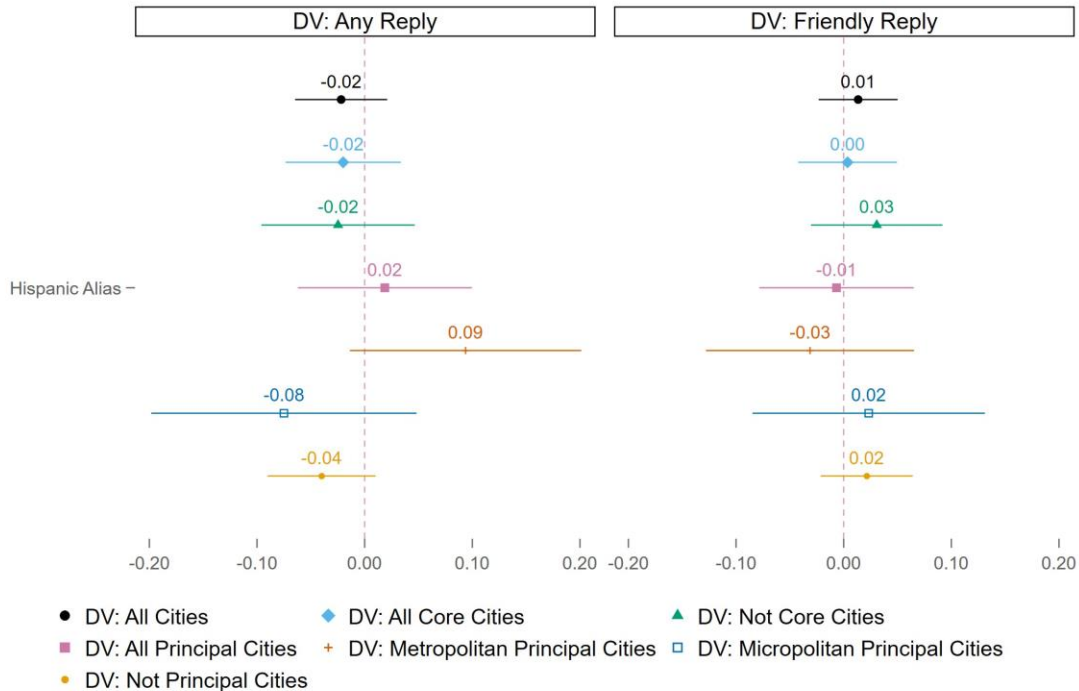
We do not find evidence that estimated discrimination is influenced by plausible moderators (Landgrave & Weller, 2020; Lowande & Proctor, 2020) such as state-level racial resentment or PHA-level professionalism. We would not interpret our results as evidence that public housing bureaucrats do not engage in racial/ethnic discrimination, because discrimination can occur in multiple, unstudied ways. At the same time, our results should lead us as a profession to revise downward our confidence in Einstein and Glick (2017)'s original findings that housing agencies were less friendly in their responses to Hispanic names. We have no reason to believe that Einstein and Glick (2017)'s results were incorrect at the time. It is possible that as audit experiments have become known among political actors (Landgrave, 2020), those same actors have shifted their behavior to reduce observed discrimination; i.e. our null results may be driven by the Hawthorne effect. We believe this is unlikely for two reasons. First, our study was fielded multiple years after Einstein and Glick (2017) so we would need to believe that street-level bureaucrats are still being affected by knowledge of the original study despite it being years later and there likely being at least some turnover in the staff at the public housing agencies. Second, audit studies of elites and the public have been conducted for a much longer time, and we continue to find discrimination in such studies suggesting that the Hawthorne effect may not readily change behavior. However, if researchers are concerned about the Hawthorne effect in this domain, then audit experiments need to be combined with other methodological tools (Gaddis, 2019; List, 2004).

City Type Heterogeneity

One of the areas in which our study differed from the original Einstein and Glick (2017) study is that their study only included public housing agencies in core cities, and we included all available cities in our sample. We believe that auditing all available cities is a strength of our research design, because it increases the precision of our estimates. However, core cities are the largest city in their respective counties, and our different results than Einstein and Glick (2017) may be a result of including smaller cities in our sample, because smaller cities may differ in multiple ways than larger cities. To address this concern, we estimated the treatment effects separately for core and non-core cities and based on 'principal city' status -- whether a given city in our sample was the principal city in a micropolitan (pop < 50,000) or metropolitan (pop > 50,000) area as defined by the US Census.

We used these approaches to explore the possibility of moderation by city type, because they each capture slightly different ways to measure the type of city. Results are presented in Figure 4 below.

Figure 4 – Analysis by Core/Principal City Status



Results from core and non-city cities are near indistinguishable from one another, but we do find some suggestive evidence of differences by whether a city is a metropolitan principal city, micropolitan principal city, or not a principal city. Micropolitan principal (p -value = 0.672) and non-principal (p -value = 0.326) cities appear to be less likely to reply to Hispanic aliases than metropolitan principal cities. In **Appendix B**, we further explore this heterogeneity and find suggestive evidence that city population is positively correlated with Hispanic aliases receiving a higher rate of replies than their white counterparts. We urge caution in interpreting this relationship because a city's population is bundled with multiple other factors that we are unable to disentangle with our present data limitations. Further study is required to fully understand the underlying cause of moderation by city population size.

Discussion and Conclusion

The original Einstein and Glick (2017) paper has been widely cited as evidence that public housing agencies discriminate against those with Hispanic names. Given the importance of this finding, we designed a replication study that built on the initial study by increasing the sample size and pre-testing the Hispanic names to ensure that people perceived them as intended. Consistent with the prior study, our results also do not find discrimination in overall reply rates and reply rates within 24 hours (Einstein & Glick, 2017). Despite our study featuring almost three times as many public housing agencies as the original study and therefore greater ability to detect differences in behavior, we find no effect of the Hispanic alias on the rate of friendly responses from public housing agencies.

Taken together, the results suggest that on average there is little evidence of discrimination against Hispanics (vs. non-Hispanic whites) by public housing agencies in communications. This is largely consistent with Einstein and Glick (2017)'s original findings; they also failed to find discrimination in two of the three outcomes they analyzed. We think this raises two important points, one substantive and one methodological. Substantively, we believe that despite the absence of evidence for discrimination in this context that it is inappropriate to infer that there is not discrimination, in general, against Hispanics by public housing agencies

because these studies only examine one aspect – initial communication – of a long process to acquire public housing benefits. Future studies should investigate evidence of discrimination in other aspects of street-level bureaucracy or examine other types of bureaucracies. Audit studies should also explore other potential aspects of identity that are correlated with race and/or ethnicity, but are distinct (e.g., nativity status), which may explain differential treatment for historically disadvantaged groups. Hispanic names more strongly associated with migrants (vs. native born Hispanics names) may be discriminated against. Audit studies have been useful in allowing us to test for discrimination in one aspect of government-public interaction, but future studies should combine audits with additional methodological tools (Gaddis, 2019; Lahey & Oxley, 2021; List, 2004).

Methodologically, the two studies together illustrate the importance of conducting replication or follow-up studies to determine if identified effects persist over time. In this context we lack a good explanation for why the effects differ across time, but temporal variation has been identified as an important avenue for research (Munger, 2019). Another important reason to conduct replications is that our ability to conduct high-powered meta-analyses depends on the presence of multiple studies that can be combined into a meta-analysis, and the incentive to conduct these studies will depend on the ability to publish their results. Journals rarely publish replication studies, and there is evidence that reviewers are particularly unlikely to support publication of replications with null results, which provides little professional incentive to conduct such research even if there is scientific rationale for doing so (Berinsky et al., 2021). As we demonstrate in this paper, findings may not replicate across time, and we should encourage the replication of prior studies to determine if there is temporal or other variation in causal effects.

Acknowledgments

We thank the reviewers and editorial team for their feedback on our manuscript. This study was part of the Center for Open Science's Systematizing Confidence in Open Research and Evidence (SCORE) project. All remaining errors are our own.

Notes

1. Throughout we use “white” to denote the non-Hispanic aliases used in our study, and as shown in Table 1, the non-Hispanic aliases used in our study were overwhelmingly perceived as white.
2. The difference in sample size partly reflects the fact that the original study also included Black aliases. We only tested for white/Hispanic ethnic differences because Einstein and Glick (2017) found no statistically significant evidence of discrimination against Blacks.
3. The inability to publish (and therefore lack of incentive to conduct) replication studies also makes it such that either insufficient studies for a meta-analysis exist or meta-analyses may feature studies that are not sufficiently similar to be considered together.
4. If one wants to identify discrimination based on names but is less concerned about attributing it to a specific attribute associated with a name, such as race, linguistic background, etc, then it is less important to know the precise attributes a name manipulates. In our context, we focus on the white versus Hispanic comparison as that was the focus of the original study (Einstein & Glick, 2017).
5. Racial resentment and Trump vote share are highly correlated; the pairwise correlation coefficient is 0.8007.
6. We use PHA scores as a measure of professionalism, but it is important to note the distinction between legislative professionalism (Squire, 1992) and professionalism as measured by PHAS scores. Legislative professionalism is measuring salary compensation, session length and staffer support. PHA scores meanwhile consider factors such as quality of housing stock. See: https://www.hud.gov/program_offices/public_indian_housing/reac/products/prodphas/prodphas-scores

References

- Abramitzky, R., Boustan, L., & Eriksson, K. (2020). Do Immigrants Assimilate More Slowly today than in the past? *American Economic Review: Insights*, 2(1), 125–41.
- Berinsky, A. J., Druckman, J. N., & Yamamoto, T. (2021). Publication Biases in Replication Studies. *Political Analysis*, 29(3), 370–384.
- Bischof, D., Cohen, G., Cohen, S., Foos, F., Kuhn, P. M., Nanou, K., Visalvanich, N., & Vivyan, N. (2021).

- Advantages, challenges and limitations of audit experiments with constituents. *Political Studies Review*, 20(2), 192–200.
- Blair, G., Cooper, J., Coppock, A., & Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, 113(3), 838–859.
- Brambor, T., Clark, W. R., & Golder, M. (2005). Understanding interaction models: Improving empirical analyses. *Political Analysis*, 14(1), 63–82.
- Butler, D. M. (2014). *Representing the advantaged: How politicians reinforce inequality*. Cambridge University Press.
- Butler, D. M., & Homola, J. (2017). An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments. *Political Analysis*, 25(1), 122–130.
- Coppock, A. (2018). Avoiding Post-Treatment Bias in Audit Experiments. *Journal of Experimental Political Science*, 6(1), 1–4.
- Costa, M. (2017). How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials. *Journal of Experimental Political Science*, 4(3), 241–254.
- Dafoe, A., Zhang, B., & Caughey, D. (2018). Information equivalence in survey experiments. *Political Analysis*, 26(4), 399–416.
- Einstein, K. L., & Glick, D. M. (2017). Does race affect access to government services? An experiment exploring street-level bureaucrats and access to public housing. *American Journal of Political Science*, 61(1), 100–116.
- Gaddis, S. M. (2017). How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies. *Sociological Science*, 4, 469–489.
- Gaddis, S. M. (2018). An Introduction to Audit Studies in the Social Sciences. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Springer International Publishing.
- Gaddis, S. M. (2019). Understanding the “How” and “Why” Aspects of Racial-Ethnic Discrimination: A Multimethod Approach to Audit Studies. *Sociology of Race and Ethnicity*, 5(4), 443–455.
- García, M. M., & Sadhwani, S. (2022). ¿Quién Importa? State Legislators and Their Responsiveness to Undocumented Immigrants. *Political Research Quarterly*, 0(0). <https://doi.org/10.1177/10659129221137825>
- Gomila, R. (2020). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, 150(4), 700–709.
- Grimmelikhuijsen, S., John, P., Meijer, A., & Worthy, B. (2019). Do Freedom of Information Laws Increase Transparency of Government? A replication of a field experiment. *Journal of Behavioral Public Administration*, 2(1), 1–10.
- Hamel, Brian T., and Derek E. Holliday. "Unequal Responsiveness in City Service Delivery: Evidence from 42 Million 311 Calls." Forthcoming in *Quarterly Journal of Political Science*
- Lahey, J. N., & Mosquera, R. M. (2022). Age and the labor market for Hispanics in the United States. *National Bureau of Economic Research Working Paper Series*, No. 30171. <https://doi.org/10.3386/w30171>
- Lahey, J. N., & Oxley, D. R. (2021). Discrimination at the Intersection of Age, Race, and Gender: Evidence from an Eye-Tracking Experiment. *Journal of Policy Analysis and Management*, 40(4), 1083–1119.
- Landgrave, M. (2020). Can We Reduce Deception in Elite Field Experiments? Evidence from a Field Experiment with State Legislative Offices. *State Politics & Policy Quarterly*, 20(4), 489–507.
- Landgrave, M. (2021). Do Politicians Ethnically Discriminate Against Hispanics? Evidence from a Field Experiment with State Legislative Offices. *Legislative Studies Quarterly*, 46(3), 621–636.
- Landgrave, M., & Weller, N. (2020). Do more professionalized legislatures discriminate less? The role of staffers in constituency service. *American Politics Research*, 48(5), 571–578.
- Landgrave, M., & Weller, N. (2022). Do Name-based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment. *Political Analysis*, 30(1), 142–148.
- List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics*, 119(1), 49–89.
- List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4), 773–793.
- Lowande, K., & Proctor, A. (2020). Bureaucratic Responsiveness to LGBT Americans. *American Journal of Political Science*, 64(3), 664–681.
- Munger, K. (2019). The Limited Value of Non-Replicable Field Experiments in Contexts With Low Temporal Validity. *Social Media+ Society*, 5(3).
- Seidel, J., & Xu, Y. (2016). *MHTEXP: Stata module to perform multiple hypothesis testing correction procedure*.
- Smith, C. W., Kreitzer, R. J., & Suo, F. (2020). The dynamics of racial resentment across the 50 US states. *Perspectives on Politics*, 18(2), 527–538.
- Squire, P. (1992). Legislative professionalization and membership diversity in state legislatures. *Legislative Studies Quarterly*, 17(1), 69–79.
- Thomas, J. C. (1986). The personal side of street-level bureaucracy: Discrimination or neutral competence? *Urban Affairs Quarterly*, 22(1), 84–100.
- Zittel, T., Louwse, T., Helboe Pedersen, H., & Schakel, W. (2021). Should we conduct correspondence study field experiments with political elites? *International Political Science Review*, 0(0). <https://doi.org/10.1177/01925121211026489>.

Appendices

Contents

Appendix A – Email Script	11
Appendix B – Additional Analysis.....	12
Appendix C – Anonymous Pre-Registration Plan.....	18
Appendix D – Additional Name Perceptions & Analysis	24

Appendix A – Email Script

The email script we used is presented in **Appendix Figure 1**. This is the exact same script used by Einstein and Glick (2017).

Appendix Figure 1 – Email Text

Email Title: Help Applying

Email Body:

Hello,

My name is [Name] and I'm trying to figure out how to apply for public housing. I believe I may be eligible.

Can you direct me to information about applying for public housing here? I also heard there might be a wait list for this program? How long is it?

Thanks, [Name]

Appendix B – Additional Analysis

In the main manuscript we present the average treatment effect of having bearing a Hispanic name, versus a white name, on (a) whether any reply is received, (b) does the reply occur within 24 hours, and (c) reply friendliness. We present the same results in tabular form in Appendix Table 1 below. We provide tabular results for main text Figures 2-3 in Appendix Tables 2-3.

The main manuscript presented results from a linear probability model. In addition to the linear probability model, we present the main analysis using logit estimation in Appendix Table 4. We also adjust for multiple comparisons in Appendix Table 5. To adjust for multiple comparisons we use the MHTEXP stata package (Seidel & Xu, 2016) to implement the proposed solution by List, Shaikh, and Xu (2019) to multiple hypothesis testing. Across specifications we fail to find statistically significant evidence that discrimination that Hispanics are discriminated against using both the linear probability model and a logit.

Appendix Table 1 – Average Treatment Effect of Hispanic Alias, Tabular Results

VARIABLES	(1) Any Reply	(2) Reply within 24 Hours	(3) Friendly Re- ply
Hispanic Alias	-0.022 (0.022)	0.028* (0.015)	-0.008 (0.020)
Constant	0.530*** (0.016)	0.121*** (0.011)	0.282*** (0.014)
Observations	2,099	2,099	2,099
R-squared	0.000	0.002	0.000

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 2 – Effect of Hispanic Name by Racial Resentment, Tabular Results

VARIABLES	(1) Any Reply	(2) Reply within 24 Hours	(3) Friendly Reply
Hispanic Alias	-0.023 (0.022)	0.027* (0.015)	-0.009 (0.019)
2016 Racial Resentment, Z-Score	0.031 (0.021)	0.000 (0.015)	0.001 (0.019)
Hispanic Alias * Racial Resentment	-0.013 (0.022)	0.015 (0.015)	-0.012 (0.019)
2016 Trump Vote Share, Z-Score	-0.087*** (0.019)	-0.021 (0.013)	-0.056*** (0.017)
Percent Hispanic - State, 2019 5YR ACS	-0.002* (0.001)	-0.001** (0.001)	-0.002*** (0.001)
Constant	0.552*** (0.020)	0.141*** (0.014)	0.312*** (0.018)
Observations	2,099	2,099	2,099
R-squared	0.018	0.006	0.018

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 3 – Effect of Hispanic Name by PHA Professionalism, Tabular Results

VARIABLES	(1) Any Reply	(2) Reply within 24 Hours	(3) Friendly Reply
Hispanic Alias	-0.032 (0.023)	-0.016 (0.020)	-0.016 (0.020)
PHA Professionalism, Z-Score	0.031** (0.016)	0.003 (0.014)	0.003 (0.014)
Hispanic Alias * PHA Professionalism	-0.002 (0.023)	0.005 (0.020)	0.005 (0.020)
Percent Hispanic, 2018 5YR ACS	0.000 (0.001)	-0.000 (0.001)	-0.000 (0.001)
Population, 2018 5YR ACS	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Constant	0.526*** (0.018)	0.286*** (0.016)	0.286*** (0.016)
Observations	1,946	1,946	1,946
R-squared	0.005	0.001	0.001

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 4 – Average Treatment Effect of Hispanic Alias, Logit Results

VARIABLES	(1) Any Reply	(2) Reply within 24 Hours	(3) Friendly Re- ply
Hispanic Alias	-0.087 (0.087)	0.238* (0.128)	-0.042 (0.097)
Constant	0.121* (0.062)	-1.980*** (0.095)	-0.935*** (0.069)
Observations	2,099	2,099	2,099

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 5 – Average Treatment Effect of Hispanic Alias, Adjusted for Multiple Hypothesis Testing

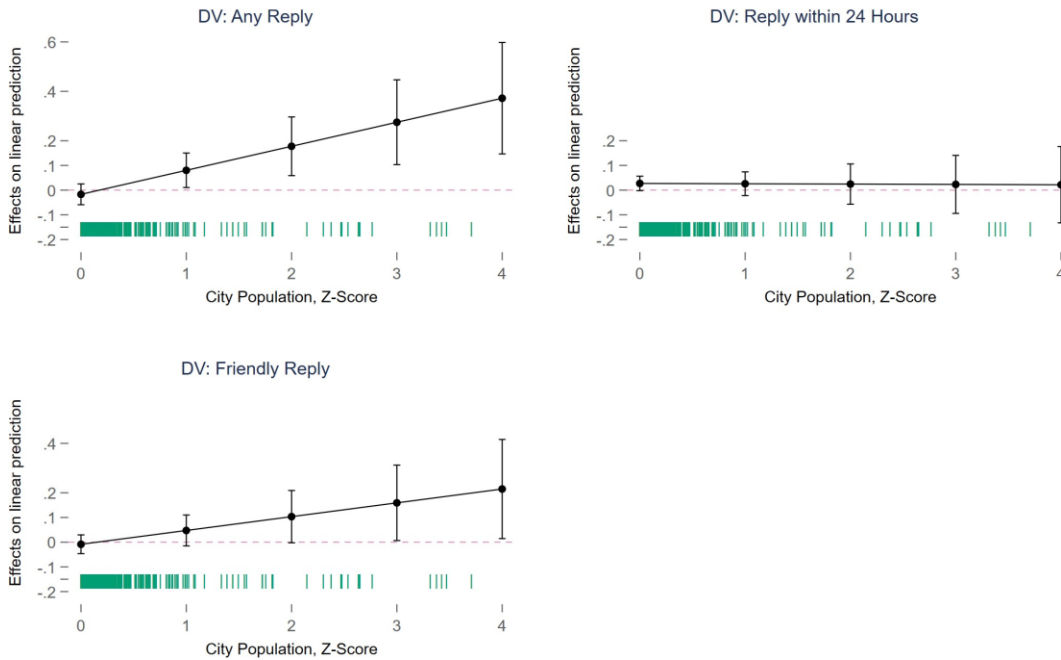
	Any Reply	Reply within 24 Hours	Friendly Re- ply
Hispanic Alias Treatment	0.022	0.028	0.008
p-value	0.317	0.066	0.661
sharpened q-value	0.952	0.197	1.000
mhtexp FWER p-value	0.635	0.197	0.661
Sample size	2,099	2,099	2,099

Note: Appendix Table 5 was computed using the mhtexp package developed by (Seidel & Xu, 2016) to implement (List et al., 2019)'s proposed solution to the problem of multiple hypothesis testing in experimental social science.

In response to the suggestions of reviewers we estimated discrimination by city type as reported in the main appendix, which lead us to consider city population as a moderator as well. In **Appendix Figure 2** we estimate the moderation of the Hispanic alias by city population size. City population size was standardized (z-score) for ease of interpretation. Notably the z-score range extended to 34 because of the inclusion of a few mega-cities, but we restrict our analyses to 4 because this better reflects the bulk of cities in our sample. As noted below, Hispanic aliases are more likely to be replied to, and be replied to in a friendly manner, in larger cities compared to their white counterparts. There is no observed moderation in speed of reply.

We urge caution against interpreting **Appendix Figure 2** causally because population size is bundled with a multitude of factors. That is, we are not sure if the moderation is due to population per se, or unobserved factors bundled with population size such as social capital or cosmopolitan attitudes. With our present data limitations, we are unable to formally test what population size is really capturing.

Appendix Figure 2 – Effect of Hispanic Name by City Population Size



Note: Controls include 2016 Trump Vote Share, State Racial Resentment, and 2019 State Hispanic Share. n=2,099.

In the main analyses we account for post-treatment bias in analyzing friendly replies by assuming that unobserved replies would have been unfriendly (column 2). As an additional robustness check we re-run the average treatment effect in **Appendix Table 6** assuming that unobserved replies would have been friendly (column 3) and subset to observed replies only (column 1). While the direction of the effect differs across models, none are statistically significant.

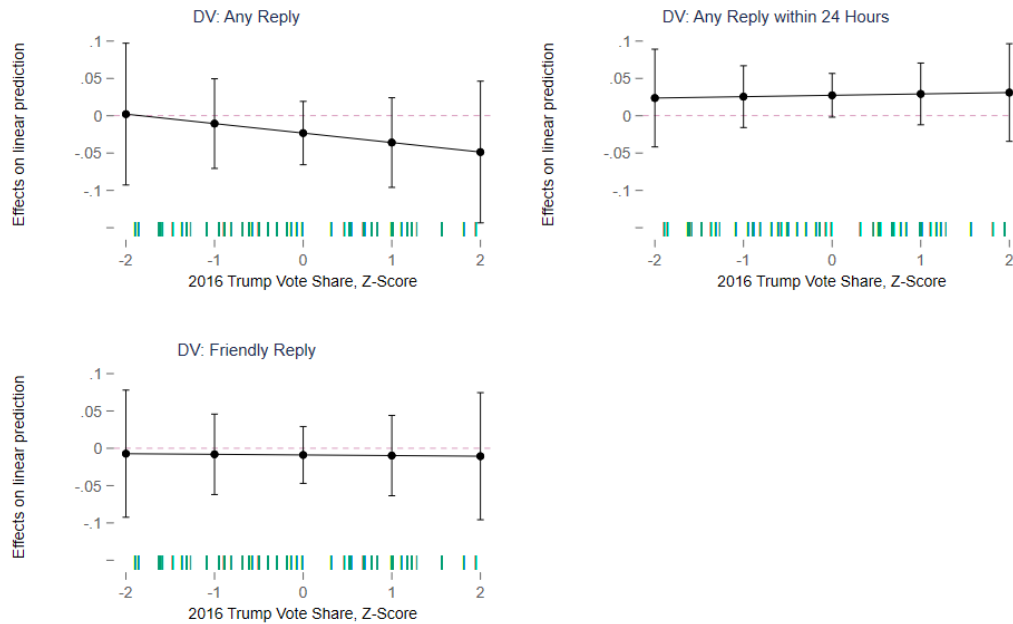
Appendix Table 6 – Average Treatment Effect of Hispanic Alias on Friendly Reply, Adjusted for Post-Treatment Bias

VARIABLES	(1) Conditional Reply	(2) Assume Unfriendly reply	(3) Assume Friendly reply
Hispanic Alias	0.006 (0.030)	-0.008 (0.020)	0.013 (0.019)
Constant	0.532*** (0.021)	0.282*** (0.014)	0.752*** (0.013)
Observations	1,090	2,099	2,099
R-squared	0.000	0.000	0.000

Standard errors in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

We test the potential moderation of Hispanic names by Trump vote share in **Appendix Figure 3**. We find no statistically significant effects across any of our outcome variables.

Appendix Figure 3 – Effect of Hispanic Name by Trump Vote Share



Note: Controls include State-level Racial Resentment, and 2019 State Hispanic Share. n=2,099.

Appendix C – Anonymous Pre-Registration Plan

DESCRIPTION:

The claim selected for replication from Einstein & Glick (2017) is that, all else equal, bureaucrats will be less friendly to Hispanics than they are to whites. This reflects the following statement from the paper's abstract: "Moreover, we observe substantial differences in email tone. Hispanic housing applicants were 20 percentage points less likely to be greeted by name than were their black and white counterparts." The claim is tested by e-mailing public housing authorities at their publicly available e-mail addresses (or online contact forms) using an audit study design. Each housing authority received an e-mail on one of two days in the same week during 2014. The authors e-mailed all public housing authorities that could plausibly be matched with a core city in a metropolitan or micropolitan area. Each housing authority was randomly assigned (via random number generator) to receive an e-mail from one of six different accounts with putatively white, Hispanic, and black names. For each racial/ethnic group, the authors chose one male and one female name. The dependent variable, Friendliness, is measured as whether the e-mailer is addressed by proper name. This is the same definition as used by Einstein and Glick (2017). The coefficient of interest for testing the claim is on the Hispanic name indicator. The Hispanic name indicator variable is substantial, negative, and highly significant (coefficient = -0.82, SE = 0.22, $p < .01$)

HYPOTHESIS: The e-mailer will be less likely to be addressed by proper name in the Hispanic name condition than in the white name condition.

STUDY TYPE:

Experiment

In this experimental study we will use multiple aliases to contact public agencies (Housing agencies and perhaps others) to determine whether the Hispanic aliases are less likely to be addressed by their proper name than the white aliases.

We will collect the email address from each county's (or other jurisdiction's) public housing agency using the federal Housing and Urban Development website. We will then randomize whether an agency receives an email from a Hispanic or white alias in which we request help with determining eligibility for federal housing assistance. We will block randomize within each U.S. state to ensure that there is an equivalent number of Hispanic and white names contacting the housing agencies within each state.

The names we use will be Joe Garcia or Michael Hernandez for the Hispanic names and John Ryan or Nicholas Smith for the white names. These names were chosen based on prior studies in which we pre-tested names across a variety of factors and these names vary the Hispanic vs. white dimension while being relatively similar across perceived class and migrant status, which helps to ensure that any differences in observed behavior are a result of the intended manipulation and not some unintended one.

After sending the emails to each housing agency we will wait one week for responses. Upon receipt of email responses we will code each response based on whether the respondent personalizes their response by using the emailer's name in the response, which means use of either a first or last name.

BLINDING:

This study is single-blind, because the housing agencies we contact will not know that they are part of a study. This reduces the possibility that housing agencies are responding because they believe they are part of an academic study, which would reduce the experimental realism of the study. To make it less likely that our study is detected by the housing agencies we are using four aliases that were not used by Einstein and Glick (2017), and we send our emails using a Gmail address, which is common and unlikely to elicit suspicion. From the paper: "It is also plausible that the prominence of the Fair Housing Act made respondents suspicious that they were being studied by academics or the government. While we cannot disprove this possibility, we believe that several pieces of evidence suggest that it is unlikely..." (p. 113)

STUDY DESIGN:

The original experiment was a 3 (white, Hispanic, black) x 2 (male, female) between subject design. The focal analysis collapses across gender, and the selected comparison is white names compared to Hispanic names. Given the intent of the replication study and the possible sample size limitations, we modified the original study by choosing to only use male and white/Hispanic names. In the initial study the authors did not find a difference based on male/female names so we eliminate this factor, and we eliminate the Black aliases as those are not a focus of our replication study. We use two different white and Hispanic names to ensure that our results are not driven by differences across only two names. These changes will also increase the number of housing authorities assigned to the treatments, which will increase the power of our study. "Each housing authority was randomly assigned (via random number generator) to receive an e-mail from one of six different accounts with putatively white, Hispanic, and black names (Table SI1 in the supporting information shows balance in the average demographics for the communities assigned to each treatment). For each racial/ethnic group, we chose one male and one female name to address any possible gender interactions emerging from the disproportionately female-headed house-hold composition of public housing." (p. 104)

RANDOMIZATION:

We will use block randomization at the state level. Randomization will be conducted using a random number generator, and we will assign each housing authority to receive an email from one of the four aliases (two white and two Hispanic names) chosen for the experiment.

DATA COLLECTION PROCEDURES:

“We use the case of affordable housing and our own audit-style experiment of over 1,000 housing authorities to test whether street-level bureaucrats discriminate when citizens attempt to access substantive programs and services...” (p. 101)

“We e-mailed public housing authorities at their publicly available e-mail addresses (or online contact forms) using an audit study design. Each housing authority received an e-mail on one of two days in the same week during 2014 (footnote: The e-mails were sent on two separate days as a consequence of Google Mail batching limits.). We e-mailed all public housing authorities that could plausibly be matched with a core city in a metropolitan or micropolitan area (n = 1,017).” (p. 103)

The following is an excerpt from the paper related to the email text. Note that we asked the authors about the subject line of the email but did not receive an answer:

Our e-mail text was the following:

Hello,

My name is X and I'm trying to figure out how to apply for public housing. I believe I may be eligible.

Can you direct me to information about applying for public housing here? I also heard there might be a wait list for this program. How long is it?

Thanks, X

In December and January 2019/20, we will collect the email addresses of all housing authorities in the United States that are listed on the Housing and Urban Development website (https://www.hud.gov/program_offices/public_indian_housing/pha/contacts). According to the information posted on the Housing and Urban Development website (<https://www.huduser.gov/periodicals/ushmc/spring95/spring95.html>) there are about 3,400 public housing agencies in the United States.

We will contact all of the housing authorities with listed email addresses, but we will not limit ourselves to the metro/micropolitan restriction from the original study. The original study does not provide sufficient information to identify which housing agencies the authors contacted and therefore our plan will be to contact all of the ones whose email addresses we can identify via the HUD website. The theory does not specify any reason to limit analysis to larger areas, and as a result of random assignment of aliases PHAs we expect observed and unobserved variables to be equivalent between the two experimental groups and therefore we can simply compare average rates of response without concern for statistical adjustments to the data.

In some of the analyses in the original paper the authors did adjust for demographic differences across the public housing agencies by using data from the American Community Survey at the “place” level as done in the original Einstein and Glick paper. This will reduce our sample size as some of the public housing agencies are from locations that do not match to a specific “place.”

The analysis in this replication is conducted only on the email responses and therefore the sampling plan requires us to consider the response rate to the emails sent out as part of the audit study. We assume that the response rate will be similar to that in Einstein and Glick, which was just below 60%. Therefore, if we contact 3,500 agencies we would expect to receive about 2,000 responses, which is sufficient for our analysis as shown in the power calculation section.

SAMPLE SIZE:

The initial target analytic sample size is 738 agencies. If a statistically significant effect is not observed after the first round of data collection, a second round will begin. The second round of data collection will increase the analytic sample with data from an additional 923 agencies for a pooled sample of 1,661 agencies.

SAMPLE SIZE RATIONALE:

Power calculations were done in accordance with the guidelines of the Social Sciences Replication Project (SSRP). The first round of data collection achieves 90% power to detect 75% of the original effect size. The pooled sample, if necessary after testing the effect on the first round of data, achieves 90% power to detect 50% of the original effect size. Details on how the power analysis was performed are available here

STOPPING RULE:

The planned analytic sample size is 738 agencies. As discussed earlier, we will email all of the PHAs for which we can find valid email addresses. After collecting all of the email responses the planned analyses will be run on randomly-selected subsets of 738 agencies if we have more responses than that. If a significant effect is found, sampling stops. If a significant effect is not found, then we will use the additional responses that were collected to increase the sample with an additional 923 agencies, for a pooled analytic sample of 1,661 agencies. Sampling will stop after the second round of data collection regardless of a significant effect.

MANIPULATED VARIABLES:

Each housing authority was randomly assigned (via random number generator) to receive an e-mail from one of six different accounts with putatively white, Hispanic, and black names (Table S11 in the supporting information shows balance in the average demographics for the communities assigned to each treatment). For each racial/ethnic group, we chose one male and one female name to address any possible gender interactions emerging from the disproportionately female-headed house-hold composition of public housing.” (p. 104) “The six names were as follows: Brett Smith, Emily Smith (white); Tyrone Johnson, Shanice Johnson (black); and Santiago Martinez, Gabriela Martinez (Hispanic). For the white and black e-mailers, we chose names that were among the 20 most distinctively black and white names in Levitt and Dubner (2010).” (p. 104).

NOTE: that if a direct replication is performed with a sample of U.S. public housing agencies again, it might be advisable to change the names of the email senders in order to avoid suspicion, given that the original study was conducted recently (see below for guidelines on name creation).

While we generally followed previous studies’ practices in selecting names, we made a couple of important adjustments. We were, when possible, attentive to the age that names implied. One name prominently used in a recent audit study—Deshaun—came into use almost exclusively after 1970 according to data from the Social Security Administration.” (p. 104)

We will choose two white and two Hispanic names for our study. Rather than using both male and female names we will pick only male names, because the original study did not find a difference based on the gender of the name and by eliminating it as a variable we more narrowly focus on race. We will use Joe Garcia and Michael Hernandez for the Hispanic name and for white names we will use John Ryan or Nicholas Smith. In a prior name perception study conducted on Mturk we confirmed that a person with Hispanic names are much less likely to be considered white than the white names.

MEASURED VARIABLES:

“In the statistical models below, we include a variable indicating whether we had to use the HUD web-site to find an e-mail address (the Hard Email variable).” (p. 105)

“Friendliness is a less widely used variable, and quantifying it is somewhat more challenging. We use what we believe to be the most easily comparable (and least subjective) measure across e-mails: whether the e-mailer is addressed by proper name. We were lenient in coding “yes.” A named salutation could be as casual as “Hi Brett” or as formal as “Dear Ms. Martinez.”” (p. 105)

“We obtain the proportion white, black [footnote 9: “We calculate the proportion non-Hispanic white and black.”], and Hispanic at the “place” level from the American Community Survey’s 2012 5-year estimates. In addition, we incorporate other relevant demographic controls from the census into our statistical models: city poverty rate and population.” (p. 106)

INDICES:

There are no indices for this study

STATISTICAL MODELS:

The authors' publicly available materials include the following note in the readme: "Privacy concerns restrict us from publicly posting the full data...Instead we include a slightly modified dataset and code that will enable public replication of some results in full...and approximations of others (the results that do require demographics)."

- The blinded data is found here.
- The codebook for the blinded data is found here.
- The script to analyze the blinded data is found here.
- The full set of public materials for this paper have been gathered on the OSF is here.

The specification for the focal logit model is found in the third column of Table 1 below. Excerpts from the paper related to the focal model are below the table:

TABLE 1 Logit models for the three Dependent Variables, with All Three Racial Groups and Other Demographic Controls Variables

	(1) Response	(2) Response in 24 Hours	(3) Proper Name Greeting
Black Name	0.13 (0.17)	0.29 (0.26)	-0.21 (0.22)
Hispanic Name	-0.21 (0.16)	0.10 (0.25)	-0.82** (0.22)
Percent Black	-0.31 (0.52)	-0.17 (0.87)	0.71 (0.74)
Percent Hispanic	-0.32 (0.43)	-0.92 (0.69)	1.33* (0.63)
Poverty Rate	-1.65 (1.21)	1.28 (2.00)	-0.93 (1.71)
Log Population	0.13 (0.07)	-0.01 (0.10)	-0.04 (0.09)
Hard Email	-0.46** (0.14)	-0.14 (0.22)	-0.42* (0.19)
Constant	-0.50 (0.75)	1.33 (1.16)	0.88 (0.99)
Observations	978	551	549
Log Likelihood	-650.7	-286.8	-367.5

Note: Standard errors in parentheses. Base category is white. Demographic variables from 2012 American Community Survey.
**p<.01, *p<.05.

"The model we report in the text, in the third column of Table 1, further demonstrates the negative impact a Hispanic name has on receiving a named greeting...The Hispanic name indicator variable is substantial, negative, and highly significant." (p. 108)

We will conduct multiple analyses in this paper. The first analysis is a simple test of whether the Hispanic names are less likely to receive a proper name greeting than the white names. We will conduct this using a two-tailed difference in proportions test as done by Einstein and Glick by pooling the two Hispanic names together and the two white names together... In the Einstein and Glick paper. The outcome variable is the proportion of all responses that are considered friendly; the numerator is the number of friendly responses and the denominator is all received responses.

We will also conduct the same (or substantially similar) analyses as done by Einstein and Glick in which they use a logit model to estimate the probability of a response from a housing agency adjusting for the ACES place-area estimates for percent black, percent Hispanic, poverty rate, and the log of the state's population.

If our audit study yields more than the 738 responses needed based on the power analysis, then we will conduct the analyses on a random subset of 738 responses. Furthermore, we can draw multiple random samples from all of the responses and conduct the analysis on each sample. This provides a good estimate of the variability in the estimated treatment effect.

Important note is that in their initial paper they did not adjust their analysis for post-treatment bias, which is necessary when analyzing the text of responses because it is not random whether a housing agency responds or not. A variety of approaches are available for adjusting for the non-random response rate (Coppock 2018), but the most straightforward adjustment is to assume that all non-responses would have been non-friendly -- that is failed to use a proper name in the response. This adjustment leads to a smaller estimated treatment effect of about 12% rather than 20%, because this adjustment adds all of the initial emails sent to the denominator but does not increase the number of friendly responses included in the numerator. We will conduct secondary analyses using this post-treatment bias correction.

TRANSFORMATIONS:

In our replication study we need to code the friendliness of a response. We are following the coding criteria established by Einstein and Glick, which coded a response as friendly if the response used either the first or last name of the recipient in the greeting of the email.

In our secondary analysis we will adjust for non-random, non-response as suggested by Coppock (2019) by using both the responses we received and those we did not receive. The adjustment requires that we assume all of the non-responses would have been non-friendly, and therefore these responses are counted as part of the denominator but not the numerator of the proportion of friendly responses. That is, the outcome variable becomes $(\# \text{ friendly responses}) / (\# \text{ total emails sent})$, and we examine the difference in this proportion between the white and Hispanic names.

INFERENCE CRITERIA:

Criteria for a successful replication attempt for the SCORE project is a statistically significant effect ($\alpha = .05$, two tailed) in the same pattern as the original study. For this replication attempt, this criteria is met by a negative and statistically significant coefficient on the 'Hispanic Name' indicator in the focal regression model.

DATA EXCLUSIONS:

Note that the focal model (column 3 of Table 1) only includes the sample of housing authorities that responded to the email in the first place. We will also code the emails for whether they appear to have come as a result of an "automatic" response. We will code this based on the presence of text such as "automatic reply," "will respond later," "out of office," etc. We will exclude automatic replies from our analysis.

DEVIATIONS AND OTHER INFORMATION:

We have two planned deviations from the original study. First, as discussed elsewhere we intend to only use male names in the replication.

Second, we will attempt to collect all available housing agencies via the HUD website. Einstein and Glick do not make it clear how they sampled their housing agencies or which agencies were included/excluded. Therefore, we cannot ensure that our sample of agencies is the same as their sample. However, given random assignment of the treatment the exact sample of housing agencies is not crucial, because randomization ensures that the agencies receiving the Hispanic and white names are equivalent in expectation. Additionally, the theory underlying their experiment concerns the behavior of public agencies and is not specific to housing agencies or the agencies they studied. Therefore, differences in the sampled agencies should not matter for the replication project.

Appendix D – Additional Name Perceptions & Analysis

In addition to pre-testing names based on perceived race/ethnicity and nativity (see above), we pretested names based on perceived socio-economic status and gender. As shown below, purportedly Hispanic names are more likely to be viewed as working or lower middle class than their white counterparts. This is not a flaw in our design so much as it is a reflection that Hispanics in the contemporary United States are viewed as coming from a lower SES than whites, all else equal. With minor variation, all four names selected were overwhelming perceived as being male.

Appendix Table 7 – Perceived Socio-Economic Status by Name

Race/Ethnic-	Name			
	Nicholas Smith	John Ryan	Joe Garcia	Michael Hernandez
Working Class	16.92%	16.42	22.73	40.30
Low Middle	4.62	10.45	25.76	19.40
Middle Class	64.62	61.19	36.36	32.84
Upper Class	13.85	11.94	15.15	7.46

Appendix Table 8 – Perceived Gender Identity by Name

Race/Ethnicity	Name			
	Nicholas Smith	John Ryan	Joe Garcia	Michael Hernandez
Male	93.85%	98.53	93.94	97.01
Female	1.54	0	6.06	1.49
Other	4.62	1.47	0	1.49

It is possible that name-specific effects exist. To address this concern, name specific responses are displayed in the below coefficient plot. As noted by some, while John Ryan and Nicholas Smith are both perceived as white, they differ in perceived nativity/citizenship. There is minimal evidence to suggest that nativity/citizenship influenced responses between the two aliases. Results from a linear probability model find that Nicholas Smith was slightly less likely to receive a reply (coef: -.0490135, p-value = 0.114) than John Ryan, but there is no evidence of a difference in friendliness (coef: -.0149162, p-value = 0.594).

Appendix Figure 4 – Responses by Name

