What's In a Name? An Experimental Analysis of Signaling Race, Ethnicity, and Gender Using Names*

Matthew Hayes† Elizabeth Mitchell‡

Abstract

A growing body of research uses names to cue experimental subjects about race, ethnicity, and gender. Yet, researchers have not explored the myriad of characteristics that might be signaled by these names. We developed two large, publicly available databases of the attributes associated with common American first names. The first provides perceived racial distinctiveness scores for 1,000 names. The second includes ratings of 89 names on 21 additional characteristics, including personal traits as well as nonracial group memberships. We show that the traits associated with first names vary widely, even among names associated with the same race and gender. Researchers using names to signal group memberships are thus likely cuing a number of other attributes as well. We demonstrate the importance of name selection by replicating DeSante (2013). We conclude by outlining two approaches researchers can use to choose names that successfully cue race (and gender) while minimizing potential confounds.

*The authors would like to thank Alexa Bankert, Diana Z. O'Brien, Leslie Schwindt-Bayer, Beth Simas, Randy Stevenson, Rick Wilson, and numerous conference participants for valuable feedback and suggestions. We also thank Christopher DeSante for sharing his original experimental materials with us.

†Assistant Professor, Department of Political Science, Rice University.

‡Ph.D. Candidate, Department of Political Science, University of California, Berkeley.

Discrimination on the basis of race, gender, and ethnicity remains a serious problem in American public life, yet measuring such discrimination has proven difficult due to social desirability bias. In recent years, scholars have turned to audit studies and survey experiments that use a minimal cue of race, ethnicity, or gender: individuals' names. Doing so has allowed scholars to shed light on potential biases by employers (Bertrand and Mullainathan, 2004), elected officials (Butler and Broockman, 2011), and ordinary citizens (DeSante, 2013) without the potential confound of social desirability bias.

Despite the prevalence with which names are used to indicate social group memberships, it is not clear that group memberships of interest are the only attributes triggered when people encounter particular names. Fryer Jr and Levitt (2004), for example, raise the concern that experiments using distinctive African American names may also be cuing subjects about individuals' class backgrounds, which could potentially compromise these studies' internal validity (but see Butler and Homola, 2017). Thus, studies showing racial bias might be detecting both taste-based discrimination against minority groups and bias on the basis of perceived traits of those group members, such as their class or educational status.[1]

In this study, we provide the most in-depth assessment of the groups, traits, and stereotypes that are invoked by the use of common American first names across different racial, ethnic, and gender categories. Using ratings from online respondents, we create two datasets. In our first dataset, we measure the perceived racial distinctiveness of 1,000 common American first names across the four largest racial and ethnic groups in the United States. In our second dataset, we measure not only how people perceive the racial distinctiveness of 89 names, but also how they perceive those names on an additional 21 attributes such as warmth and competence. Using these two datasets, we find that the traits associated with first names vary widely, even within the same race and gender, suggesting that researchers using names to cue particular group memberships are likely cuing other attributes as well. We then demonstrate the importance of selecting a race or gender prime by replicating the findings of DeSante (2013) using purposefully-selected first names. Finally, we propose two methods with which researchers can choose names that cue race and gender while minimizing potential confounds.

[1]It is important to note that such perceptions are, in and of themselves, a manifestation of racial bias. Indeed, our data show that African American names are rated lower than white names on a number of desirable traits, such as honesty, competence, hardworking, and intelligent. This effect is especially pronounced for African American women's names. For more detailed analysis, see Section S2.2 in the Supplemental Information.

Data and methods

To assess what traits are cued when individuals encounter racially- and gender-distinct names, we developed two publicly available databases of the attributes associated with common American first names. To build these datasets, we began by obtaining a list of given names among people born in the United States between the years 1955 and 1990 from the U.S. Social Security Administration (SSA). We merged the SSA data with data from Tzioumis (2018), which draws on proprietary mortgage loan data to provide the distribution of self-reported race for 4250 common first names. To identify the gender of given first names, we used information on the gender of name recipients included in the SSA data.

Next, we selected the 1,000 names that represent the largest numbers of Americans in each of four major racial and ethnic groups–white, black, Latino, and Asian American–born between 1955 and 1990. We estimated the number of nameholders in each group by multiplying the number of Americans born with each name from the SSA data by the percent of nameholders belonging to each racial group from the Tzioumis (2018) data. We then selected the 125 names that accounted for the most members of each race-gender group. These 1,000 names represent our main sample for measuring racial distinctiveness.

Because of the prohibitive costs associated with collecting information on a wide range of attributes for our list of 1,000 names, we also created a second, smaller sample of names on which to measure a wider range of characteristics. In the interest of selecting names that could be used to cue a particular racial or ethnic group, we limited the data to names that were likely to be seen as distinctive to one of the racial or ethnic groups in the data [2]. Based on the availability of common distinctive names, we limited this analysis to the categories Hispanic/Latino, Non-Hispanic White, and Non-Hispanic Black. We then randomly sampled from the list of racially distinctive names so that we had 10 names from each race for each gender. This list of 60 names formed our initial sample. For the purposes of the replication study described below, 29 further names were selected to increase the attribute diversity of our sample of white and black women's names. The additional black women's names were selected through further sampling using the same procedure described above. The additional white women's names were selected via sampling with an eye towards increasing socio-economic diversity (Figlio, 2005; Levitt and Dubner, 2006). The full set of names we used in this study can be found in Table S9 in the Supplemental Information.

[2] Distinctive names were defined as names for which a proportion above some threshold of people with that name fall into a single racial or ethnic group. The threshold proportion differed by group: for the initial sample, names were considered distinctively Hispanic/Latino if at least 50% of name recipients were Hispanic/Latino, black names if at least 40% of name recipients were black, and white names if at least 80% of name recipients were white.

We began surveying respondents recruited from Amazon's Mechanical Turk service in August 2018 and respondents recruited from Lucid Theorem in January 2020. Rolling data collection has continued periodically to the present. Each subject is asked to rate a number of randomly selected names from one of the two name samples described above. Respondents from Mechanical Turk provided ratings for the smaller sample of distinctive names; to date, we have collected a total of 9,000 respondents from Mechanical Turk, which represent 130,000 ratings of 89 names. Respondents from Lucid Theorem rated the larger sample of 1,000 names; to date, 2,149 respondents have provided 258,000 ratings.

For each name, respondents were asked to respond to either "How well does the word [TRAIT] describe the name [NAME]" or "How likely is it that someone named [NAME] is [GROUP]." For the larger sample of 1,000 names, respondents were asked only about how likely it was that each name belonged to a particular racial or ethnic group. The additional trait and group ratings for the smaller sample of 89 names can be found in Figure S4 in the Supplemental Information. Each trait or group rating was scored on a 5-point scale ranging from "Not well/likely at all" (1) to "Extremely well/likely" (5).

After collecting these ratings, we estimated each name's score on each of the different attributes using a mixed effects model. We predict attribute ratings with fixed effects for name and random effects for each rater. We interpret the coefficient on a name's fixed effect (i.e. the intercept estimated for that name) as that name's level of the attribute in question. This approach allows us to estimate, for example, a name's perceived competence by combining all the ratings of that name's competence while accounting for individual raters' tendency to rate all names as competent. It also allows us to calculate standard errors for our estimates that take account of raters providing multiple ratings.

Racial Associations of Common Names

We first present the results of the 1,000 name sample rated by Lucid Theorem respondents on four attributes: their likelihood of being held by someone who is white, black, Hispanic, or Asian American. Figure 1 shows the rated likelihood of each name belonging to each race by the true likelihood of each name belonging to each race, as measured by the percentage of nameholders belonging to that race in the Tzioumis (2018) data. The positive slopes in each of the panels of Figure 1 suggest that our raters, in general, perceive the race of names accurately: the larger the proportion of people with a name belong to a racial group, the likelier our raters were to say someone with that name belonged to that race.

In a dataset available in the Supplemental Information, we assign each name a perceived race based on the racial group that received the highest likelihood rating for that name. We also assign each name a racial

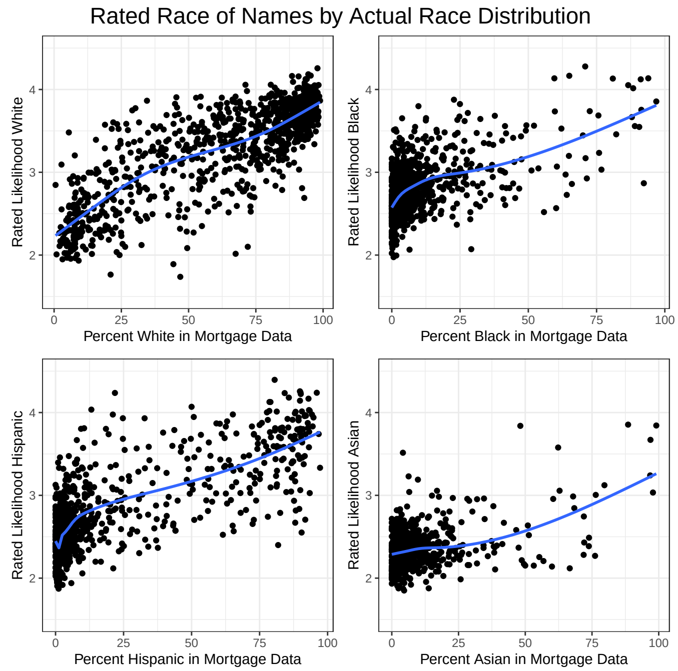

Figure 1: Raters accurately perceive the racial associations of common American names.

distinctiveness score based on the difference in rating between the highest and second-highest racial group. For example, the name Alberto has a Hispanic rating of 4.01 and a rating of 2.59 for its second-highest racial group, white. The name Alberto is therefore classified as Hispanic with a distinctiveness score of 1.42. This allows researchers to identify the names that most clearly cue membership in each racial group. We discuss recommendations for the use of this data in the Recommendations section below.

Intra-group variation in ratings: Replication study

The data presented in the previous section can help researchers choose names that communicate membership in a racial group. However, names that are racially distinctive are often distinctive along other dimensions as well. In this section, we demonstrate that names selected for their racial attributes can also cue other traits. A dataset available in the Supplemental Information contains ratings of 89 names on 25 attributes, including race, gender, class, and partisanship, as well as traits like competence and likability. As is discussed below, even names that are seen as clearly within the same race and gender group vary widely on other attributes. To illustrate the implications of this variation, we partially replicate and extend the findings of DeSante (2013), a case in which names that vary in nonracial traits produce substantively different results.

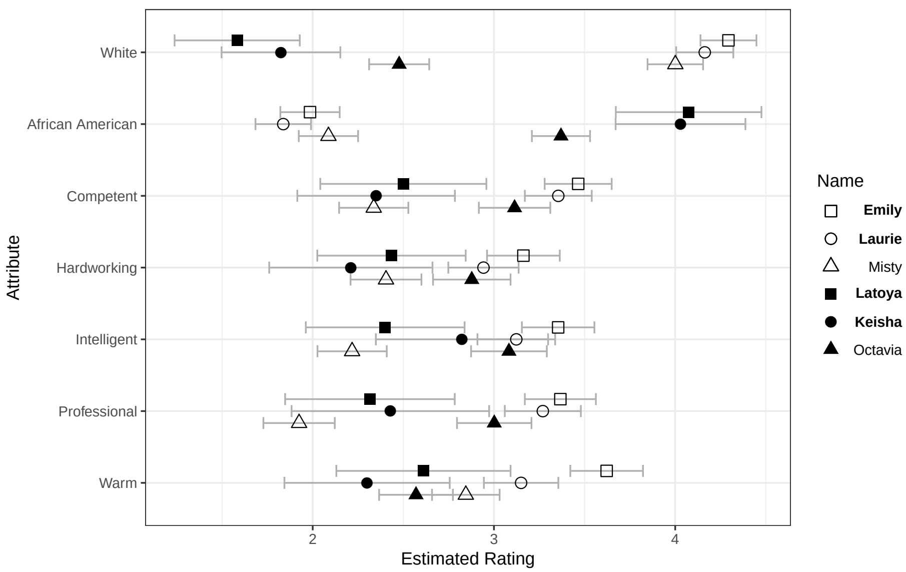

Figure 2: Average trait and group ratings by first name for selected women's names

In his article, DeSante (2013) presented respondents with pairs of hypothetical welfare applicants and asked them to allocate welfare funds to the applicants they viewed, generally finding evidence for racial discrimination. DeSante used the names Emily and Laurie to cue respondents to a white applicant, and Latoya and Keisha to cue respondents to a black applicant. Figure 2 presents ratings of DeSante's names on several key dimensions based on responses from our Mechanical Turk sample. From our data collection, it is clear that the names DeSante chose are good options for signaling race. Emily and Laurie are readily identified as white. Latoya and Keisha are similarly well-suited to signal that an individual is African American. These names, however, likely trigger additional traits and considerations.Figure 2 shows that Emily and Laurie are rated as more competent, hardworking, intelligent, professional, and warm than either Latoya and Keisha. While these differences in evaluation likely indicate racial bias in and of themselves, they also suggest a more complicated mechanism than simple racial phenotype for the racial discrimination found by DeSante (2013). In other words, it is possible that these traits are, in part, the mechanism through which racial bias is operating.

We can, however, select racially-distinct names that do not trigger these attributes. Figure 2 also includes a racially distinct black name (Octavia) and white name (Misty) that run counter to the trends in the names chosen by DeSante. Unlike Latoya and Keisha, Octavia is perceived as a competent ($\mu = 3.11$), hardworking ($\mu = 2.88$), and professional ($\mu = 3.00$) person. Misty, in contrast, is perceived as less competent ($\mu = 2.34$), hardworking ($\mu = 2.40$), and professional ($\mu = 1.92$) than either Emily or Laurie. By selecting names purposefully to minimize these trait differentials, we can begin to identify the mechanisms through which racial bias operates (e.g., phenotype preference or taste-based discrimination vs. white respondents' biased perception that African Americans are less competent or hardworking). If part of the racial bias uncovered by DeSante (2013) is operating via inferences about worker competence, then we should see smaller treatment effects based on the traits accompanying racially-distinct names.

This leads us to hypothesize that both worker competence and worker race will affect the amount of welfare funding awarded by respondents. We expect that high-competence names (Emily and Octavia) should be awarded more funding than low-competence names (Misty and Keisha). While we expect a racial gap in funding to persist, we expect that the difference in funding between black and white names should be smaller for a high-competence black name (Octavia) than for a low-competence black name (Keisha).

Data and Methods

To test these hypotheses, we fielded an experiment modeled after the treatment used in DeSante (2013) on Amazon's Mechanical Turk in September, 2019. In the original study, all respondents saw two welfare applications with either no names, two white names (Laurie and Emily), a white and a black name (Laurie and Keisha), or two black names (Latoya and Keisha). These applications were also randomly assigned to have a "Worker Quality Assessment" of "Poor," "Excellent," or no assessment. Respondents were then asked to allocate money from a \$1,500 budget to each application. Any money left unallocated would go to "offsetting the state budget deficit."

In order to isolate the effects of race versus the traits people ascribe to racially-distinct names, our replication is a simplified version of this experiment. In our replication, respondents viewed two welfare applications identical in appearance to the original experiment. Rather than randomizing both applications, all respondents viewed the same baseline application of "Emily" who was rated as "Excellent" compared to a second application. We used a $2 \times 2 \times 2$ factorial design for this second application, randomizing the race (white/black), competence (high/low), and quality assessment (excellent/poor) of the second applicant using the names Laurie, Misty, Octavia, and Keisha as our cues for race and competence.[3] Respondents were then asked to allocate funding to the two applicants or to offsetting the deficit. We collected a total 1,200 respondents across these 8 experimental conditions.[4]

Figure 3 presents the main results of our replication for treatments with a "Poor" worker quality

[3] A discussion of how our design differs from DeSante (2013) can be found in Section S3.1 of the Supplemental Information.
[4] Pre-registration information with the Open Science Foundation can be found at `URL BLINDED FOR REVIEW`

assessment.[5] The first two panels are for high status white and black names, respectively. What we find here is remarkable consistency; there are effectively no differences for Laurie or Octavia who are rated as poor workers. The average amount allocated to Emily in these conditions was \$675 when paired with poor Laurie and \$683 when paired with poor Octavia, and the average amounts received by Laurie and Octavia are nearly identical: \$530.25 and \$530.50. This suggests that racial bias is not uniform; Octavia and Laurie appear to be punished equally for having a poor assessment of their worker quality.

When respondents viewed excellent Emily paired with poor Keisha, however, they allocated significantly more (\$745) to Emily than in any other condition ($p = 0.025$, $p = 0.008$, and $p = 0.021$ compared to Laurie, Misty, and Octavia, respectively). This provides some evidence of racial bias consistent with the original findings, although the form of that bias appears to be an increased reward for a white applicant rather than an increased penalty for a black applicant. The only condition in which we see some evidence of punishment is when the applicant is a poorly-rated Misty. In that condition, Misty received less money (\$493) than other comparably-rated names ($p = 0.058$, $p = 0.081$, and $p = 0.084$ compared to Keisha, Laurie, and Octavia, respectively), with respondents allocating correspondingly more to offset the state deficit.

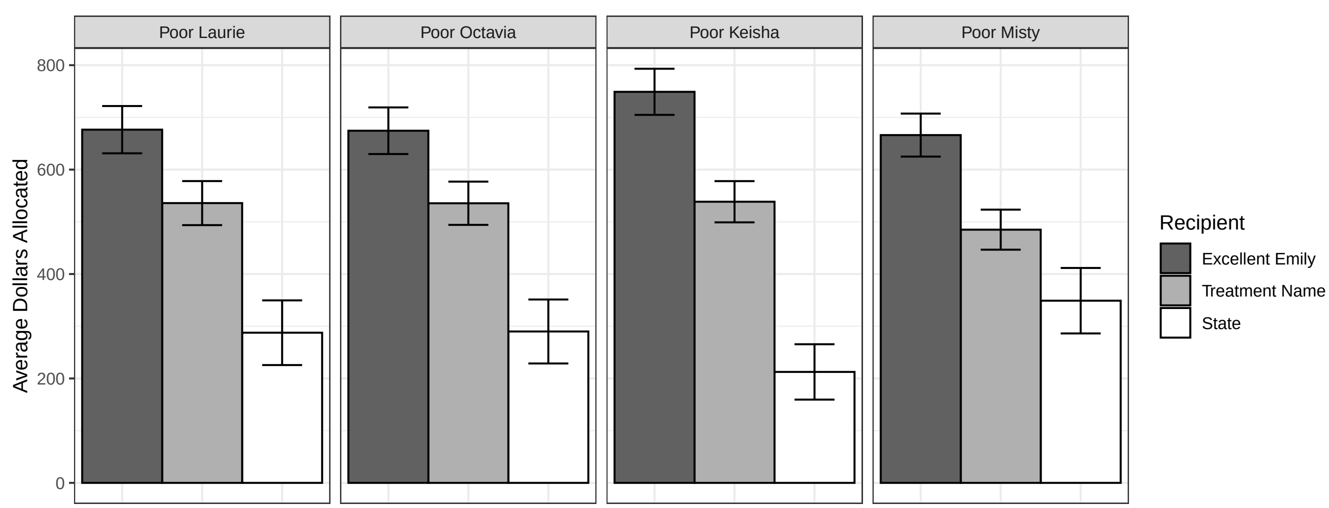

Figure 3: Average allocation by treatment condition

Taken together, our findings suggest that the selection of names used to cue race is important, and can

[5] Results of "Excellent" worker quality assessments can be found in the Appendix. They reveal minimal evidence for race or status differences.

determine the type of bias we detect in our experiments. Using a high-status black name, like Octavia, is likely to minimize racial differences. Indeed, if DeSante (2013) had used only white and black names of comparable perceived status or competence, it is entirely possible that no evidence of racial bias would have emerged, as in our findings for Laurie versus Octavia. If lower status black names are compared to higher status white names, we are likely to find larger treatment effects and a pro-white or anti-black bias (in our Keisha condition). Finally, it is even possible to mistake class- or trait-based bias as anti-white bias if researchers select relatively lower status white names, such as Misty, compared to higher status black names, such as Octavia.

Recommendations for Future Research

In combination, our large, publicly available names dataset and empirical analyses yield important findings. On the one hand, we find that racially- (and gender-) distinct names do send fairly clear signals about group membership. This provides greater confidence in the use of such cues in experimental research. On the other hand, we demonstrate that names trigger a variety of attributes in addition to just race and gender. When individuals encounter a name such as "Emily," for example, they are likely inferring not just that she is a white woman, but also that she is relatively warm, competent, and intelligent. Importantly, we show that there are notable differences in these inferences across names within the same racial group.

What, then, should experimenters do? Rather than abandon the enterprise of measuring racial—and also gender—effects using names, we argue that scholars should treat the complicated nature of such cues as a feature and not a bug. We recommend that experimenters interested in studying racial effects should take one of two approaches in selecting names, depending on their objectives: a matched approach or a randomization approach.

The Matched Approach: In some cases, researchers are interested in isolating the effect of perceived racial phenotype alone, independent of other attributes correlated with race. Researchers in this situation should seek to minimize the differences between names used to cue different races, with special attention to differences on traits that they would theoretically expect to be associated with the outcome of interest. In this situation, we suggest researchers use the many-attribute data from our smaller sample of names to select names that are racially-distinctive, but similar on other characteristics that might affect evaluations. For example, audit studies exploring race-based job-market discrimination should select names that differ on race but are similar on traits desirable to employers, such as competence or laziness.

The Randomization Approach: In other cases, researchers are not interested in isolating the effects of race independent of other relevant attributes. Instead, they may wish to signal race in a way that incorporates differences on nonracial traits and thus reflects how people with racially distinctive names may be judged in general. We suggest that researchers in this situation use the larger sample of names rated on racial attributes to select a group of names that are distinctive for the race(s) of interest. Then, in implementing their experiment, researchers should randomize across these names (i.e., create multiple versions of the treatment containing different realizations of racially distinctive names). For example, a researcher creating a profile of a black candidate in a survey experiment could use our list of the fifty most distinctively black names, and display a randomly selected name for the candidate from this list to each respondent. This approach allows for a fuller representation of the universe of racially distinctive names and should prevent the idiosyncratic attributes of any particular name from unduly influencing the results. Using either of these approaches allows experimenters to measure racial bias while minimizing the impact of potential unanticipated confounds.

More generally, our work shows that paying close attention to the nature of treatments cuing race (or gender[6]) can help us identify the scope of bias and uncover the mechanisms that drive it. DeSante's (2013) original findings, for example, suggest significant bias against black women welfare applicants. We find that the results are more nuanced. Although racial bias clearly exists, it operates at least in part through stereotypes of African Americans as less competent and hardworking. And these stereotypes are not applied equally across all African American names. By separating out the potential pathways through which race, gender, and ethnicity operate, we can make significant strides in understanding the nature of prejudice.

[6]With our data, both approaches can also be applied to gender and politics scholarship. Scholars interested in gender stereotyping can, for example, use our data to choose male and female names that are matched on race and personal traits. Or, they can randomize across male and female names with different attributes.

References

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *The American Economic Review* 94(4):991–1013.

Butler, Daniel M and David E Broockman. 2011. "Do politicians racially discriminate against constituents? A field experiment on state legislators." *American Journal of Political Science* 55(3):463–477.

Butler, Daniel M and Jonathan Homola. 2017. "An empirical justification for the use of racially distinctive names to signal race in experiments." *Political Analysis* 25(1):122–130.

DeSante, Christopher D. 2013. "Working twice as hard to get half as far: Race, work ethic, and Americas deserving poor." *American Journal of Political Science* 57(2):342–356.

Figlio, David N. 2005. Names, expectations and the black-white test score gap. Technical report National Bureau of Economic Research.

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

Levitt, Steven D and Stephen J Dubner. 2006. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. William Morrow chapter 6.

Tzioumis, Konstantinos. 2018. "Demographic aspects of first names." *Scientific data* 5:180025.