Investigator Characteristics and Respondent Behavior in Online Surveys[*]

Ariel White[†] Anton Strezhnev[‡] Christopher Lucas[§]

Dominika Kruszewska[¶] Connor Huff[‖]

October 11, 2017

Abstract

Prior research demonstrates that responses to surveys can vary depending on the race, gender, or ethnicity of the investigator asking the question. We build upon this research by empirically testing how information about researcher identity in *online surveys* affects subject responses. We do so by conducting an experiment on Amazon's Mechanical Turk in which we vary the name of the researcher in the advertisement for the experiment and on the informed consent page in order to cue different racial and gender identities. We fail to reject the null hypothesis that there is no difference in how respondents answer questions when assigned to a putatively black/white or male/female researcher.

[*]Authors are listed in reverse-alphabetical order and contributed equally. For helpful comments, we thank Matthew Blackwell, Joshua Kertzer, Gary King, Jonathan Ladd, Robert Schub, Dustin Tingley, and participants of the Harvard Experimental Political Science Graduate Student Conference and MPSA. For generously funding this research, we thank the Harvard Experiments Working Group, the Harvard Center for American Political Studies, and the Multidisciplinary Program in Inequality and Social Policy at the Harvard Kennedy School. This study was reviewed by Harvard University's IRB with protocol # IRB15-2260 on July 1, 2015. It was pre-registered on EGAP with ID 20151014AB. The data and code required to replicate all analyses in this article can be found on Dataverse (DOI:10.7910/DVN/R8PNCP).

[†]Assistant Professor of Political Science, Massachusetts Institute of Technology, arwhi@mit.edu

[‡]Ph.D. Candidate, Department of Government, Harvard University, astrezhnev@fas.harvard.edu

[§]Ph.D. Candidate, Department of Government, Harvard University, clucas@fas.harvard.edu

[¶]Ph.D. Candidate, Department of Government, Harvard University, dkruszewska@fas.harvard.edu

[‖]Ph.D. Candidate, Department of Government, Harvard University, cdezzanihuff@fas.harvard.edu

1 Introduction

Researchers conducting in-person and telephone surveys have long found that the ways in which respondents answer questions can vary depending on the race, gender, or ethnicity of the interviewer (Hatchett and Schuman, 1975; Cotter, Cohen and Coulter, 1982; Reese et al., 1986; Huddy et al., 1997; Davis, 1997; Davis and Silver, 2003; Adida et al., 2016). This is generally argued to occur for two main reasons. First, the provision of information about the investigator could create demand effects whereby the subjects guess the purpose of the study or the interviewer's views and change their responses to align with this perceived purpose.[1] Second, potential subjects may be more or less comfortable answering questions from researchers with a particular identity and subsequently either refuse to participate in studies, decline answering certain questions, or censor the ways in which they answer, all of which could substantively change the results of survey research. Researchers often seek to mitigate these concerns when designing surveys.[2]

In this paper we build upon prior research and empirically test whether researcher identity affects survey responses in *online survey platforms*. We do so by varying information about the researcher—conveyed through their name—in both the advertisement for survey participation and the informed consent page. We take this approach for two reasons. First, the inclusion of researcher names at each of these junctures is common practice. Second, an emerging strain of research throughout the social sciences demonstrates how inferences made from names can affect behavioral outcomes even in the absence of in-person or telephone interactions.[3] We

[1]We can consider concerns about social desirability bias to fall into this category.

[2]For example, Grewal and Ritchie (2006), Schaeffer, Dykema and Maynard (2010), and Survey Research Center (2010) explicitly advise researchers to consider interviewer effects as part of the research design, though more recent research demonstrates that demand effect concerns might be overstated for survey experiments (Mummolo and Peterson, 2017). See also Berrens et al. (2003) for discussion of the advantage of internet surveys in reducing interviewer bias compared to telephone or in-person surveys, and Bush and Prather (2016) for how the mode of technology used to conduct surveys can substantively affect survey responses.

[3]See for example, Bertrand and Mullainathan (2004), Butler and Broockman (2011), White, Nathan and Faller (2015), Levine Einstein and Glick (2017), Edelman, Luca and Svirsky (2015), and most recently Butler and Homola (2017).

go on to test how this variation in the researcher name affects the ways in which respondents answer questions in online surveys.

The experiment conducted in this paper contributes to an expanding strain of research exploring the composition and attributes of online survey pools.[4] Our findings help to interpret the substantive results of prior studies that used online surveys,[5] and also provide guidelines for researchers as they move forward. In this study, we fail to reject the null hypothesis of no difference in respondents' behavior when assigned to a putatively black/white or female/male researcher. Our estimates suggest that there could be a substantively small difference between question responses for putatively male and female researchers, but given the high power of the experiment, we are able to bound the substantive size of the effect. We conclude that these differences are likely substantively negligible for most researchers. In general, the results of this paper demonstrate that researchers need not worry that using their own names in either survey advertisements or online consent forms will substantively affect online survey results.

2 Experimental Design

In the experiment each respondent was "treated" by exposure to one researcher name intended to cue race and gender, appearing first in the advertisement for the survey and then in the consent form inside the survey. The experiment was conducted on Amazon's Mechanical Turk (MTurk), where it is common for researchers' names to appear at both of these points.[6] To generate the names associated with each of these manipulations, we combined three commonly

[4]See, for example, Berinsky, Margolis and Sances (2014); Chandler, Mueller and Paolacci (2014); Krupnikov and Levine (2014); Clifford, Jewell and Waggoner (2015); Huff and Tingley (2015); Mullinix et al. (2015); Levay, Freese and Druckman (2016); Leeper and Thorson (2015).

[5]A few prominent examples of political science articles published using online samples drawn from Mechanical Turk have been published in the *American Political Science Review* (Tomz and Weeks, 2013), *American Journal of Political Science* (Healy and Lenz, 2014), *Comparative Political Studies* (Charnysh, Lucas and Singh, 2014), *International Organization* (Wallace, 2013), and *the Journal of Conflict Resolution* (Kriner and Shen, 2013).

[6]Readers will note that this design captures two stages: first, selection into the survey, and second, the ways in which respondents answer questions conditional on having selected into the survey. In Appendix Section J we present results from a different experiment in which we randomize the name of the researcher *only on the consent*

used lists of racially distinct first and last names.[7] We crossed the lists of first and last names to produce many possible combinations[8] and drew two names for each of the four manipulation categories (black men, black women, white men, and white women). The full list of names used in this experiment is presented in Table 1.

Black Men	Deshawn Booker	Tyrone Robinson
Black Women	Ebony Gaines	Deja Washington
White Men	Connor Schroeder	Brett Walsh
White Women	Molly Ryan	Laurie Yoder

Table 1: Names used for each of the four investigator name manipulations, based on lists from Bertrand and Mullainathan (2004), Fryer, Jr. and Levitt (2004), Word et al. (2008)

We then created accounts under the names of our hypothetical researchers ("Ebony Gaines", "Brett Walsh", etc.) and recruited subjects through these named accounts. We also included these researcher names on the consent forms for our study. This dual approach is both realistic and methodologically useful. Many Institutional Review Boards require that the researcher include their name on the consent form, and as shown in Table 2, a large number of researchers post studies on platforms such as MTurk under their own names. Given these practices, the substantive nature of treatment is consistent with common practices for researchers using the MTurk survey pool. Moreover, the research design allows us to measure how knowledge about researchers' identities can shape not only the nature of responses, but the overall response rate.[9] Posting the survey from named researcher accounts means that potential respondents

form with a generic account name. Doing so allows us to estimate the effect of varying the researcher name only in the consent form where there is no initial selection step. The results from this second experiment are substantively consistent with what we present in the remainder of this paper.

[7]First names were drawn from a combination of lists found in Bertrand and Mullainathan (2004) and Fryer, Jr. and Levitt (2004), while last names were drawn from lists in Word et al. (2008) and Bertrand and Mullainathan (2004). Our instrument did not include a manipulation check but the studies from which we drew the list of names have shown that they are racially-distinctive enough for respondents to make inferences about the person's racial identity. We are thus confident that the names we used were highly informative about the race and gender of the individual conducting the study.

[8]We omitted a few randomly-generated names that already belonged to celebrities, such as Jermaine Jackson.

[9]The results for this are presented in Appendix Section D.

see the name of the researcher before deciding whether or not to participate, so it allows us to capture the selection process that may occur in real studies.

Search Term	Accounts Using Real Names	Total Accounts	Proportion Using Name
Politics	3	5	.6
Survey	121	169	.72
Research	21	41	.51
Academic	5	9	.56
University	1	4	.25
Political Science	1	1	1
Psychology	23	31	.74
Economics	1	1	1
TOTAL	176	261	.67

Table 2: Displays the number of unique accounts on MTurk using real names. To calculate these amounts, we searched for the specified term then scraped all account names on August 15, 2016. Next, we manually classified all unique account names as either a real, identifiable name or any other naming scheme (lab name, nonsensical string, etc.).

However, including the treatment in the recruitment process poses design challenges. We could not simply post all treatment conditions simultaneously, because users would then see eight identical surveys posted under eight different researcher names and immediately understand the purpose of the experiment. Instead, we set up the experiment such that any user could only observe one treatment condition, by pre-recruiting a pool of respondents.

First, we ran a pre-survey asking only one question[10] that captured the unique MTurk "workerID" of each respondent that opted in (N of approximately 5000). Second, we randomly assigned each of these unique identifiers to one of the eight researcher name conditions listed in Table 1. Finally, we created separate MTurk accounts under each researcher name and deployed the same survey within each account. Subjects were assigned a "qualification" within the MTurk interface, according to their assigned condition. Each survey was set such that only MTurk workers with the correct qualification could see the survey (and thus the username associated with it).[11] This meant that each potential respondent could see only one survey

[10]The question asked about the number of tasks the respondent had previously completed on Mechanical Turk.

[11]In practice, Mechanical Turk functions were done through R scripts using the MTurkR package to access the Mechanical Turk API (Leeper, 2015, 2013). This allowed us to post tasks in small batches (of 9 at a time) so as to avoid having the tasks posted to online MTurk discussion boards where workers share lucrative HIT opportunities

from their assigned researcher, and could then choose whether or not to take that survey. In summary, we posted an initial survey where we collected MTurk IDs, randomly assigned these workers to one of eight conditions where we varied the researcher name, and then *only respondents in that condition could view that HIT*.[12]

Within the survey, respondents answered a series of questions about social and political attitudes. We drew questions from Pew, Gallup, and the American National Election Survey, specifically asking about issues for which racial and gender cues may prompt different responses.[13] We chose to ask questions about race and gender, as these are two of the main areas where prior research has demonstrated that interviewer attributes can affect subject behavior. Moreover, this is the information conveyed most prominently by researchers through their names in online surveys. After all subjects had completed all study-related activities, respondents were de-briefed about the nature and purpose of the study.

3 Results

Our design allows us to test whether researcher identity shapes the sample of respondents that agree to take the survey. We find little evidence of such an effect.[14] We find substantively small differences in the number of people who take the different surveys, and no difference in respondents' backgrounds on a range of personal characteristics. We also do not find differences in survey completion rates across name; all rates were extremely high (above 97%). Therefore, we are not concerned about inducing selection bias by analyzing the set of completed surveys. We turn next to the content of survey responses.

(this could have exposed our experimental design). We posted these small batches at short, regular intervals (each HIT expired and was re-posted every 15 minutes for several days) to ensure that the tasks were continuously available to potential workers across all experimental conditions. This approach seems to have worked: regular scans of major MTurk discussion board did not reveal any postings about our HITs.

[12] "HITs" or Human Intelligence Tasks are the name MTurk gives any individual unit of work posted on the site. In this case, a HIT included a link to take our survey for some pre-specified payment amount.

[13] The full text of the outcome questions is presented in Appendix Section I.

[14] We explore the selection process in more detail in the Online Appendix.

In general, our analyses fail to reject the null hypothesis that there is no difference in how respondents answer questions when assigned to a putatively black or female researcher relative to a white or male one. We estimate all of our treatment effects using linear regression models, regressing outcome on the indicator of treatment. Robust standard errors are estimated using a bootstrapping procedure. Following our pre-analysis plan, our rejection levels for accepting that the effects differ from zero are calibrated to yield an expected number of false discoveries of $\alpha = .05$, adjusting for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).[15] This adjustment is important since we dramatically increase the chances of a false positive finding by testing for multiple outcomes (Benjamini and Hochberg, 1995).[16] To avoid the appearance of "fishing" for significant p-values across many outcomes, we cannot simply follow a rule of rejecting any null hypothesis when $p < .05$. We focus on estimating only the average treatment effects of the researcher race and gender treatments, and, consistent with our pre-analysis plan, only investigate possible treatment effect heterogeneity as exploratory rather than confirmatory results.[17]

Our first set of outcome questions examines whether assignment to a putatively female/black (relative to male/white) investigator changes reported affect towards, or support for policies meant to help, women/blacks.[18] For the race dimension of treatment, we estimate treatment effects on three distinct outcomes: expressed racial resentment (as measured by the 0 to 1 scale developed by Kinder and Sanders (1996)), willingness to vote for a black president, and

[15]For the Benjamini-Hochberg procedure, we order our m p-values from smallest to largest $P_{(1)}, ..., P_{(m)}$ and find the largest k such that $P_{(k)} \leq \alpha \times \frac{k}{m}$. This ensures that our false discovery rate, that is is the expected share of rejected nulls that are 'false positives,' is controlled at $\alpha = .05$. Note that under this procedure, we would not reject any nulls if all p-values are $> .05$ and reject all nulls if all p-values are $< .05$.

[16]While Benjamini and Hochberg (1995) analyze the case where hypotheses are independent, Benjamini and Yekutieli (2001) show that the procedure also properly controls the false-discovery rate under positive dependence between hypotheses. This is likely to be the case under our set of tests as each question can be seen as measuring different elements of an individual's latent affect towards a group. Moreover, simulation studies by Groppe, Urbach and Kutas (2011) show good performance of the Benjamini-Hochberg method even under violations of independence.

[17]For a discussion of potential treatment effect heterogeneity by race/gender, see Appendix Section G.

[18]In Appendix Sections C-E, we also report results for selection into the survey itself, survey completion, and attention check passage rates, finding no substantive differences across the treatment conditions.

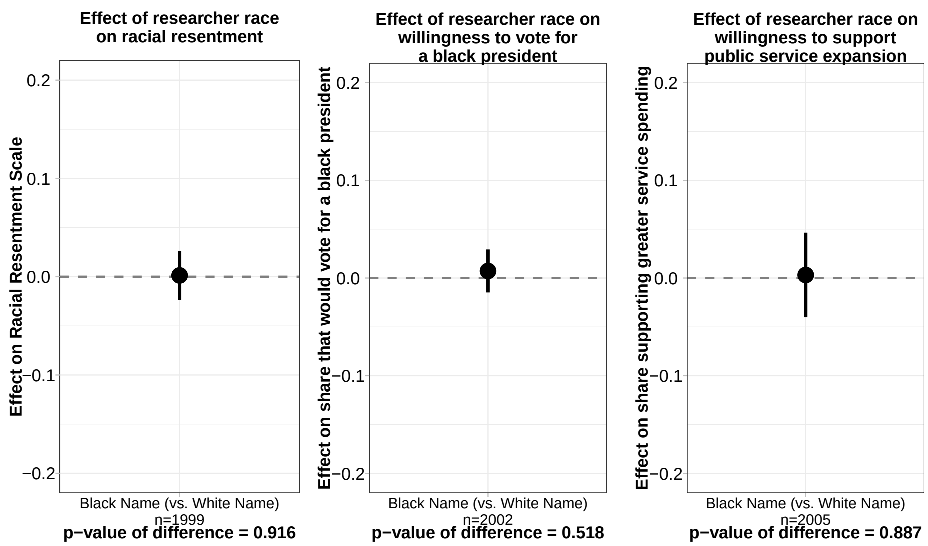
support for social service spending. On gender, we examine respondents' beliefs regarding the role of women in society, willingness to vote for a woman presidential candidate, and the same social service spending outcome. In selecting our first two outcome questions, we sought questions that were both commonly used in online surveys but also directly related to each of our treatments. The social spending measure was included as a facially non-racial measure that could still have racial or gendered overtones. This allowed us to test whether respondents would think of social spending as disproportionately benefiting minorities and women, and so potentially answer in either raced or gendered ways depending on the putative race or gender of the researcher.

We designed our experiment to target a sample size of 2000 total respondents.[19] For the race treatment, we find no evidence that black versus white researcher names yield different responses on the outcome questions. Figure 1 plots the expected difference in outcomes for each of these three questions for respondents assigned to a hypothetical black researcher name relative to respondents assigned to a hypothetical white researcher name. For all three outcomes, the difference in outcome between the two treatment groups is not statistically significant at $\alpha = .05$. We fail to reject the null of no effect for all outcomes at the $\alpha = .05$ level.

For the gender treatment, when we adjust for multiple comparisons we fail to reject the null hypothesis that there is no difference between putatively male or female researchers. Figure 2 plots the difference in expected values for each of the the outcomes between the female researcher and male researcher treatment conditions. While we fail to reject the null, we should note that for all outcomes, respondents under the female researcher treatment condition were about two to four percentage points more likely to express affective/policy support for women. The individual p-value for the null of no effect on the gender equality outcome question fell just below the commonly used threshold of .05. The p-values for the null for the other two outcomes, however, fall just above the typically used threshold. Under our pre-registered design, using the Benjamini-Hochberg correction for multiple testing, we fail to

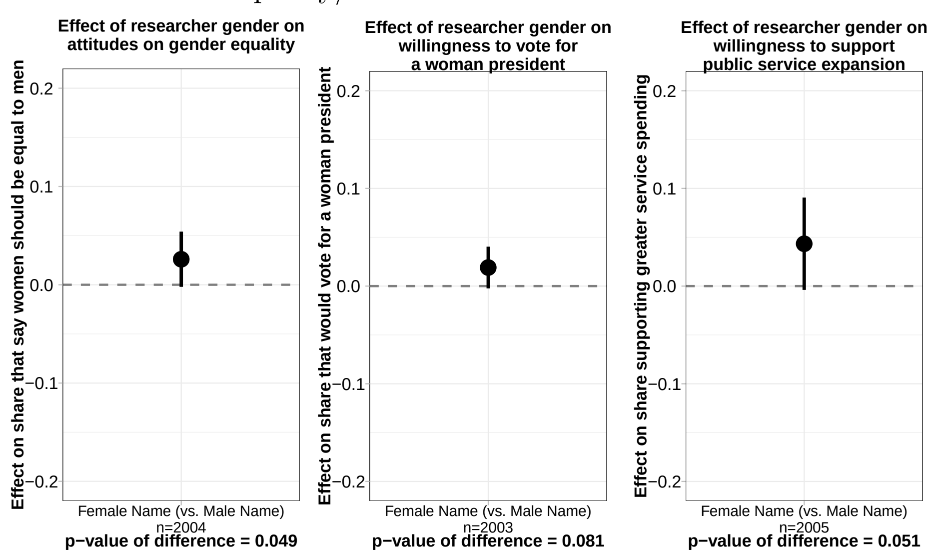
[19] Our final sample consists of 2006 unique respondents that we could confirm had completed the overall survey. We omit responses from one respondent who requested that their responses not be used after the debriefing process.

reject the null for all three outcomes.[20] We cannot conclude that assignment to a putatively female researcher name significantly increased the likelihood that respondents would exhibit more woman-friendly attitudes on gender-related questions.



Note: Lines denote 95% multiple comparison-adjusted confidence intervals (Benjamini and Yekutieli, 2005).

Figure 1: Differences in policy/attitude outcomes for researcher race treatment.



Note: Lines denote 95% multiple comparison-adjusted confidence intervals (Benjamini and Yekutieli, 2005).

Figure 2: Differences in policy/attitude outcomes for researcher gender treatment.

Despite our failure to reject the null, we note that the point estimates for the direction of

[20]This is because the threshold level for rejecting the null of no effect on only the gender equality outcome falls to $\alpha = .05 \times \frac{1}{3} = .016$ – below the p-values that we observe.

the effect are consistent with our original hypothesis. In general, respondents assigned to a putatively female investigator were, in-sample, more likely to express beliefs that were more supportive of women's equality. Given this, how concerned should researchers be about these estimates? Power calculations for our design suggest a relatively small upper bound for any "true" effect. For a study of our sample size, and accounting for the multiple comparison adjustment, we conclude that it is unlikely that we would have failed to reject had the true effect of any one of these outcomes been greater than 5 percentage points.[21] While it is not possible to "affirm" a null hypothesis, the high power of our study is such that our null finding implies any real effect is likely to be bounded close to 0.

4 Discussion and Conclusion

In this paper, we demonstrate that researchers using online survey platforms such as Amazon's Mechanical Turk generally need not be concerned that information conveyed through their name in either the advertisement for the HIT or the informed consent page will subsequently affect their results. Our study is designed to address both elements of investigator bias: inferences about the purposes of the study and comfort with the investigator, either of which we might expect to affect the willingness of respondents to take the survey in the first place, overall effort, and the types of answers given. We fail to reject the null hypothesis that researchers' race or gender (cued through names) have no effect on respondents' survey behaviors. While our evidence suggests there might be a small "true effect" of researcher gender, our power calculations demonstrate that this effect, if any, is quite small and likely not substantively meaningful for most researchers.

There are at least two plausible explanations for why the results of this paper diverge from the substantively meaningful effects found in research on other survey platforms. First, it could be the case that either the strength or substance of the treatment differs between online survey platforms and other modes of conducting surveys (such as in-person or telephone). That is, interacting in person with a black/white or male/female researcher might have a stronger

[21] For a more detailed discussion of the power calculations, see Appendix Section F.

effect on respondent behavior than simply reading their names. Substantively, this means that even if respondents do notice the putatively black/white or male/female name assigned to the researcher through treatment, the act of reading this name is simply not enough to change their subsequent survey responses.

Second, it could be the case that respondents in online survey platforms are less likely to take treatment.[22] If this were the case, it could be in part driven by the fact that our respondents were recruited via Amazon's Mechanical Turk, where the financial incentives for respondents to complete tasks as quickly as possible might lead them to quickly skim through the consent page.[23] This means they would be less likely to notice the researcher name and thus less likely to respond to it. Even if respondents were prone to bias, it could be masked by the fact that few respondents actually read the names in the first place. The present study is unable to adjudicate between these two potential explanations.

For researchers conducting studies on MTurk and similar online platforms, this distinction will not matter. Nevertheless, the two different mechanisms have important implications both for the external validity of the present study as well as further research on the attributes of online survey pools. In particular, researchers should be cautious in applying the results of this study when either (1) they provide more information about themselves than simply their name in the advertisement for the survey and on the informed consent page (that is, they have a stronger treatment), or (2) respondents in their sample pay more attention throughout all stages of the survey than MTurk respondents (i.e. there is higher treatment uptake). In our experience, the first point is unlikely to occur across different survey platforms since few platforms provide more researcher information than MTurk. However, whether and how much attention varies across different survey platforms and how this substantively affects results is an open question and interesting area of further research.

[22] This explanation would be consistent with the null finding presented in this paper on the effect of researcher name on likelihood of selecting into the survey, though this evidence is not sufficient to rule out the first explanation.

[23] We note, though, that MTurk workers' financial incentives could operate in either direction. The existence of services like Turkopticon, used to keep track of individual requester accounts and their payment practices, suggests that Turkers might be even more motivated than other survey takers to notice researcher names.

References

Adida, Claire L., Karen E. Ferree, Daniel N. Posner and Amanda Lea Robinson. 2016. “Who’s Asking? Interviewer Coethnicity Effects in African Survey Data.” *Comparative Political Studies* 49(12):1630–1660.

Benjamini, Yoav and Daniel Yekutieli. 2001. “The Control of the False Discovery Rate in Multiple Testing Under Dependency.” *Annals of Statistics* 29(4):1165–1188.

Benjamini, Yoav and Daniel Yekutieli. 2005. “False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters.” *Journal of the American Statistical Association* 100(469):71–81.

Benjamini, Yoav and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

Berinsky, Adam J., Michele F. Margolis and Michael W. Sances. 2014. “Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys.” *American Journal of Political Science* 58(3):739–753.

Berrens, Robert P., Alok K. Bohara, Hank Jenkins-Smith, Carol Silva and David L. Weimer. 2003. “The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples.” *Political Analysis* 11(1):1–22.

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

Bush, Sarah and Lauren Prather. 2016. “An Experiment on the Use of Electronic Devices to Collect Survey Data.” International Methods Colloquium.

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.

Chandler, Jesse, Pam Mueller and Gabriele Paolacci. 2014. “Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers.” *Behavior Research Methods* 46(1):112–130.

Charnysh, Volha, Christopher Lucas and Prerna Singh. 2014. “The Ties That Bind: National Identity Salience and Pro-Social Behavior Toward the Ethnic Other.” *Comparative Political Studies* 48(3):267–300.

Clifford, Scott, Ryan M. Jewell and Philip D. Waggoner. 2015. “Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?” *Research & Politics* 2(4):1–9.

Cotter, Patrick R., Jeffrey Cohen and Philip B. Coulter. 1982. "Race-of-Interviewer Effects in Telephone Interviews." *Public Opinion Quarterly* 46(2):278–284.

Davis, Darren W. 1997. "The Direction of Race of Interviewer Effects among African-Americans: Donning the Black Mask." *American Journal of Political Science* 41(1):309–322.

Davis, Darren W. and Brian D. Silver. 2003. "Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge." *American Journal of Political Science* 47(1):33–45.

Edelman, Benjamin G, Michael Luca and Dan Svirsky. 2015. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." *Harvard Business School NOM Unit Working Paper* (16-069).

Fryer, Jr., Roland G. and Steven J. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *The Quarterly Journal of Economics* CXIX(August):767–806.

Grewal, Ini and Jane Ritchie. 2006. Ethnic and Language Matching of the Researcher and the Research Group During Design, Fieldwork and Analysis. In *Health and Social Research in Multiethnic Societies*, ed. James Y. Nazroo. Oxon: Routledge pp. 65–81.

Groppe, David M., Thomas P. Urbach and Marta Kutas. 2011. "Mass Univariate Analysis of Event-Related Brain Potentials/Fields II: Simulation Studies." *Psychophysiology* 48(12):1726–1737.

Hatchett, Shirley and Howard Schuman. 1975. "White Respondents and Race-of-Interviewer Effects." *The Public Opinion Quarterly* 39(4):523–528.

Healy, Andrew and Gabriel S. Lenz. 2014. "Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy." *American Journal of Political Science* 58(1):31–47.

Huddy, Leonie, Joshua Billig, John Bracciodieta, Patrick J. Moynihan and Patricia Pugliani. 1997. "The Effect of Interviewer Gender on the Survey Response." *Political Behavior* 19(3):197–220.

Huff, Connor and Dustin Tingley. 2015. ""Who Are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2(3):1–12.

Kinder, Donald R and Lynn M Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. University of Chicago Press.

Kriner, Douglas L. and Francis X Shen. 2013. "Reassessing American Casualty Sensitivity The Mediating Influence of Inequality." *Journal of Conflict Resolution* 58(7):1174–1201.

Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1(01):59–80.

Leeper, Thomas. 2013. "Crowdsourcing with R and the MTurk API." *The Political Methodologist* 20:2–7.

Leeper, Thomas and Emily Thorson. 2015. "Minimal Sponsorship-Induced Bias in Web Survey Data." https://s3.us-east-2.amazonaws.com/tjl-sharing/assets/SurveySponsorship.pdf.

Leeper, Thomas J. 2015. *MTurkR: Access to Amazon Mechanical Turk Requester API via R.* R package version 0.6.5.1.

Levay, Kevin E., Jeremy Freese and James N. Druckman. 2016. "The Demographic and Political Composition of Mechanical Turk Samples." *SAGE Open* 6(1):1–17.

Levine Einstein, Katherine and David M. Glick. 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.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(02):109–138.

Mummolo, Jonathan and Erik Peterson. 2017. "Demand Effects in Survey Experiments: An Empirical Assessment." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2956147.

Reese, Stephen D., Wayne A. Danielson, Pamela J. Shoemaker, Tsan-Kuo Chang and Huei-Ling Hsu. 1986. "Ethnicity-of-Interviewer Effects Among Mexican-Americans and Anglos." *Public Opinion Quarterly* 50(4):563–572.

Schaeffer, Nora Cate, Jennifer Dykema and Douglas W. Maynard. 2010. Interviewers and Interviewing. In *Handbook of Survey Research*, ed. Peter V. Marsden and James D. Wright. Emerald Bingley pp. 437–470.

Survey Research Center. 2010. *Guidelines for Best Practice in Cross-Cultural Surveys.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/.

Tomz, Michael and Jessica Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(4):849–865.

Wallace, Geoffrey P.R. 2013. "International Law and Public Attitudes Toward Torture: An Experimental Study." *International Organization* 67(01):105–140.

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

Word, David L., Charles D. Coleman, Robert Nunziata and Robert Kominski. 2008. "Demographic Aspects of Surnames from Census 2000." https://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf.

Online Appendix for *Investigator Characteristics and Respondent Behavior in Online Surveys*

Contents

A Additional Reporting Standards Information **2**

B Recruitment and Randomization Procedure **3**

C Completion Rates **4**

D Selection Into Treatment **6**

E Attention Checks **9**

F Power Calculations **10**

G Exploratory Analysis of Effect Heterogeneity **11**

H Survey Questions (Order Randomized) **13**

I Pre-Analysis Plan **15**

I.1 Introduction: A Second Experiment on Researcher Identity in Online Surveys . . . 15

I.2 Experimental Survey Design . . . 16

I.3 Statistical Tests . . . 18

I.4 Appendix: Survey Text . . . 19

I.4.1 Informed Consent 19
I.4.2 Outcome Questions [Question order randomized] 21

J A Second Experiment: Randomization in Only the Consent Form **27**

A Additional Reporting Standards Information

Most of the requirements of the JEPS reporting standards are met by our main paper and various parts of this Appendix. In this section, we include several additional details that should aid in the understanding or replication of our study.

Recruitment Dates We recruited respondents and implemented the study between March 3 and April 4, 2016.

Incentives We paid survey participants between \$0.05 and \$1 for each wave of the study.

Evidence of Random Assignment Here we present group means of background covariates for respondents assigned to our different treatment conditions, as evidence of balance (suggesting our random assignment to treatment conditions worked as expected).

Variable	N	Male Researcher	Female Researcher	T-statistic
Prop. Male	2005	0.508	0.468	1.782
Prop. Democrat	2004	0.437	0.456	-0.839
Prop. Republican	2004	0.220	0.198	1.232
Prop. Voted in 2012	2005	0.703	0.729	-1.261
Prop. White	2002	0.820	0.805	0.837
Prop. Black	2002	0.070	0.065	0.389

Table A1: Covariate means by treatment conditions - Researcher gender

Variable	N	White Researcher	Black Researcher	T-statistic
Prop. Male	2005	0.484	0.491	-0.337
Prop. Democrat	2004	0.441	0.453	-0.544
Prop. Republican	2004	0.206	0.211	-0.307
Prop. Voted in 2012	2005	0.714	0.718	-0.191
Prop. White	2002	0.809	0.815	-0.365
Prop. Black	2002	0.074	0.061	1.170

Table A2: Covariate means by treatment conditions - Researcher race

Outcome Means/Standard Deviations Here we present means and standard deviations for all of our outcome measures.

Outcome	Mean	Std. Dev.	Number of Responses
Racial Resentment	0.468	0.280	1999
Would vote for a black president	0.934	0.249	2002
Would vote for a woman president	0.938	0.241	2003
Women's role: Equal	0.904	0.295	2004
Women's role: Home	0.053	0.225	2004
Expand public services	0.619	0.486	2005
Cut public services	0.381	0.486	2005

Table A3: Summary statistics for outcome variables

B Recruitment and Randomization Procedure

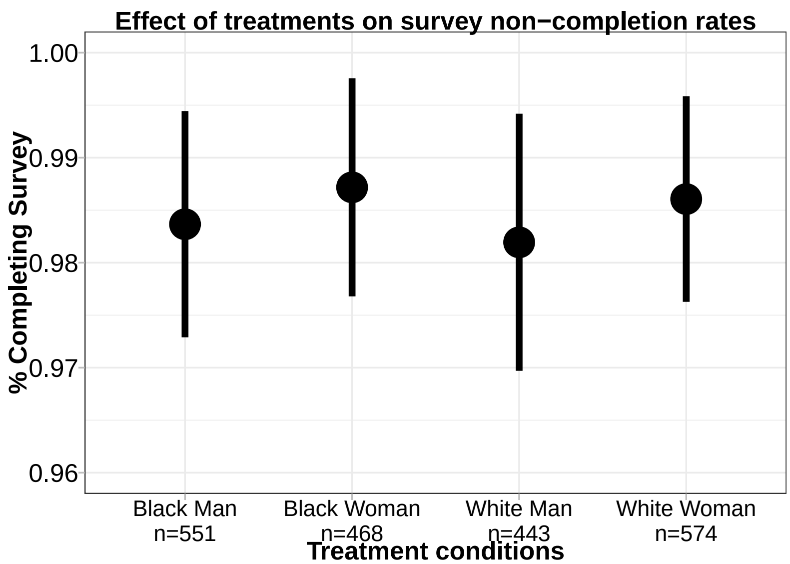
In order to assign Mechanical Turk workers to each of our treatment conditions, we first needed to obtain a sufficient number of candidate worker IDs that we could assign to each researcher name treatment. We recruited these respondents via an initial HIT which yielded a total of 5858 respondents. We then removed respondents that had either taken a previous version of one of our pilot studies, or took the survey multiple times. Doing so reduced the number of candidate respondents to 4659. We allocated respondents uniformly to the researcher name treatments in two waves, the first wave containing 3210 and the second wave containing 1449 respondents.

Of the 4659 respondents assigned to take our HIT, we received 2110 response to all versions of our survey. However, some of these respondents either accidentally or intentionally took the survey twice, due to a lag between the time that a new version of a HIT was posted and when we were able to remove a respondent from being qualified to take the HIT. We were able to recover MTurk IDs for most respondents and removed any duplicate HITs beyond the first taken by the respondent. For those responses where we could not record an MTurk ID, possibly due to respondents disabling some Javascript code execution on their browsers, we drop responses based on duplicated IP addresses. While this is a bit conservative as multiple users can in theory share the same IP address if they are on the same network (such as a public library), we want to avoid including multiple observations from the same source. In all cases, the observations with duplicated IPs had the same treatment assignment, suggesting that we are picking up on the same person entering the survey multiple times. Our replication code provides the exact mechanism that we used to complete this de-duplication task. However, for privacy reasons, we do not release the IP addresses or the Mechanical Turk IDs of the respondents. We convert both IP address and MTurk ID to an arbitrary numerical indicator in the replication files that accompany this paper. After removing duplicates, we are left with 2036 unique observations. In total, 2016 of our 2036 respondents completed the survey and reached the end of the task. Of these 2016, we could match 2006 to one of the 4659 Mechanical Turk IDs we assigned at the beginning. We drop responses from one of these participants due to a request from that person after they completed the survey, for a total of 2005 responses. These 2005 constitute our sample of "completed" surveys.

C Completion Rates

An important element of non-response is the question of whether respondents are differentially failing to complete the survey based on which treatment condition they receive. We do not find strong evidence that this is the case. Figure A1 plots the share of respondents under each treatment category that we could identify had completed the survey via Mechanical Turk. We fail to reject the null hypothesis of no differences in completion rates between each of the four treatment categories ($p > .05$). Also, in general, completion rates were very high (between 98 and 99 percent), so we do not have significant concerns over potential post-treatment bias induced by conditioning only on those respondents that finished the survey.[1]

[1] Note that in estimating these differences, we are conditioning on individuals selecting into the survey in the first place. We discuss selection into treatment in Section D below.



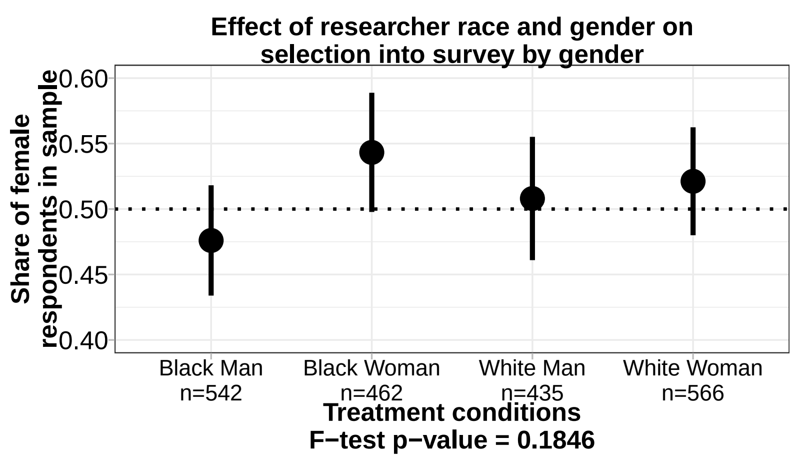
Note: Lines denote 95% confidence intervals.

Figure A1: Completion rates conditional on starting survey across four treatment categories.

D Selection Into Treatment

One of the mechanisms through which cues of researcher identity might affect survey results and responses is through differential selection into treatment. Some individuals may be more or less likely to take a survey after receiving information about the researcher, and differences in responses to survey questions may therefore be attributable to differences in respondents' covariate distributions.

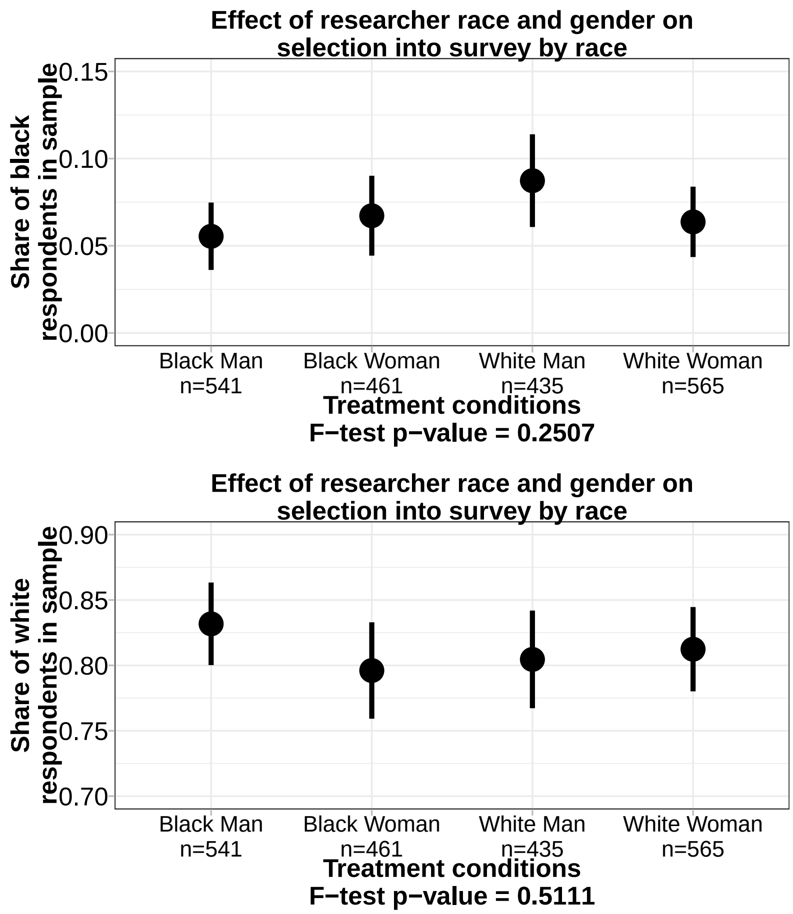
We do find evidence that some of our Mechanical Turk HITs were more likely to be taken by respondents. We test the null hypothesis that the observed counts in each of the four treatment categories are statistically indistinguishable from a uniform distribution by calculating a test statistic, the difference between the largest and smallest number of respondents in a treatment category. We then compared the observed test statistic to the null distribution where each respondent in the survey was assigned to each of the four treatment categories with probability .25. Using Monte Carlo simulation, we calculate a p-value of 0.00018, far below the common threshold of $\alpha = .05$. Based on the test, there is evidence to conclude that there may be some difference in the "popularity" of HITs based on the researcher name in the advertisement.



Note: Lines denote 95% confidence intervals.

Figure A2: Differences in treatment groups across respondents' background covariates: Gender

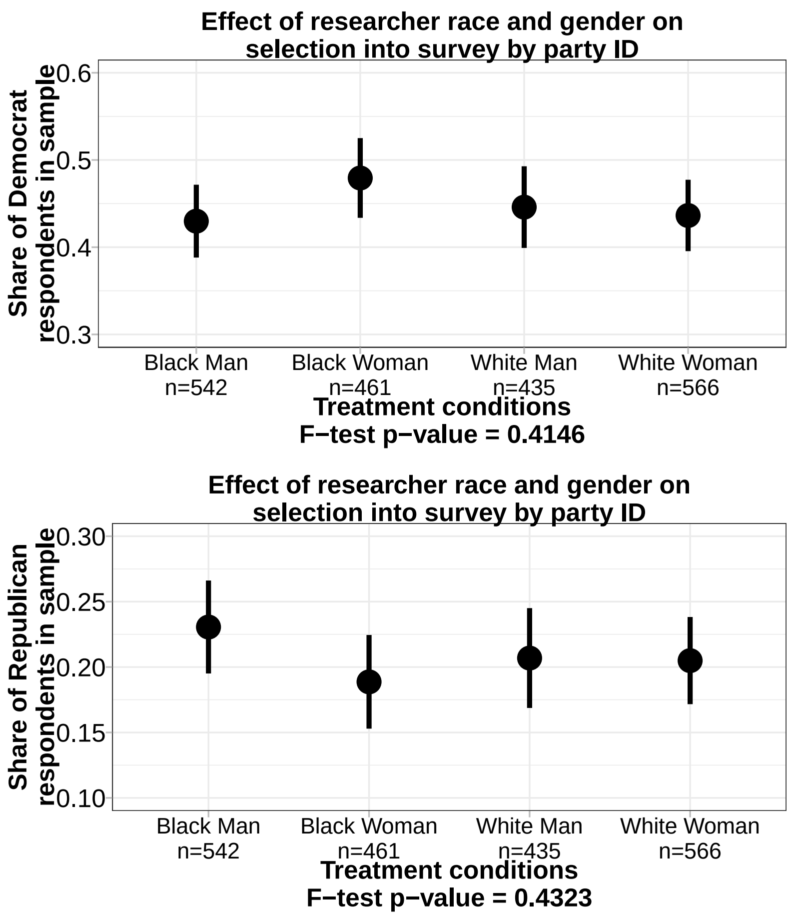
However, the magnitude of this difference is not particularly large. And more importantly, there do not appear to be significant differences between the samples for any of the four treatment categories based on the background covariates that we collected. Figure A2 plots the share of female



Note: Lines denote 95% confidence intervals.

Figure A3: Differences in treatment groups across respondents' background covariates: Race

respondents in each of the four treatment categories. Using an F-test, we cannot reject the null of no difference across all four treatment categories (at $\alpha = .05$). The same null result holds for respondents who identify as white, for those who identify as black (Figure A3) and for political party identification (Figure A4). If there is some interesting selection into treatment, it is not clear that it relates to any of these commonly observed background covariates.

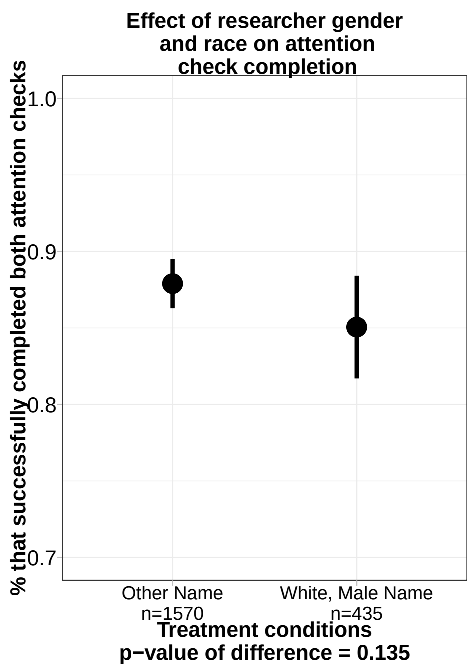


Note: Lines denote 95% confidence intervals.

Figure A4: Differences in treatment groups across respondents' background covariates: Party ID

E Attention Checks

Our final set of outcome questions test for whether respondents pay greater attention during surveys conducted by white male researchers when compared to all other treatment categories. We measure attention using two attention check questions randomly introduced in the experiment. Figure A5 reports the estimated probabilities of completing both attention checks correctly under the white male and non-white male treatments. Roughly 85 to 88 percent of respondents in both treatment groups successfully navigated the two attention check questions and the difference between the two is not statistically significant at $\alpha = .05$. Not only did treatment not appear to affect response substance (in terms of policy and attitude questions), it also did not significantly influence response quality (in terms of attention).



Note: Lines denote 95% confidence intervals.

Figure A5: Differences in attention check outcomes for researcher race/gender treatment.

F Power Calculations

This section outlines our procedure for evaluating the power of our hypothesis test of the effect of treatment on the three gender equality outcomes. To calculate the power of a hypothesis test, we need to make assumptions on the sampling distribution of the test statistics under the null. In our case, we have three large-sample difference-in-means tests between two proportions. While we fix the sample size ex-ante and vary the hypothetical "true" effect size for each treatment, we need to make an additional assumption about the population variance for each outcome. We start by using the sample variance of each of the three outcomes: beliefs about gender roles, willingness to vote for a woman president, and support for social service spending. Since we want to be somewhat conservative in our power analyses and account for potentially more variable samples, we scaled up each of the in-sample estimates by a factor of 1.5 (bounded by the maximum of .25 for a binary variable).

We evaluate power for two scenarios: one where all three nulls are false and have equal effect sizes, and one in which only one null is false (for the highest variance outcome). In the first scenario, we compute the "average power" (the share of nulls rejected) (Benjamini and Liu, 1999). In the second, we calculate the probability of rejecting the *non-null* hypothesis specifically. Figure A6 plots the two power curves for the hypothetical difference-in-means test between two proportions. We consider the case of $n = 2000$ evenly split between treatment and control. The y-axis denotes the power or expected power and the x-axis plots the hypothesized true absolute effect size. Following the Benjamini-Hochberg procedure for a false discovery rate of .05, we see that with a relatively high power of .8, we would detect effects greater than or equal to 5 percentage points. So even though we failed to reject the null hypothesis, the power of our design is such that it is unlikely that this would have happened had the true effects been of a substantial size (greater than 5%).

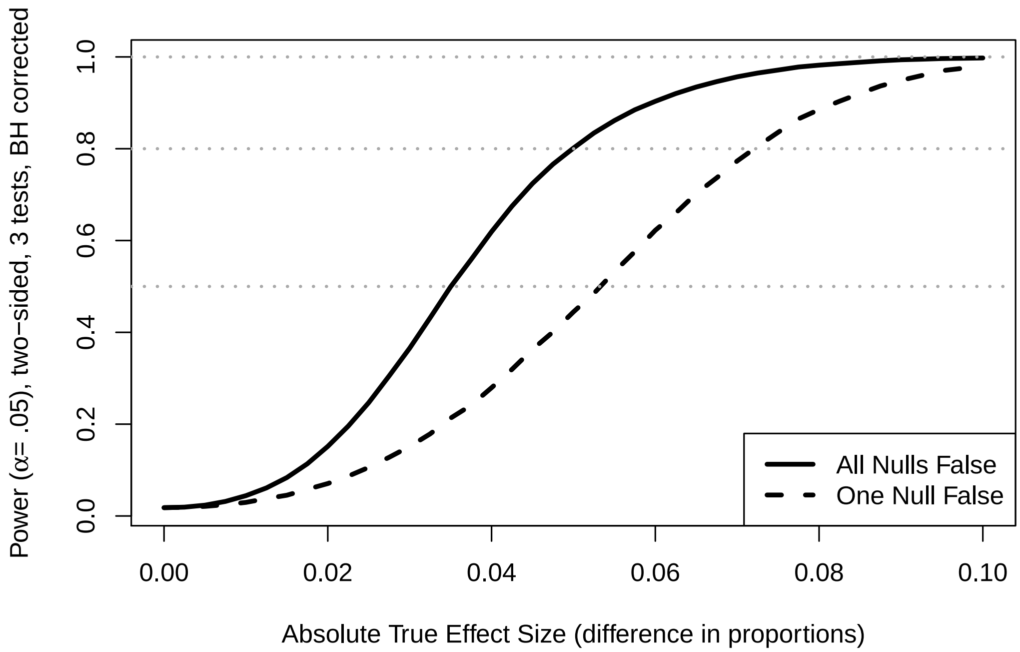
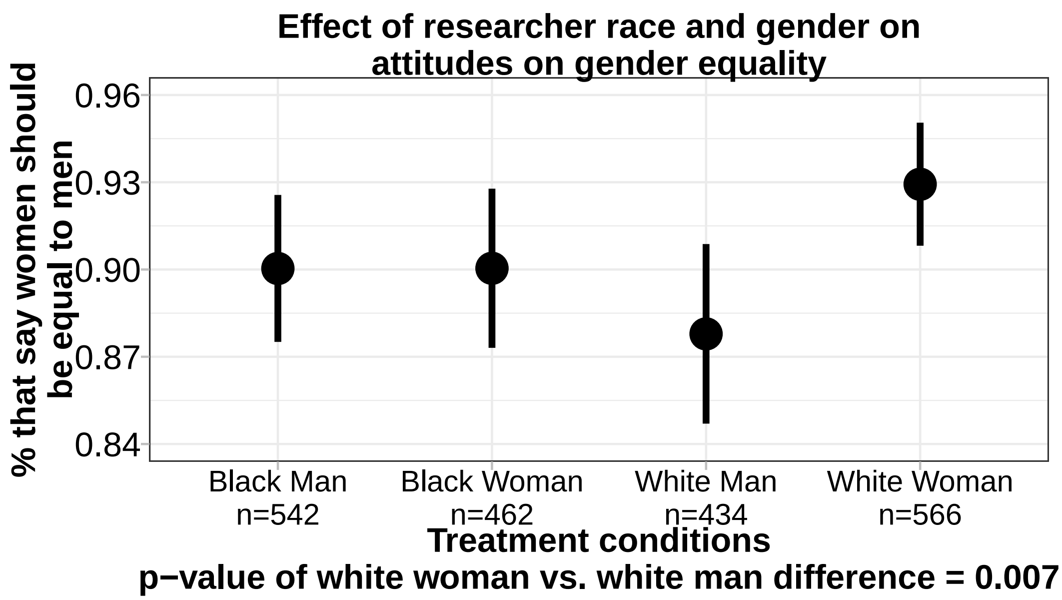


Figure A6: Power curves for three Benjamini-Hochberg adjusted large-sample two-sided difference in proportions tests. $n = 2000$ split evenly between groups.

G Exploratory Analysis of Effect Heterogeneity

Exploratory analyses provide some evidence of interesting potential effect heterogeneity when unpacking the female researcher effect on reported attitudes towards women's equality. Specifically, we find interesting variation in race. It is important to note that the analysis that follows was not registered in our pre-analysis plan. While respondents in the putative white female researcher condition are significantly more likely to state that women should be equal to men in industry, business, and government relative to respondents in the white male researcher condition, we find no such effect of gender among the black researcher conditions. Figure A7 plots the estimated proportions of respondents in all four treatment categories. The p-value of the difference between the white woman and white man conditions is quite small at .007. However, since we did not pre-specify hypotheses about race and gender interactions when registering the study, this finding should be considered exploratory. However, based on its strength, we believe it warrants further investigation.



Note: Lines denote 95% confidence intervals.

Figure A7: Exploratory analyses for effect heterogeneity across race for gender effect on attitudes towards women's equality.

H Survey Questions (Order Randomized)

Closed-Ended Survey Questions

- In the past few years, we have heard a lot about improving the position of black people in this country. How much real change do you think there has been in the position of black people in the past few years: a lot, some, or not much at all?

  Not much

  Some

  A lot

- Some people feel that women should have an equal role with men in running business, industry and government. Others feel that women's place is in the home. Where would you place yourself on this scale or haven't you thought much about this?

  Equal role

  Women's place is in the home

  Haven't thought much about this

- Between now and the 2016 Presidential Election, there will be discussion about the qualifications of presidential candidates - their education, age, race, religion, and so on. If your party nominated a generally well-qualified person for president who happened to be _____, would you vote for that person?

  Black ("Yes, would" or "No, would not")

  A woman ("Yes, would" or "No, would not")

  Catholic ("Yes, would" or "No, would not")

  Hispanic ("Yes, would" or "No, would not")

  Jewish ("Yes, would" or "No, would not")

  Mormon ("Yes, would" or "No, would not")

Gay or lesbian (“Yes, would” or “No, would not”)

Muslim (“Yes, would” or “No, would not”)

An atheist (“Yes, would” or “No, would not”)

- Some people think the government should provide fewer services, even in areas such as health and education, in order to reduce spending. Other people feel that it is important for the government to provide many more services even if it means an increase in spending. Which do you prefer?

  Cut services/spending

  More services/spending

- While taking this survey, did you engage in any of the following behaviors? Please check all that apply.

  Use your cell phone

  Browse the internet

  Talk with another person

  Watch TV

  Listen to music

- Do you agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly with these statements?

  1. Over the past few years, blacks have gotten less than they deserve. (agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly)

  2. Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors. (agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly)

3. It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites. (agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly)

4. Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class. (agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly)

I Pre-Analysis Plan

I.1 Introduction: A Second Experiment on Researcher Identity in Online Surveys

The experiment discussed in this pre-analysis plan is the second component of a broader project exploring how the perceived identity of researchers affects responses in online surveys. In particular, we analyze how information about the race and gender of the researcher, conveyed in the required consent form and researcher username, affect survey responses. While the theoretical component and hypotheses across these two experiments are identical, this second experiment presents a slightly different way of testing the hypotheses by directly varying the researcher identity via the username used to post the survey.[2] This is in contrast to the first round of the experiment in which we used a generic username called "survey researcher" and only manipulated the researcher name on the consent form.

The reasons for explicitly manipulating the researcher username are two-fold. First, this is a common practice for researchers conducting online survey experiments. Indeed, several of the authors contributing to this paper explicitly list their name when advertising for their survey. It

[2]The first experiment, with a more complete discussion of the theory and hypotheses, was also pre-registered on the Experiments in Governance and Politics website here:
`http://egap.org/design-registration/registered-designs/`.

is useful to know whether information conveyed through this common practice can affect results, in order to establish best practices for how researchers should implement survey experiments on popular online survey platforms. Second, researcher usernames are displayed prominently on online survey platforms such as Amazon's Mechanical Turk (MTurk). This means that survey respondents are more likely to be cognizant of the name of the researcher when proceeding from the survey posting to the survey itself. If, as we discuss in the related pre-registration document, we expect that perceived researcher identity can affect survey responses, then usernames represent one commonly used means through which respondents can learn the identity of the researcher.

I.2 Experimental Survey Design

Respondents will be recruited online via Amazon's Mechanical Turk (MTurk) platform. Many experimental researchers have begun using MTurk for low-cost recruitment of subjects, making it a highly relevant subject pool for investigating the effects of researcher identity within an online context. In online recruitment platforms like Mechanical Turk, accounts have an associated username. We use this username to connote race and gender.

Because the treatment is in the username, rather than in the experiment, subject recruitment must be conducted carefully. In particular, the problem is that the treatment is no longer embedded in the survey (which makes randomization relatively simple) but instead in the request for users to take the survey. We solve this problem as follows.

We pool samples from previous studies conducted by the authors on Mechanical Turk. From this existing pool of respondents, we assign each to one of eight names, listed in Table A4. Subjects are then given a qualification associated with the condition to which they are assigned. Then, we create separate MTurk accounts for each username and deploy the same survey within each account. Each deployed survey is set such that only MTurk workers with the correct qualification can see the survey (and thus the username associated with it).

To generate the names associated with each of these manipulations, we combined three commonly used lists of racially distinct first and last names. First names were drawn from a combination of lists found in Bertrand and Mullainathan (2004) and Fryer, Jr. and Levitt (2004), while last names were drawn from lists in Word et al. (2008) and Bertrand and Mullainathan (2004). We crossed the lists of first and last names to produce many possible combinations[3] and drew two names for each of the four manipulation categories (black women, white women, black men, and white men). The full list of names used in this experiment is presented in Table A4.

Black Men	Deshawn Booker	Tyrone Robinson
Black Women	Ebony Gaines	Deja Washington
White Men	Connor Schroeder	Brett Walsh
White Women	Molly Ryan	Laurie Yoder

Table A4: Names used for each of the four investigator name manipulations, based on lists from Bertrand and Mullainathan (2004), Fryer, Jr. and Levitt (2004), Word et al. (2008)

Within the survey, respondents are asked a series of questions about social and political attitudes. We draw questions from Pew, Gallup, and the American National Election Survey specifically asking about issues for which racial and gender cues may prompt different responses. The full text of the outcome questions is presented in Appendix H. We also explore the extent to which respondents are paying attention and are willing to put in effort using attention checks and open text responses. The general structure of the attention checks used in the experiment is taken from Berinsky, Margolis and Sances (2014). Respondents are also asked to complete a randomly assigned writing task, either on their attitudes towards a female president or on a time in their life when they were affected by politics. The latter prompt is sufficiently general that variation in response depth will capture respondent's general "effort" levels rather than attitudes towards a particular issue. In order to obscure the general purpose of the survey, we randomly permute some of the demographic questions with the outcome questions. However, to avoid priming party affiliation, gender or race, we leave the party ID, gender, and race questions for the end of the survey.

[3]We omitted a few randomly-generated names that already belonged to celebrities, such as Jermaine Jackson.

I.3 Statistical Tests

For Hypothesis 1, we estimate two separate treatment effects.[4] The first is the effect of assignment to a putatively female name on the probability that a respondent indicates that they believe that women should have an equal role in the workforce. The second is the effect of assignment to a putatively black name on the respondent's racial resentment scale. We expect that the effect for the former will be positive while the latter will be negative. For Hypothesis 2, we will estimate the effect of assignment to a putatively white and male name on the probability that a respondent correctly completes both of the attention check assignments. We expect this effect estimate to be positive. For estimation, we will fit a linear probability model of the outcome on treatment and compute standard errors via a nonparametric bootstrapping procedure. While not needed for identification, we will include respondent-level covariates (e.g. gender, income, education) in the regression model in order to increase the efficiency of our estimator.

Our rejection levels for two-sided hypothesis tests of whether the average treatment effects differ from zero are calibrated to correct for problems of multiple testing. We are willing to tolerate an overall Type I error rate of $\alpha = .05$. With three main hypothesis tests, we could obtain a conservative rejection threshold for each individual hypothesis test of $.05/3 = .017$ using the Bonferroni correction. This controls the Familywise Type I Error Rate and guarantees that the probability of any single erroneous rejection in the set of tests is less than or equal to .05. However, this approach sacrifices a significant amount of power. A less conservative but more powerful approach is to set a rejection threshold to control the False Discovery Rate (FDR). We use the Benjamini-Hochberg procedure to set a rejection level for the hypothesis tests (Benjamini and Hochberg, 1995). This entails a two-step procedure where we order the 3 p-values of the individual hypothesis tests from smallest to largest, $p_{(1)}, ..., p_{(3)}$ and then set our rejection level to $p_{(k)}$, where k is the largest value of

[4] Note that the hypotheses referred to in this section reference those laid out in the first preregistration document, which are as follows. H_1: Assignment to a putatively female/black investigator will increase support for policies that provide for and protect the rights of women/blacks. H_2: Attention and effort will be greatest among subjects assigned to a putatively white, male investigator.

i that satisfies $p_{(i)} \leq \frac{i}{3}\alpha$. This procedure controls the expected share of false hypothesis rejections out of the total number of rejections to be no greater than .05.

We do not specify any ex-ante interactions of the treatment effects with baseline covariates. However, because the mechanism through which any treatment effects operate are of significant interest, we will conduct exploratory analyses of potential treatment effect heterogeneity by estimating models with interactions between treatment and respondent identity variables. Among other interactions, we are interested in seeing whether any average treatment effect is primarily driven by behavior changes among men (in the case of the gender treatment) and white respondents (in the case of the race treatment). We will attempt to replicate any promising results from these exploratory analyses in a follow-up experiment that explicitly registers interactive hypotheses prior to the experiment.

Additionally, because respondents have the option to stop taking the survey after treatment is assigned but before outcomes are measured, there is concern that an analysis conditional on survey completion will be biased for the average treatment effect if treatment also affects the probability that a respondent will drop out. Although it is not possible to adjust for nonignorable drop-out in the absence of prior covariates on respondents, we will examine whether there appears to be systematic differences between treatment arms with respect to attrition and employ sensitivity analyses in the vein of Scharfstein, Rotnitzky and Robins (1999) in order to evaluate the robustness of our estimates to this potential source of bias.

I.4 Appendix: Survey Text

I.4.1 Informed Consent

Thank you for participating in this survey. Please take time to answer questions honestly and thoroughly. Your responses are essential to our research.

This research is being conducted under the supervision of BLACK/WHITE x MALE/FEMALE NAME at Harvard University. All of the information that we obtain from your session will be anonymous. We do not ask you for your name. Your name or identifying information will not be used in any reports of the research. There will be no direct benefit to you from participation in this study other than the agreed-upon financial compensation. We hope, however, that the research will benefit society by improving our understanding of the factors that influence people's decision making.

If you want to receive the findings of this study, you may contact NAME HERE (GENERIC EMAIL). Complete contact information is as follows.

NAME HERE

Department of Government

Harvard University

Cambridge, MA 02138

GENERIC EMAIL HERE

If you have questions about your rights or about research-related harm, or if your questions, concerns, suggestions, or complaints are not being addressed by the researchers above, please contact:

Director of IRB Operations

Harvard University Committee on the Use of Human Subjects in Research

1414 Massachusetts Avenue, Second Floor

Cambridge, MA 02138

Phone: 617-496-5593

jjaeger@fas.harvard.edu

The nature and purpose of this study have been satisfactorily explained to me and I (participant) agree to become a participant in the study described above. I understand that I am free to discontinue participation at any time if I so choose. Answering yes continues the survey, answering no ends the survey.

Yes

No

I.4.2 Outcome Questions [Question order randomized]

- Do you agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly with these statements?

  - Over the past few years, blacks have gotten less than they deserve.

  - Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.

  - It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites.

  - Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class.

- Some people feel that women should have an equal role with men in running business, industry and government. Others feel that women's place is in the home. Where would you place yourself on this scale or haven't you thought much about this?

  Equal role

  Womens place is in the home

  Havent thought much about this

- Between now and the 2016 Presidential Election, there will be discussion about the qualifications of presidential candidates - their education, age, race, religion, and so on. If your party nominated a generally well-qualified person for president who happened to be ____, would you vote for that person?

Black ("Yes, would" or "No, would not")

A woman ("Yes, would" or "No, would not")

Catholic ("Yes, would" or "No, would not")

Hispanic ("Yes, would" or "No, would not")

Jewish ("Yes, would" or "No, would not")

Mormon ("Yes, would" or "No, would not")

Gay or lesbian ("Yes, would" or "No, would not")

Muslim ("Yes, would" or "No, would not")

An atheist ("Yes, would" or "No, would not")

- Some people think the government should provide fewer services, even in areas such as health and education, in order to reduce spending. Other people feel that it is important for the government to provide many more services even if it means an increase in spending. Which do you prefer?

  Cut services/spending

  More services/spending

Demographic Questions [Order permuted with outcome questions]

- What is the highest level of education you have completed?

  Less than High School

  High School / GED

  Some College

  2-year College Degree

  4-year College Degree

Masters Degree

Doctoral Degree

Professional Degree (JD, MD)

- What is your yearly household income, putting together the income of all the members of your household?

  Less than 30,000

  30,000 39,999

  40,000 49,999

  50,000 59,999

  60,000 69,999

  70,000 79,999

  80,000 89,999

  90,000 99,999

  100,000 or more

Attention Checks [permuted with the outcome questions]

- When a big news story breaks people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to this survey. To show that you've read this much, please ignore the question and select Reuters website and Huffington Post as your two answers.

  When there is a big news story, which is the one news website you would visit first? (Please only choose one)

New York Times website	The Drudge Report	The Associated Press (AP) website
Huffington Post	Google News	Reuters website
Washington Post website	ABC News website	National Public Radio (NPR) website
CNN.com	CBS News website	USA Today website
FoxNews.com	NBC News website	New York Post Online
MSNBC.com	Yahoo! News	None of these websites

- We are very interested to know what political issues people think are the most relevant today. People often have different attitudes about what issues the United States government should focus on addressing and we would like to understand more about this public debate. We also want to know if people are paying attention to this question. To show that you've read this much, please ignore the question and select Energy and Global trade as your two answers.

  Which of the following issues do you think should be the highest priority for President Obama and Congress in 2015? (Please only choose one)

Terrorism	Reducing crime	Energy
Economy	Poor and needy	Influence of lobbyists
Jobs	Military	Transportation
Education	Immigration	Money in politics
Social Security	Environment	Scientific research
Budget deficit	Race relations	Global warming
Health care costs	Moral breakdown	Global trade
Medicare	Tax reform	None of these issues

Open Response Questions

[Randomly assign respondents to one of the two questions below]

- Please write a few sentences about what you think about the United States potentially having a female president.

- Please write a few sentences about a time that politics affected your life.

Demographic Questions [placed at end of survey]

- Generally speaking, do you consider yourself to be a(n):

  Democrat

  Republican

  Independent

- → [if independent] As of today do you lean more to the Republican Party or more to the Democratic Party?

  Republican

  Democrat

- In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you – did you vote in the last presidential election in 2012?

  I don't remember.

  No, I did not vote.

  Yes, I voted.

- → [if yes] Who did you vote for in the last presidential election?

  Mitt Romney

  Barack Obama

Other (specify)

- What is your ethnicity? Select all that apply.

  Black

  White

  Hispanic

  Asian

  Native American

  Other (specify)

- What is your gender?

  Male

  Female

  Other (please specify)

- What do you think is the purpose of this research?

  [Open Response]

- While taking this survey, did you engage in any of the following behaviors?

  Use your cell phone

  Browse the internet

  Talk with another person

  Watch TV

  Listen to music

Debriefing (sent to all subjects by email AFTER all responses are collected)

The purpose of this study was to learn about how researcher identity influences subjects' responses. In order to learn about this, we needed to manipulate the identity of the researcher. Though the contact information you were given was in fact real, the name of the researcher was not. This study was conducted by Connor Huff, Dominika Kruszewska, Christopher Lucas, Anton Strezhnev, and Ariel White, all at Harvard University. If you have any questions about this study, please contact them at the following location.

Connor Huff, Dominika Kruszewska, Christopher Lucas, Anton Strezhnev, Ariel White

Department of Government

Harvard University

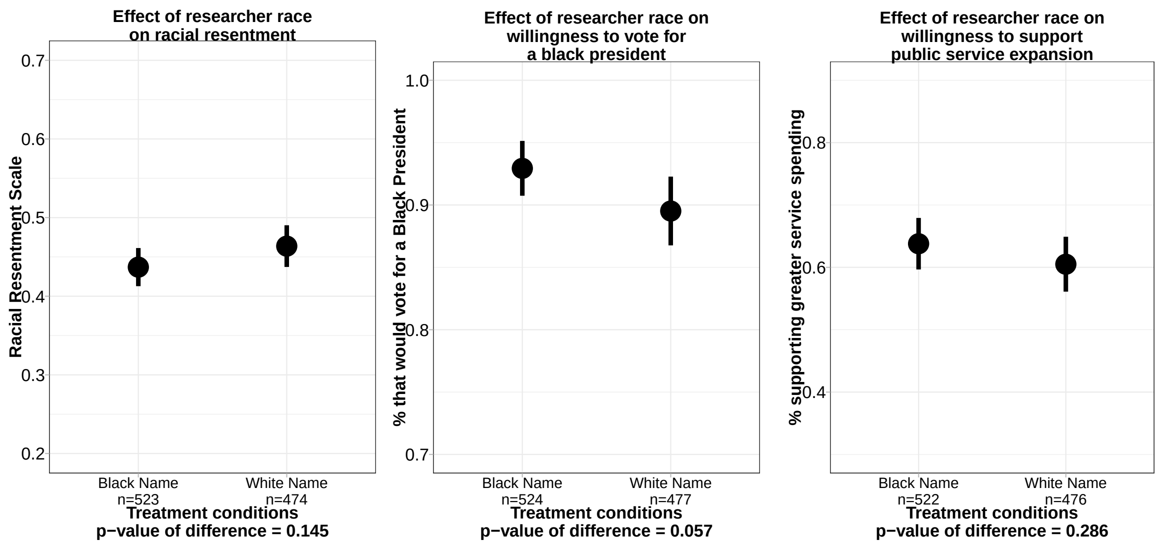
Cambridge, MA 02138

GENERIC EMAIL HERE

J A Second Experiment: Randomization in Only the Consent Form

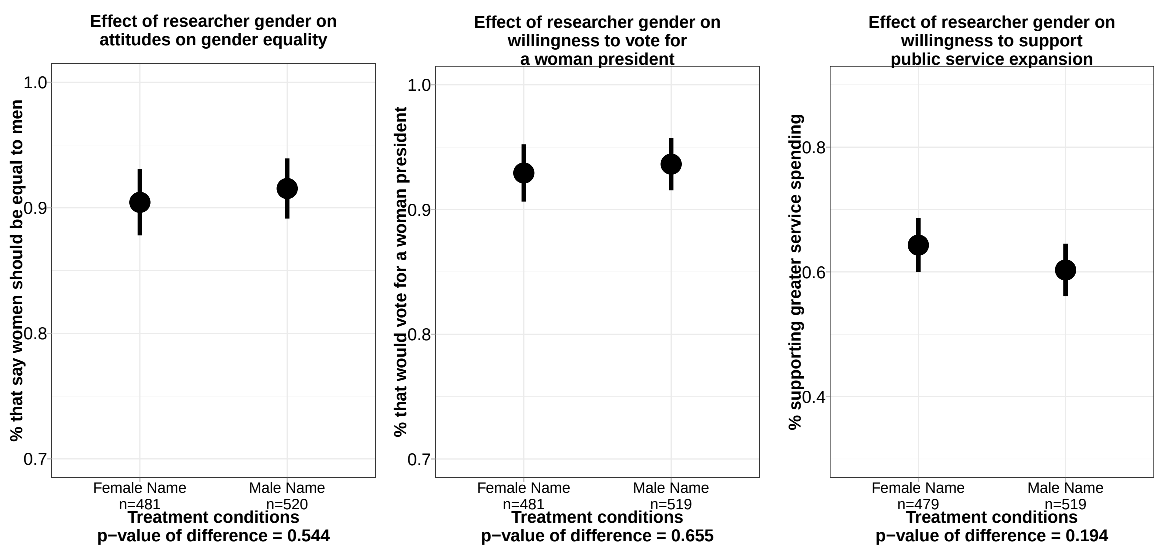
In addition to the experiment presented in this paper, we ran a second experiment in which we used a generic account name to post the initial HIT and then varied the researcher name only in the consent form.[5] This allowed us to (1) test how respondents react to researcher identity when presented with a more subtle treatment, and (2) estimate the effect of researcher name on survey responses without the initial selection stage into the survey conditional on the name posted in the HIT. This second experiment was also fielded on Amazon's Mechanical Turk on a sample of 1000 respondents. We found no statistically discernible effect of treatment on any of the outcomes of interest. The full results for the pilot experiment are presented in Figures A8-A11.

[5]This experiment was pre-registered on EGAP with study ID 20150717AA.



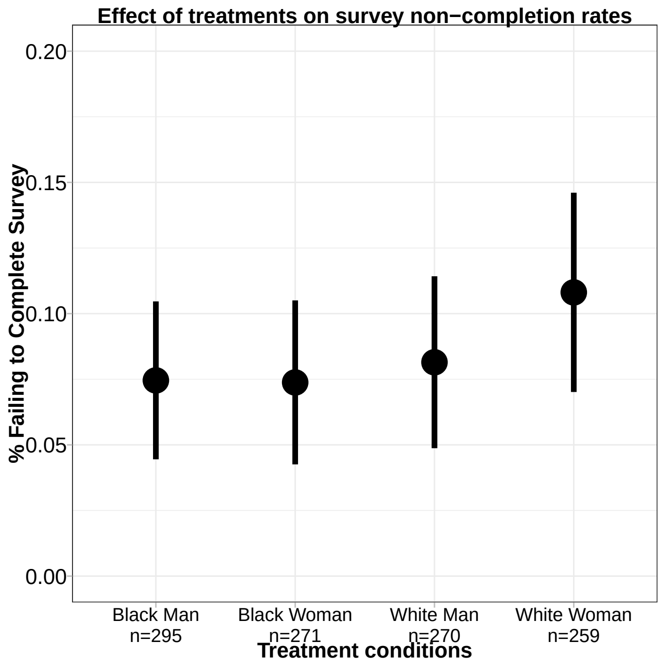
Note: Lines denote 95% confidence intervals.

Figure A8: Differences in policy/attitude outcomes for researcher race treatment in consent form only randomization experiment.



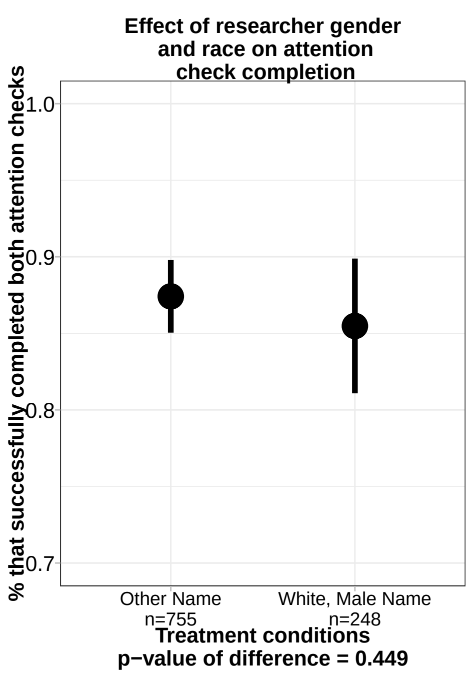
Note: Lines denote 95% confidence intervals.

Figure A9: Differences in policy/attitude outcomes for researcher gender treatment in consent form only randomization experiment.



Note: Lines denote 95% confidence intervals.

Figure A10: Estimated drop-out rates for experiment 1 across treatment conditions in consent form only randomization experiment.



Note: Lines denote 95% confidence intervals.

Figure A11: Differences in attention check outcomes for researcher race/gender treatment in consent form only randomization experiment.

References

Benjamini, Yoav and Wei Liu. 1999. "A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence." *Journal of Statistical Planning and Inference* 82(1):163–170.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

Berinsky, Adam J., Michele F. Margolis and Michael W. Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

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

Fryer, Jr., Roland G. and Steven J. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *The Quarterly Journal of Economics* CXIX(August):767–806.

Scharfstein, Daniel O, Andrea Rotnitzky and James M Robins. 1999. "Adjusting for nonignorable drop-out using semiparametric nonresponse models." *Journal of the American Statistical Association* 94(448):1096–1120.

Word, David L., Charles D. Coleman, Robert Nunziata and Robert Kominski. 2008. "Demographic Aspects of Surnames from Census 2000." `https://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf`.