Investigator Characteristics and Respondent Behavior in Online Surveys *

Ariel White[†] Anton Strezhnev[‡] Christopher Lucas[§]
Dominika Kruszewska[¶] Connor Huff[‖]

April 16, 2016

Abstract

Prior research has demonstrated how the results of surveys can vary depending on the race, gender, or ethnicity of the investigator asking the question. This paper extends the logic of that research to empirically test how information about researcher identity in *online surveys*, conveyed in the advertisement for the experiment and on the informed consent page, affects subject responses. We do so by conducting an experiment on Amazon's Mechanical Turk in which we vary the name of the researcher 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, 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.

[†]arwhite@fas.harvard.edu

[‡]astrezhnev@fas.harvard.edu

[§]clucas@fas.harvard.edu

[¶]dkruszewska@fas.harvard.edu

[‖]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 depend 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). These findings are most salient for questions about group attitudes or policies like affirmative action where attributes of the researcher substantively interact with the survey questions of interest.[1] These considerations have played a central role in the design and implementation of a range of survey-based research.[2]

In this paper we extend the logic of prior research to 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 has demonstrated how inferences made from names can affect behavioral outcomes even in the absence of in-person or telephone interactions.[3]

In this paper we experimentally test whether researcher identity, as cued through researcher names, shapes a range of online survey behaviors. In doing so, we contribute to an expanding

[1]This is generally argued to occur through two main mechanisms. First, provision of information about the investigator could create demand effects whereby the subjects guess at the purpose of the study or the interviewer's views and change their responses to align with this perceived purpose (concerns about social desirability bias also fall into this category). Second, potential subjects may be more or less comfortable speaking to researchers with a particular identity and select into the sample accordingly.

[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. 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.

[3]See for example, Bertrand and Mullainathan (2004), Butler and Broockman (2011), White, Nathan and Faller (2015), Einstein Levine and Glick (N.d.), or Edelman, Luca and Svirsky (2015).

strain of research exploring the composition and attributes of online survey pools.[4] Given the prominence of online surveys as a key source of data for a range of published research throughout top political science journals,[5] these studies contribute to better understanding how we should interpret the substantive results of these prior studies while also providing 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. While it is important for researchers to consider how information given to subjects could affect survey behavior, we believe researchers need not worry that their names will substantively affect online survey results.

2 Experimental Design

The experiment proceeds as follows. 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,

[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 (N.d.).

[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 F we present results from a different experiment in which we randomize the name of the researcher *only on the consent 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.

we combined three commonly 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. Many researchers post studies on platforms such as MTurk under their own names, so the nature of treatment is consistent with common practices for researchers using the MTurk survey pool. It also 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 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.

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

[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).

[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 B.

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

"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 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.

[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 (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.

[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 G.

3 Results

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 fully saturated linear regression models with robust standard errors. Following our pre-analysis plan, our rejection levels for accepting that the effects differ from zero are calibrated to yield an expected number of Type I errors of $\alpha = .05$, adjusting for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).[14] This adjustment is important since we dramatically increase the chances of a false positive finding by testing for multiple outcomes (Benjamini and Hochberg, 1995). 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.[15]

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.[16] 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 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

[14] 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 never reject for any p-value $> .05$.

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

[16] In Appendix Sections A-C, 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. As we discuss in the Appendix, the discussion of these results is exploratory since we did not preregister any hypotheses for these outcome questions.

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 racialized or gendered ways depending on the putative race or gender of the researcher.

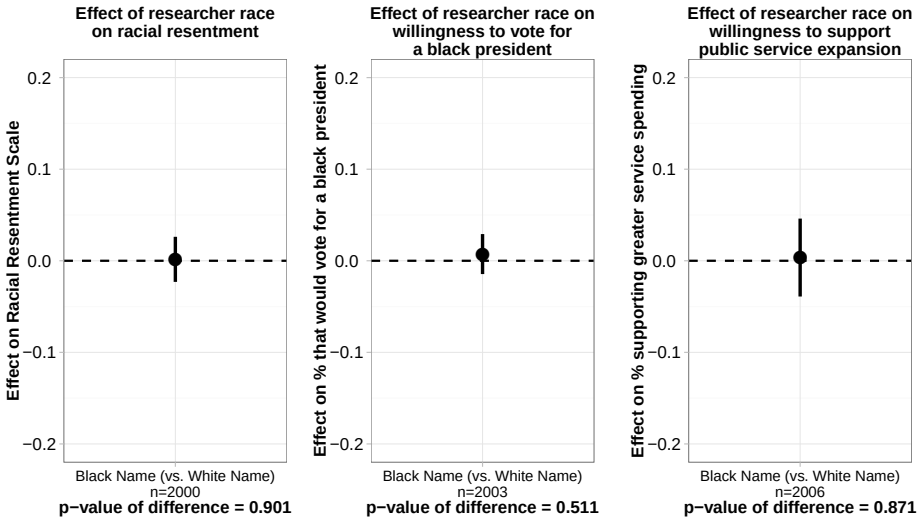
We designed our experiment to target a sample size of 2000 total respondents.[17] For the race treatment, we find no evidence that black versus white researcher names yield different responses on the outcome questions. Figure 1 plots the expected difference in outcomes for each of these three questions 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-values for the null of no effect on the gender equality and woman president outcome questions fell just below the commonly used threshold of .05. The p-value for the null for support for a woman president, however, falls just above the typically used threshold. Under our pre-registered design, using the Benjamini-Hochberg correction for multiple testing, we fail to reject the null for all three outcomes.[18] We cannot conclude that assignment to a putatively female researcher name significantly increased the likelihood that respondents would

[17] Our final sample consists of 2006 unique respondents that we could confirm had completed the overall survey.

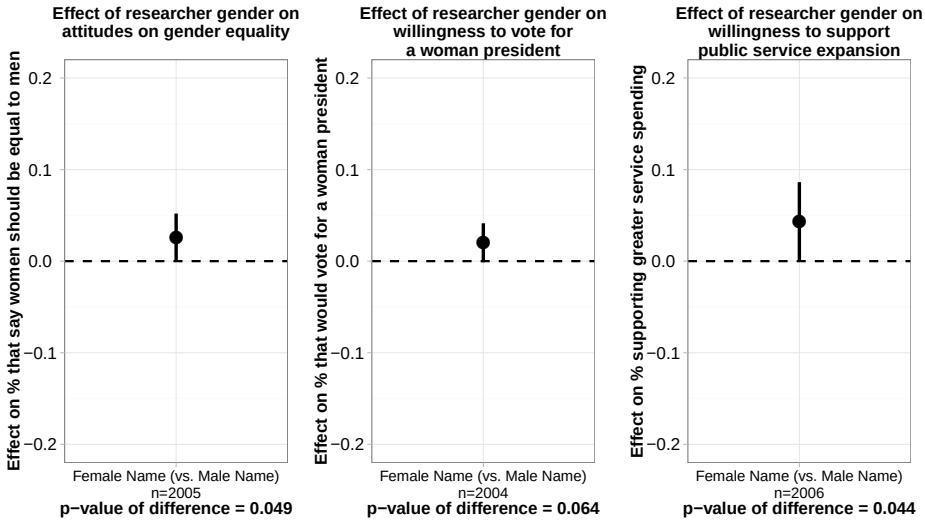
[18] This is because the threshold level for rejecting the null of no effect on only the gender equality and service spending outcomes falls to $\alpha = .05 \times \frac{2}{3} = .033$ – below the p-values that we observe.

exhibit more woman-friendly attitudes on gender-related questions.



Note: Lines denote 95% confidence intervals.

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



Note: Lines denote 95% confidence intervals.

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

Despite our failure to reject the null, it is noteworthy that the point estimates for the direction of the effect are consistent with our original hypothesis – respondents assigned to a putatively female investigator were in-sample more likely to express beliefs that were more supportive of women's equality. Given our study's power and null result, what can we say about the range of potential "true" effects? Using power calculations from our pre-analysis stage, we can conclude that it is extremely unlikely that the true effect on any of these outcomes

is greater than 4 percentage points.[19] Not only did we fail to reject the null hypothesis, the power of our design makes it very unlikely that this would have happened if the true effect were substantial (greater than 4%). While it is not possible to "affirm" a null hypothesis, the high power of our study is such that our null finding provides a meaningful bound on any true effect – no more than 4 or 5 percentage points.

4 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. We fail to reject the null hypothesis that researchers' race or gender (cued through names) do not affect 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.

References

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.

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.

[19] For a more detailed discussion of the power calculations see Appendix Section D.

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* .

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):2053168015622072.

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).

Einstein Levine, Katherine and David M. Glick. N.d. "Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing." *American Journal of Political Science*. Forthcoming.

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.

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 Respnse." *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).

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* .

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):2–7.

Leeper, Thomas and Emily Thorson. N.d. "Minimal Sponsorship-Induced Bias in Web Survey Data." *Working Paper*. Forthcoming.

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):2158244016636433.

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.

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. **URL:** *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 PR. 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* (February):1–14.

Word, David L, Charles D Coleman, Robert Nunziata and Robert Kominski. 2008. "Demographic aspects of surnames from census 2000." *Unpublished manuscript* .

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

Ariel White* Anton Strezhnev† Christopher Lucas‡
Dominika Kruszewska§ Connor Huff¶

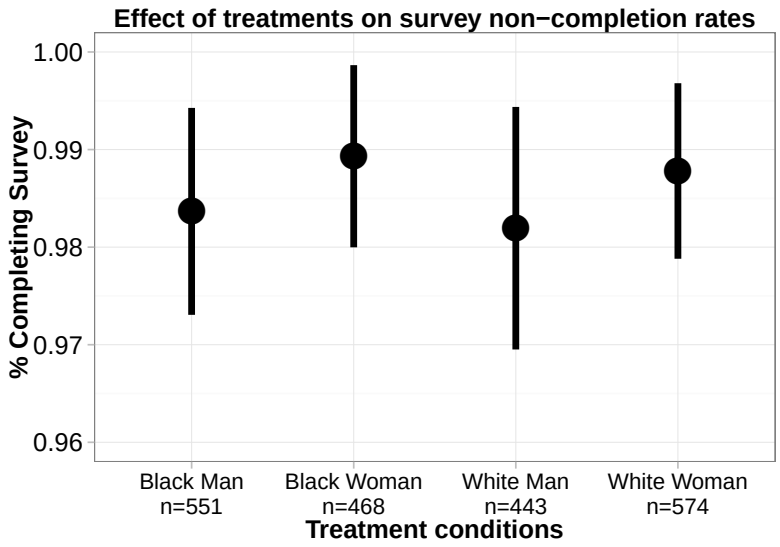
April 16, 2016

A 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. Because of our randomization scheme (we are unable to block), we have a slight imbalance in the number of observations under any treatment condition. To some extent, this selection effect should be considered a part of the overall effect of researcher identity – assignment to one name relative to the other could affect the distribution of responses not only by changing individual level behavior, but also by changing the composition of respondents who ultimately take up the survey. However, we also may be concerned about differential completion rates between treatment categories among those respondents who chose to begin our survey. We do not find strong evidence that this is the case. Figure 1 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

*arwhite@fas.harvard.edu
†astrezhnev@fas.harvard.edu
‡clucas@fas.harvard.edu
§dkruszewska@fas.harvard.edu
¶cdezzanihuff@fas.harvard.edu

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.



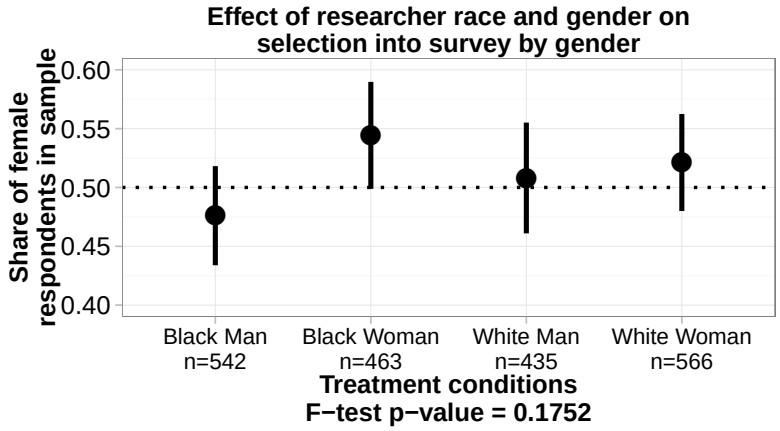
Note: Lines denote 95% confidence intervals.

Figure 1: Completion rates conditional on starting survey across four treatment categories.

B 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.

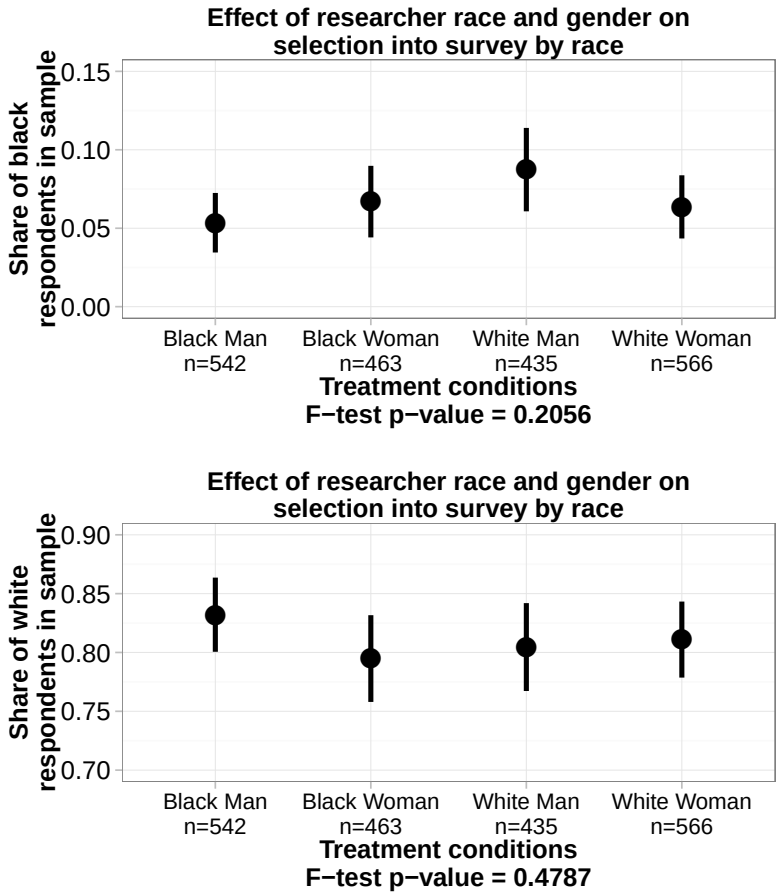
We do find evidence that some of our Mechanical Turk HITs were more likely to be taken by respondents. We test the null hypothesis that the observed counts in each of the four treatment categories are statistically indistinguishable from a uniform distribution by calculating a test statistic, the difference 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 0.00012, 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 2: Differences in treatment groups across respondents' background covariates: Gender

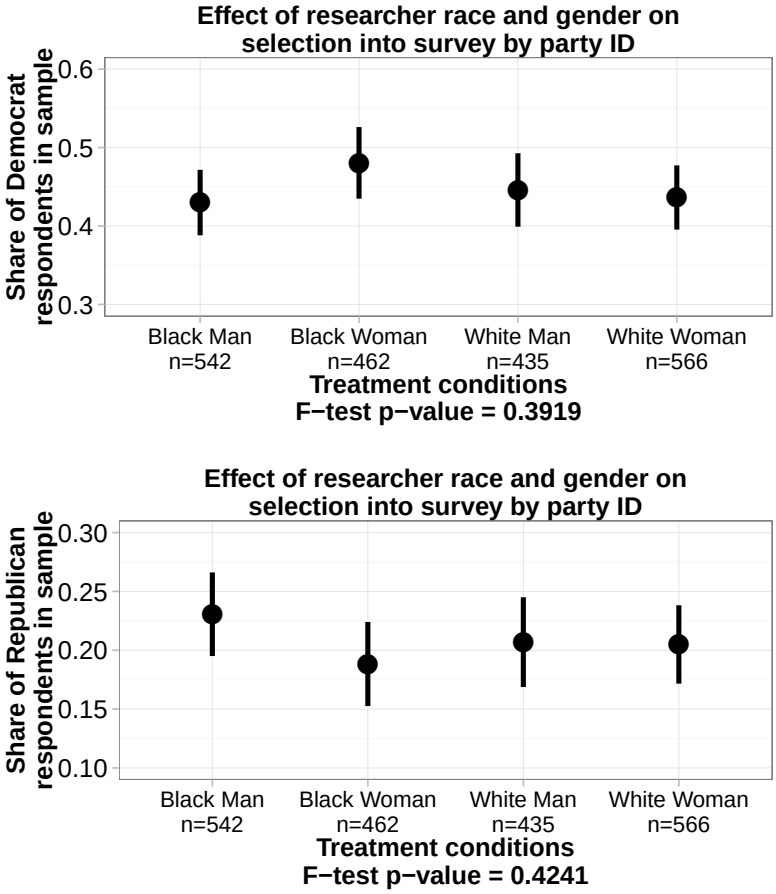
However, the magnitude of this difference is not particularly large. And more importantly, there do not appear to be significant differences between the samples for any of the four treatment categories based on the background covariates that we collected. Figure 2 plots the share of female



Note: Lines denote 95% confidence intervals.

Figure 3: 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 3) and for political party identification (Figure 4). If there is some interesting selection into treatment, it is not clear that it relates to any of these commonly observed background covariates.

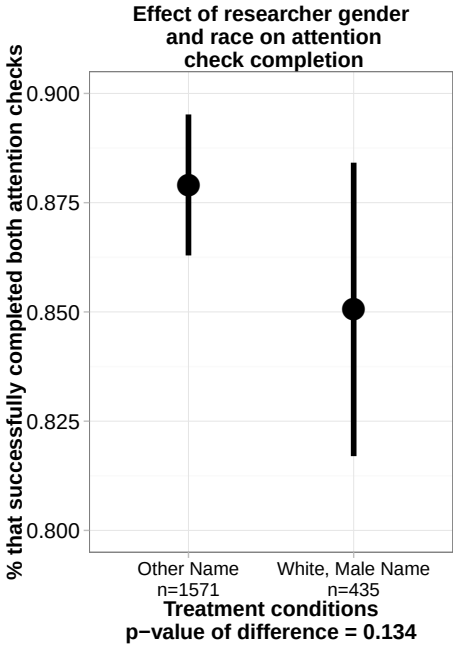


Note: Lines denote 95% confidence intervals.

Figure 4: Differences in treatment groups across respondents' background covariates: Party ID

C 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 5 reports the estimated probabilities of completing both attention checks correctly under the white male and non-white male treatments. Roughly 90-92 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 5: Differences in attention check outcomes for researcher race/gender treatment.

D Power Calculations

Figure 6 plots a simple power curve for a hypothetical difference-in-means test between two proportions. We consider $n = 2000$ evenly split between treatment and control and estimate the pooled population variance of the outcome using the sample variance from our experiment. The y-axis denotes the probability of rejecting the null hypothesis and the x-axis plots the true absolute effect size. Supposing a single two-sided z-test at $\alpha = .05$, we see that our study would be essentially guaranteed (with probability $\approx .99$) to detect any effect greater than 6 percentage points. With a relatively "high" power of .8, we would detect an effect greater than or equal to 4 percentage points. So even though we failed to reject the null hypothesis, the power of our design is such that it is very unlikely that this would have happened had the true effect been of a substantial size (greater than 4%). While it is not possible to "affirm" a null hypothesis, the high power of our study is such that our null finding provides a meaningful bound on any true effect – no more than 4 or 5 percentage points.

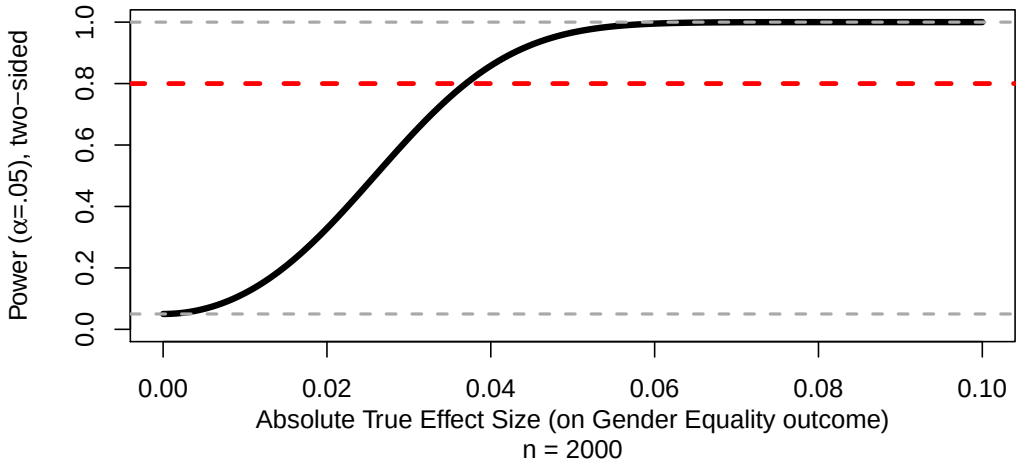
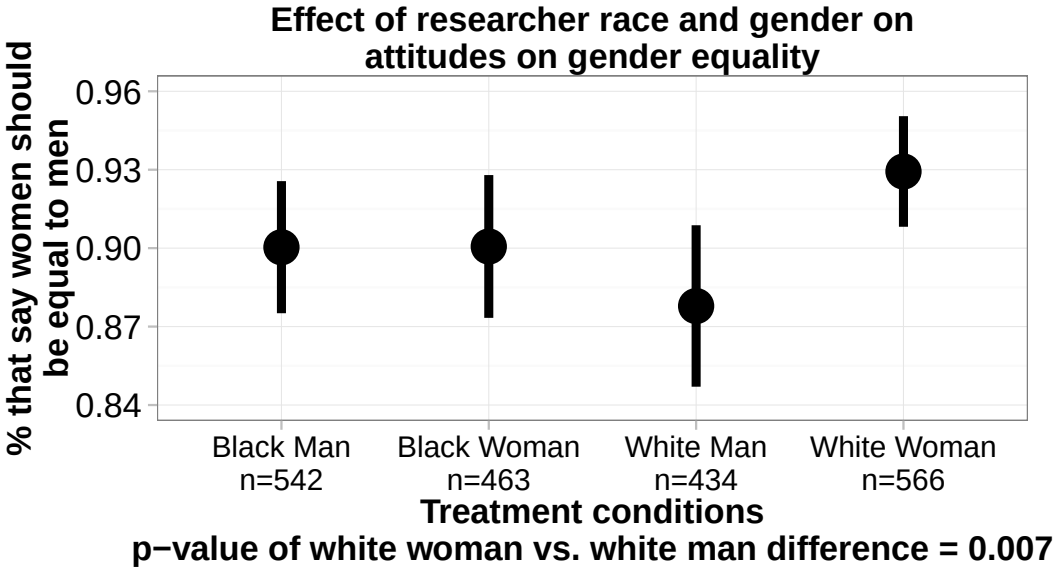


Figure 6: Power curve for large-sample two-sided difference in proportions given $n = 2000$ split evenly between groups.

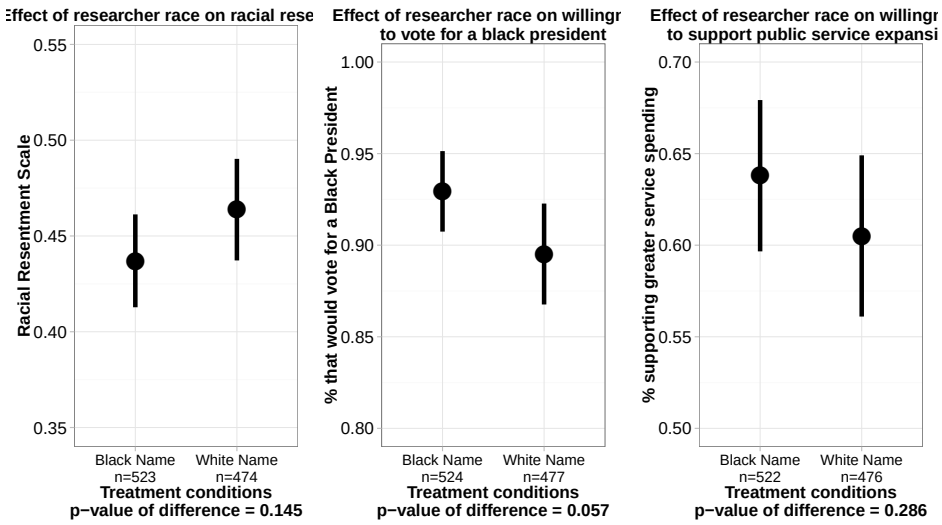
E Exploratory Analysis of Effect Heterogeneity

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 7 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 7: Exploratory analyses for effect heterogeneity across race for gender effect on attitudes towards women's equality.

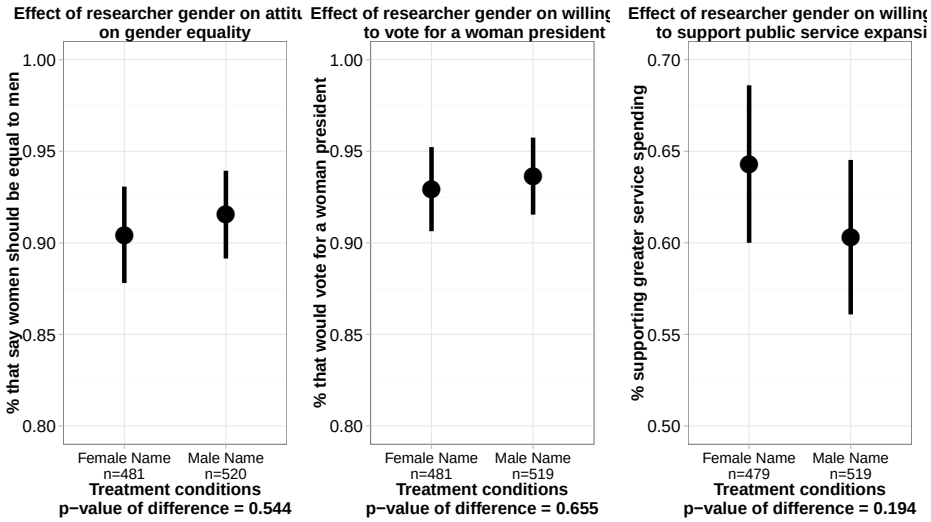


Note: Lines denote 95% confidence intervals.

Figure 8: Differences in policy/attitude outcomes for researcher race treatment in consent form only randomization experiment.

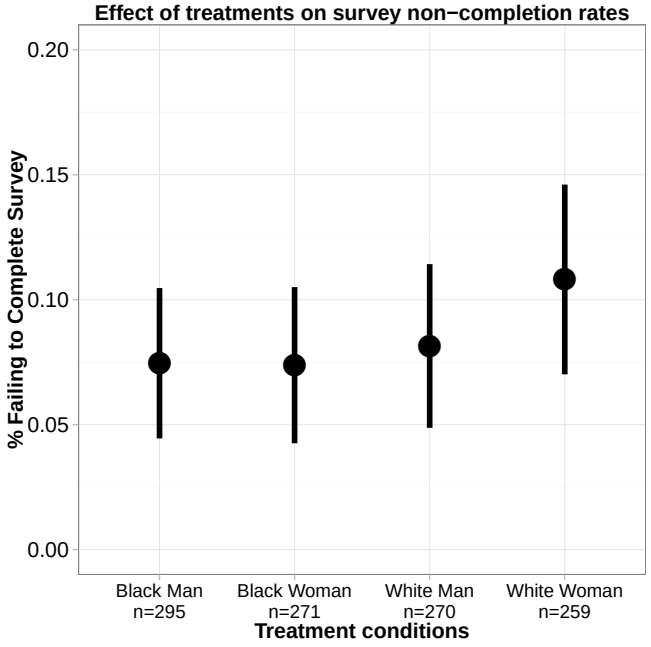
F A Second Experiment: Randomization in Only the Consent Form

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. 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 Experiment 1 are presented in Figures 8-11.



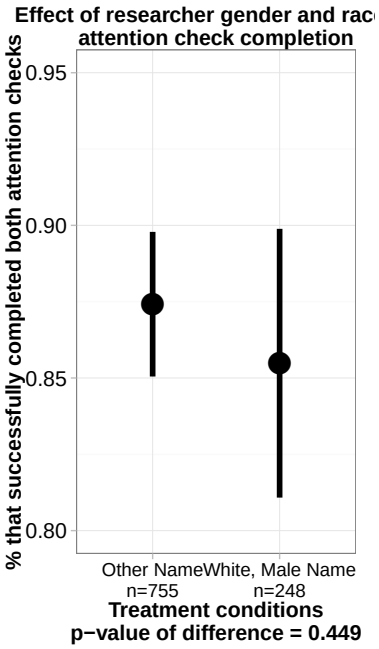
Note: Lines denote 95% confidence intervals.

Figure 9: Differences in policy/attitude outcomes for researcher gender treatment in consent form only randomization experiment.



Note: Lines denote 95% confidence intervals.

Figure 10: Estimated drop-out rates for experiment 1 across treatment conditions in consent form only randomization experiment.



Note: Lines denote 95% confidence intervals.

Figure 11: Differences in attention check outcomes for researcher race/gender treatment in consent form only randomization experiment.

G 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)