*American Political Science Review*, Page 1 of 13
doi:10.1017/S0003055418000837 

# Demand Effects in Survey Experiments: An Empirical Assessment

JONATHAN MUMMOLO *Princeton University*
ERIK PETERSON *Texas A&M University*

*Survey experiments are ubiquitous in social science. A frequent critique is that positive results in these studies stem from experimenter demand effects (EDEs)—bias that occurs when participants infer the purpose of an experiment and respond so as to help confirm a researcher's hypothesis. We argue that online survey experiments have several features that make them robust to EDEs, and test for their presence in studies that involve over 12,000 participants and replicate five experimental designs touching on all empirical political science subfields. We randomly assign participants information about experimenter intent and show that providing this information does not alter the treatment effects in these experiments. Even financial incentives to respond in line with researcher expectations fail to consistently induce demand effects. Research participants exhibit a limited ability to adjust their behavior to align with researcher expectations, a finding with important implications for the design and interpretation of survey experiments.*

A long-standing critique of social science experiments is that evidence which supports researcher expectations is an artifact of "experimenter demand effects" (EDEs) (Iyengar 2011; Orne 1962; Sears 1986; Zizzo 2010). The concern is that experimental subjects infer the response researchers expect and behave in line with these expectations—and differently than they otherwise would. The result is biased evidence that supports a researcher's hypotheses only due to the efforts of subjects. Concern over EDEs and related phenomena (e.g., so-called "Hawthorne" effects[1]) is evidenced by the considerable effort researchers expend guarding against them. These countermeasures range from subtle attempts to disguise experimental treatments and key outcome measures, to deceptive statements aimed at masking a study's intent.

While the concept of EDEs originated to critique laboratory experiments in psychology (Orne 1962), the threat they pose is now highly relevant for political science given the widespread use of survey experiments across the field (see, e.g., Gaines, Kuklinski, and Quirk 2007 and Mutz 2011). A particular concern is that survey experiments frequently utilize online subject pools, such as Amazon's Mechanical Turk, with a potentially high capacity to produce demand effects. Respondents in these settings often have extensive prior experience participating in social science research and are attentive to researcher expectations to ensure they receive positive assessments of their performance and compensation for their work (Goodman, Cryder, and Cheema 2013; Krupnikov and Levine 2014). In a highly influential study,[2] Berinsky, Huber, and Lenz (2012, 366) recommend researchers avoid revealing their intentions in online survey experiments due to concerns about EDEs (see also Paolacci and Chandler 2014, 186). They write:

> M-Turk respondents…may also exhibit experimental demand characteristics to a greater degree than do respondents in other subject pools, divining the experimenter's intent and behaving accordingly (Orne 1962; Sears 1986). To avoid this problem and the resulting internal validity concerns, it may be desirable to avoid signaling to subjects ahead of time the particular aims of the experiment. Demand concerns are relevant to any experimental research, but future work needs to be done to explore if these concerns are especially serious with respect to the M-Turk respondent pool…

If present, EDEs could undermine experimental results in an array of major literatures in political science. Yet there is little evidence demonstrating (1) the existence of EDEs in survey experiments or (2) the degree to which EDEs distort findings from these studies (but see de Quidt, Haushofer, and Roth 2018 and White et al. 2018).

Replicating five prominent experimental designs that span all empirical subfields of political science, we assess the severity and consequences of demand effects by randomly assigning participants to receive information about the purpose of each experiment before participating. This information takes various forms across the different studies and includes hints about the focus of the experiment, explicit statements that relay the

Jonathan Mummolo, Assistant Professor of Politics and Public Affairs, Department of Politics and Woodrow Wilson School of Public and International Affairs, Princeton University, jmummolo@princeton.edu.

Erik Peterson, Assistant Professor of Political Science, Department of Political Science, Texas A&M University, erik.peterson@tamu.edu.

The authors are grateful for feedback from Adam Berinsky, Cheryl Boudreau, Amber Boydstun, John Bullock, Brandice Canes-Wrone, Justin Esarey, Justin Grimmer, Erin Hartman, Samara Klar, Neil Malhotra, Nolan McCarty, Tali Mendelberg, Sean Westwood, and attendees of the 2017 Society for Political Methodology (PolMeth) annual meeting. Replication materials can be found on the American Political Science Review Dataverse at: https://doi.org/10.7910/DVN/HUKSID.

Received: April 27, 2018; revised: September 28, 2018; accepted: November 6, 2018

[1] The terms "Hawthorne" and "demand" effects are often used interchangeably. We view them as related but distinct, with "Hawthorne" effects denoting changes in behavior due to the knowledge one is being observed, and EDEs referring to participants' efforts to validate a researcher's hypotheses. We also distinguish EDEs from a "social desirability" bias towards normatively positive responses that may or may not coincide with researcher aims.

[2] As of September 2018, Berinsky et al. (2012) had over 2,000 citations on Google Scholar.

hypothesis advanced in the original study and a directional treatment scheme where different groups are provided with opposing expectations about the anticipated direction of the treatment effect. We conduct these experiments on convenience samples from Amazon's Mechanical Turk, where the potential for demand effects is thought to be particularly severe, as well as more representative samples from an online survey vendor. Across five surveys that involve more than 12,000 respondents and over 28,000 responses to these experiments, we fail to find evidence for the existence of EDEs in online survey experiments. That is, on average, providing respondents with information about the hypothesis being tested does not affect how they respond to the subsequent experimental stimuli.

To examine a most-likely case for EDEs, we also include conditions where respondents are given both information about experimenter intent *and* a financial incentive for responding in a manner consistent with researcher expectations. When this added incentive is present, we are sometimes able to detect differences in observed treatment effects that are consistent with the presence of EDEs. But on average, pooling across all our experiments, we still see no detectable differences in treatment effects even when financial incentives are offered.

While we cannot completely rule out the existence of EDEs, we show that conditions which should magnify their presence do not facilitate the confirmation of researcher hypotheses in a typical set of experimental designs. When made aware of the experiment's goal, respondents did not generally assist researchers. These results have important implications for the design, implementation, and interpretation of survey experiments. For one, they suggest that traditional survey experimental designs are robust to this long-standing concern. In addition, efforts to obfuscate the aim of experimental studies due to concerns about demand effects, including ethically questionable modes of deception, may be unnecessary in a variety of settings.[3]

## CONCERNS ABOUT EXPERIMENTER DEMAND EFFECTS

Orne (1962) raises a fundamental concern for the practice of experimental social science research: in an attempt to be "good subjects," participants draw on study recruitment materials, their interactions with researchers, and the materials included in the experiment to formulate a view of the behavior that researchers expect of them. They then attempt to validate a researcher's hypothesis by behaving in line with what they perceive as the expected behavior in a study. These "demand effects" represent a serious methodological concern with the potential to undercut supportive evidence from otherwise compelling research designs by offering an artifactual, theoretically uninteresting explanation for nearly any experimental finding (see also Bortolotti and Mameli 2006; Rosnow and Rosenthal 1997; Weber and Cook 1972; Zizzo 2010).

While rooted in social psychology laboratory studies that involve substantial researcher–subject interaction (e.g., Iyengar 2011), concerns about EDEs extend to other settings. In particular, demand effects also have the potential to influence experimental results in the substantial body of research employing survey experiments to study topics throughout social science. In what follows, we define survey experiments as studies in which research subjects self-administer a survey instrument containing both the relevant experimental treatments and outcome measures. This encompasses a broad class of studies in which participants recruited through online labor markets (Berinsky, Huber, and Lenz 2012), survey vendors (Mutz 2011), local advertisements (Kam, Wilking, and Zechmeister 2007), or undergraduate courses (Druckman and Kam 2011) receive and respond to experimental treatments in a survey context.

This focus serves two purposes. First, these scope conditions guide our theorizing about potential channels through which demand effects may or may not occur by limiting some avenues (e.g., cues from research assistants) credited with conveying demand characteristics to experimental participants in laboratory settings (Orne and Whitehouse 2000). Second, this definition encompasses a substantial body of social science research, making a focused assessment of EDEs relevant for the wide array of studies that employ this methodological approach (see Mutz 2011, Sniderman 2011, Gaines, Kuklinski, and Quirk 2007 for discussions of the growth of survey experiments in political science).

## EXPERIMENTER DEMAND EFFECTS IN SURVEY EXPERIMENTS

Concerns over EDEs are not limited to laboratory studies and are often explicitly invoked by researchers when discussing the design and interpretation of survey experiments. In a survey experiment evaluating how seeing Muslim women wearing hijabs affects attitudes related to representation, Butler and Tavits (2017) show politicians images of men and women in which either some or none of the women wear hijabs. The authors avoid a treatment in which all the women in the image wear hijabs, "because we wanted to mitigate the possibility of a demand effect" (728). Huber, Hill, and Lenz (2012) employ a multi-round behavioral game meant to assess how citizens evaluate politicians' performance and take care to address the concern that participants will come to believe their performance in later rounds counts more than in early rounds, thereby inducing "a type of demand effect" (727).

Countermeasures to combat the threat of EDEs in survey experiments stem from a shared assumption that demand effects can be limited by obfuscating an experimenter's intentions from participants. In one approach, researchers disguise experimental treatments

[3] While we offer evidence that these features may be unnecessary to combat demand effects, they sometimes serve additional purposes beyond alleviating EDEs (e.g., addressing concerns about social desirability bias) and may still be necessary due to these other concerns.

and primary outcome measures. Fowler and Margolis (2014, 103) embed information about the issue positions of political parties in a newspaper's "letter to the editor" section, rather than provide the information directly to respondents, to minimize the possibility that subjects realize the study's focus. Hainmueller, Hopkins, and Yamamoto (2014, 27) advocate the use of "conjoint" experiments, in which respondents typically choose between two alternatives (e.g., political candidates) comprised of several experimentally manipulated attributes, in part because the availability of multiple attributes conceals researcher intent from participants. Druckman and Leeper (2012, 879) examine the persistence of issue framing effects across a survey panel and only ask a key outcome measure in their final survey to counteract a hypothesized EDE in which participants would otherwise feel pressured to hold stable opinions over time.

In a second approach, researchers use cover stories to misdirect participants about experimenter intent (e.g., Bortolotti and Mameli 2006; Dickson 2011; McDermott 2002). Kam (2007, 349) disguises an experiment focused on implicit racial attitudes by telling participants the focus is on "people and places in the news" and asking questions unrelated to the experiment's primary goal. In studies of the effects of partisan cues, Bullock (2011, 499) and Arceneaux (2008, 144) conceal their focus by telling participants the studies examine the public's reaction to "news media in different states" and "how effectively the Internet provides information on current issues."

## POTENTIAL LIMITS ON EDEs IN SURVEY EXPERIMENTS

Concerns about EDEs in survey experiments are serious enough to influence aspects of experimental design. However, there is limited empirical evidence underlying these concerns in the survey experimental context. Recent studies have begun to assess the presence and severity of demand effects in some survey settings. White et al. (2018) test whether the characteristics of survey researchers—one potential source of experimenter demand in online settings—alter experimental results. They find that manipulating the race and gender of the researcher in a pretreatment consent script has no discernible effect on experimental results. de Quidt, Haushofer, and Roth (2018) probe for demand effects in experimental designs common in behavioral economics, including dictator and trust games. They conclude that EDEs are modest in these settings. Despite these new developments, there is still limited evidence for the presence or absence of demand effects in survey experiments with attitudinal outcomes—where respondents face fewer costs for expressive responding than in behavioral games with a monetary incentive—and in situations where experimenter intent is conveyed in a direct manner, rather than indirectly through the inferences respondents make based on researcher demographics.

There are distinctive aspects of survey experiments that cast some doubt on whether the EDE critique generalizes to this setting. One set of potential limitations concerns subjects' ability to infer experimenter intent in survey experiments. Even absent a cover story, survey experiments typically utilize between-subject designs that provide no information on the experimental cell in which participants have been placed. Treatments in these studies are also embedded inside a broader survey instrument, blurring the line between the experimental sections of the study and nonrandomized material that all respondents encounter.

These features create a complicated pathway for participants to infer experimenter intent. Respondents must not only parse the experimental and nonexperimental portions of the survey instrument but, having done so, they need to reason out the broader experimental design and determine the behavior that aligns with the experimenter's intentions, even as they only encounter a single cell in the broader experimental design. If errors occur in this process, even would-be "helpful" subjects will often behave in ways that fail to validate researcher expectations.

Of course, the process through which subjects respond to an experiment's demand characteristics may not be so heavily cognitive. The primary source of demand effects in laboratory experiments are subtle cues offered by researchers during their direct interactions with experimental participants (Rosnow and Rosenthal 1997, 83; see also; Orne and Whitehouse 2000). However, the context in which many survey experiments are conducted blocks this less cognitively taxing path for demand effects to occur. Online survey experiments fit into a class of "automated" experiments featuring depersonalized interactions between researchers and subjects. Theories about the prevalence of demand effects in experimental research consider automated experiments to be a least-likely case for the presence of EDEs (Rosenthal 1976, 374–375; Rosnow and Rosenthal 1997, 83). In line with these accounts, online experiments were considered a substantial asset for reducing the presence of EDEs at the outset of this type of research (Piper 1998; McDermott 2002, 34; Siah 2005, 122–3).

A second set of potential limitations is that, even if participants correctly infer experimenter intent and interpret the complexities of the survey instrument, they may not be inclined to assist researchers. While EDEs rely on the presence of "good subjects," other scholars raise the possibility of "negativistic subjects" who behave contrary to what they perceive to be researcher intentions (Cook et al. 1970; Weber and Cook 1972) or participants who are simply indifferent to researcher expectations (Frank 1998). To the extent these other groups exhibit the on-average inclination of a subject pool, they would defy researcher expectations. While there is limited empirical evidence on the distribution of these groups in various subject pools, prior studies offer suggestive evidence that fails to align with the "good subject" perspective. Comparing findings between experienced experimental participants drawn from online subject pools (who are potentially better at discerning experimenter intentions), and more naive participants, researchers find that treatment effects are smaller among the more experienced subjects (Chandler, Mueller, and Paolacci 2014; Chandler et al. 2015;

Krupnikov and Levine 2014). At least for the online samples now common in survey experimental research, this is more in line with a negativistic, or at least indifferent, portrayal of experimental subjects than accounts where they attempt to validate researcher hypotheses.

Despite the widespread concern over EDEs in online survey experiments, our discussion highlights several elements that may limit demand effects in these studies. However, there is limited evidence to test between this account and other perspectives in which EDEs create widespread problems for survey experiments in political science. For this reason, the next section introduces a research design to empirically examine demand effects in political science survey experiments.

## RESEARCH DESIGN

We deploy a series of experiments specifically designed to assess the existence and magnitude of EDEs. We do so by replicating results from well-known experimental designs while also randomizing the degree to which the purpose of the experiment is revealed to participants. Our data come from five surveys fielded on two survey platforms (see Table 1). The first three surveys were conducted on Amazon's Mechanical Turk, which hosts an experienced pool of survey respondents (see, e.g., Berinsky, Huber, and Lenz 2012; Hitlin 2016). The last two samples were purchased from the survey vendor Qualtrics, the second of which was quota sampled to meet nationally representative targets for age, race, and gender. In cases where more than one experiment was embedded within a single survey instrument, all respondents participated in each experiment, though the participation order was randomized.

While the convenience sample of respondents from Mechanical Turk used in the first three studies may present disadvantages for many types of research, we view it as an ideal data source in this context. Prior research portrays Mechanical Turk as a particularly likely case for demand effects to occur based on the labor market setting in which subjects are recruited (e.g., Berinsky, Huber, and Lenz 2012; Paolacci and Chandler 2014). These platforms host experienced survey participants that are especially attentive to researcher expectations due to the concern that they will not be compensated for low-quality work (i.e., the requester may not approve their submission) and their need to maintain a high work approval rate to remain eligible for studies that screen on past approval rates. This attentiveness creates the possibility that, in an attempt to please researchers, respondents will react to any features of an experiment that reveal the response expected of them by the researcher. If we fail to observe EDEs using these samples, we may be unlikely to observe them in other contexts. However, in order to speak to the threat of EDEs in higher-quality respondent pools, we present results from Qualtrics samples as well. In what follows, we first outline the published studies we chose to replicate. We then describe three different experimental schemes that were employed to test for the presence and severity of demand effects.

## REPLICATED STUDIES

To test for the presence and severity of EDEs, we replicate five published studies. Two studies come from the American Politics literature. The first is a classic framing study, a substantive area where concerns over demand effects have been expressed in laboratory contexts (e.g., Page 1970; Sherman 1967). In this experiment, respondents read a hypothetical news article about a white supremacist group attempting to hold a rally in a US city (Mullinix et al. 2015; Nelson, Clawson, and Oxley 1997). In the control condition, respondents saw an article describing the group's request to hold the rally. In the treatment condition, respondents saw a version of the article highlighting the group's first amendment right to hold the rally. Both groups were then asked how willing they would be to allow the rally. The hypothesis, based on prior findings, was that those exposed to the free speech frame would be more likely to support the group's right to hold the rally.

The second experiment was inspired by Iyengar and Hahn (2009), which tests whether partisans are more likely to read a news article if it is offered by a news source with a reputation for favoring their political party (i.e., partisan selective exposure). We offered participants two news items displayed in a $2 \times 2$ table (see Figure A.2 in the Online Appendix), each with randomized headlines and sources, and asked them to state a preference for one or the other. The sources were

**TABLE 1. Survey Samples**

| Survey | Platform | Date | *N* | Sample | Included experiments | Demand treatment scheme |
|---|---|---|---|---|---|---|
| 1 | M-Turk | Jan. 2017 | 1,395 | Convenience | Framing, partisan news | Gradation |
| 2 | M-Turk | Mar. 2017 | 1,635 | Convenience | Resumé, partisan news | Directional |
| 3 | M-Turk | Jan. 2018 | 1,874 | Convenience | Democratic peace, welfare | Incentive |
| 4 | Qualtrics | Feb. 2018 | 2,374 | Only partisans | Partisan news | Incentive |
| 5 | Qualtrics | Apr. 2018 | 5,550 | Nationally rep. quotas for race, age, gender | Framing, democratic peace, welfare | Incentive |

Fox News (the pro-Republican option), MSNBC (the pro-Democrat option), and USA Today [the neutral option (Mummolo 2016)]. Responses were analyzed in a conjoint framework (Hainmueller, Hopkins, and Yamamoto 2014), in which each of the two news items offered to each respondent was treated as a separate observation.[4] The inclusion of a conjoint design—especially one with so few manipulated attributes[5]—offers another avenue for EDEs to surface, as within-subject designs are thought to contain, "the potential danger of a purely cognitive EDE if subjects can glean information about the experimenter's objectives from the sequence of tasks at hand," but may offer increased statistical power relative to between-subject experiments (Zizzo 2010, 84; see also Charness, Gneezy, and Kuhn 2012 and Sawyer 1975).

We replicate one study from International Relations, a highly cited survey experiment by Tomz and Weeks (2013) examining the role of public opinion in the maintenance of the "Democratic Peace"—the tendency of democratic nations not to wage war on one another. In this experiment, respondents assessed a hypothetical scenario in which the United States considers whether to use force against a nation developing nuclear weapons. The experiment supplied respondents with a list of attributes about the unnamed country in question, one of which is whether the country is a democracy (randomly assigned). The outcome is support for the use of force against the unnamed country.

We replicate one study from Comparative Politics concerning attitudes toward social welfare (Aarøe and Petersen 2014). In this experiment, respondents are presented with a hypothetical welfare recipient who is described as either unlucky ("He has always had a regular job, but has now been the victim of a work-related injury.") or lazy ("He has never had a regular job, but he is fit and healthy. He is not motivated to get a job."). Following this description, we measure support for restricting access to social welfare.

Finally, we replicate a resumé experiment (Bertrand and Mullainathan 2004) in which a job applicant is randomly assigned a stereotypically white or African American name. We hold all other attributes of the resumé constant and ask respondents how willing they would be to call the job applicant for a job interview. Our expectation, based on prior results, was that respondents who saw the job candidate with the stereotypically African American name would be less likely to say they would call the candidate for an interview.

In general, we are able to recover treatment effects in our replications that are highly similar in both direction and magnitude to the previous studies they are based on (see Figures B.3–B.7 in Online Appendix). The one exception is the resumé experiment, where we do not observe evidence of anti-Black bias. We suspect this difference stems from varying context. The original study was a field experiment conducted on actual employers relative to the survey experiment on a convenience sample that is used here [though a recent labor market field experiment (Deming et al. 2016), also failed to find consistent race effects]. Nevertheless, we include results from the resumé experiment below because our interest is primarily in how revealing an experiment's hypothesis affects respondent behavior, not the effect of the treatment in the original study.

## MANIPULATING THE THREAT OF DEMAND EFFECTS

Using these five experimental designs, we manipulate the risk of EDEs with three approaches, all of which involve providing respondents varying degrees of information about experimenter intentions prior to participating in one of the experiments described above. The presence or absence of this additional information was randomized independently of the treatments within each experiment. In the first approach, which we term the "Gradation" scheme, we randomly assign respondents to receive either no additional information, a hint about the researcher's hypothesis, or an explicit description of the researcher's hypothesis.

We next employ a "Directional" scheme that manipulates the anticipated direction of the expected treatment effect, assigning respondents to receive either no additional information, an explicit hypothesis stating the treatment will induce a positive shift in the outcome, or an explicit hypothesis stating the treatment will induce a negative shift in the outcome. (To make the expectations presented in the "Directional" scheme plausible, we also add a brief justification for why we hypothesize the given effect.) This directional design eliminates the possibility that we will fail to observe EDEs simply because respondents are predisposed to behave in line with researcher expectations even in the absence of knowledge of a study's hypothesis. For example, if no EDEs occur in the gradation version of the partisan news experiment, it may be because respondents were already inclined to respond positively to the politically friendly news source, making demand behavior and sincere responses observationally equivalent. The directional design breaks this observational equivalence.

Finally, we use an "Incentive" scheme that offers a financial incentive for assisting the researcher in confirming their hypothesis—an approach that maximizes the likelihood of observing EDEs, helps us adjudicate between different mechanisms that may produce or inhibit EDEs, and sheds light on the external validity of our findings (we discuss this design and its implications in greater detail below). Table 2 displays the wording of the first two EDE treatment schemes in the context of the partisan news experiment (see Online Appendix A for wording used in the other experiments).

[4] Headlines and sources were randomly drawn without replacement from lists of three total possible headlines and sources, meaning the two competing news items always contained different content. Figure B.2 in the Online Appendix displays the results of tests for balance on observables for all experiments.

[5] We note that several typical conjoint experimental designs, such as candidate choice experiments, contain more than two attributes per profile, which may serve to mask researcher intent and help guard against EDEs.

**TABLE 2. Text of EDE Treatments in Partisan News Experiment**

| Gradation scheme | | Directional scheme | |
|---|---|---|---|
| Control: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read." | Control: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read." |
| Hint: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read. The purpose of this exercise is so we can measure whether the news outlet offering an article influences how likely people are to read the article." | Positive effect: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read. The purpose of this exercise is so we can measure whether the news outlet offering an article influences how likely people are to read the article. We expect that people will be more likely to choose an article if the news source offering it is known to favor their preferred political party, since people tend to seek out information that is consistent with their personal views." |
| Explicit: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read. The purpose of this exercise is so we can measure whether people are more likely to choose a news item if it is offered by a news outlet with a reputation of being friendly toward their preferred political party." | Negative effect: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read. The purpose of this exercise is so we can measure whether the news outlet offering an article influences how likely people are to read the article. We expect that people will be more likely to choose an article if the news source offering it is known to be more critical of their preferred political party, since people often say they strive to be open minded and are willing to hear diverse points of view." |

The quantity of interest in all these experiments is a difference-in-differences. Specifically, we estimate the difference in an experiment's treatment effect due to revealing information about its purpose to participants. This quantity is represented by the following expression:

$$(E[\text{response}|\text{information, treatment}] - E[\text{response}|\text{information, control}]) - (E[\text{response}|\text{no information, treatment}] - E[\text{response}|\text{no information, control}]).$$

This estimand captures the degree to which demand effects, if present, are consequential for the conclusions produced by survey experimental research. If the traditional EDE critique is valid, offering this information should lead participants to assist in the confirmation of each hypothesis and make treatment effects in the presence of additional information about the experiment's aim larger (in absolute value) than in the absence of such information. This quantity focuses attention on the key source of concern regarding demand effects: Does their presence alter the treatment effects researchers obtain from survey experiments?

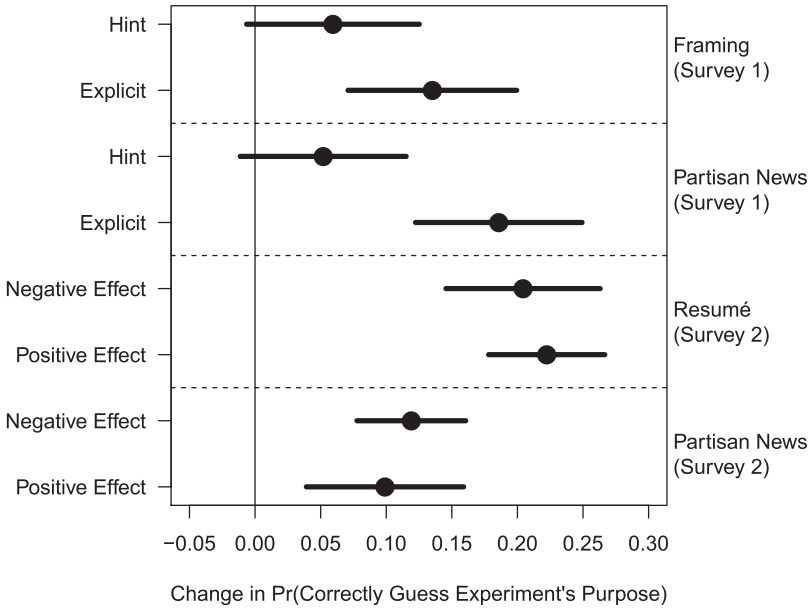
## RESULTS

A first-order concern is verifying that respondents grasped the information the demand treatments revealed about the purpose of the experiments. As a manipulation check, we measure respondent knowledge of the purpose of each experiment by asking them to choose from a menu of six or seven (depending on the experiment) possible hypotheses following each experiment.[6] Across all the studies, the mean rate of correctly guessing the hypothesis among those provided no additional information was 33%. This suggests that the actual hypotheses were not prohibitively obvious, and that it should be possible to manipulate the risk of EDEs by revealing additional information.

Figure 1 displays the results of OLS regressions of indicators for guessing the purpose of the experiment on indicators for the EDE treatment conditions. Turning first to the "Gradation" treatment scheme in the framing experiment, those in the hint and explicit conditions were six- and 14-percentage-points more likely to correctly guess the researcher's hypotheses relative to those who were given no information on the

[6] See Figures A.4–A.11 in the Online Appendix for the wording of these items and Figure B.1 for mean rates of correct guessing across experiments.

**FIGURE 1. Manipulation Check: Information Treatments Increase Risk of EDEs**




*Note*: The figure displays the effects of revealing information about an experiment's hypothesis on the probability of correctly guessing the experiment's hypothesis from a multiple choice list later in the survey. The results indicate the treatments were effective at increasing the share of respondents aware of the experiment's hypothesis, thereby increasing the theoretical risk of EDEs. Bars represent 95% confidence intervals.

experiment's purpose. We see similar results in the partisan news experiment. Compared to the baseline condition with no demand information, those in the hint and explicit conditions were five- and 19-percentage-points more likely to correctly guess the hypothesis. Even among this M-Turk sample comprised of respondents thought to be particularly attentive, an explicit statement of experimenter intent is necessary to produce large increases in awareness of a study's focus.

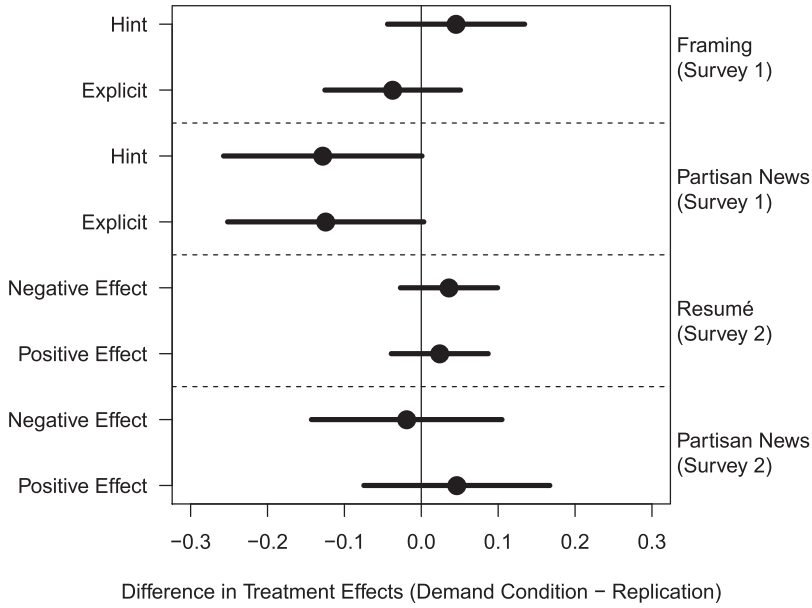
The manipulations also worked as intended in the "Directional" EDE experiments. In this case, respondents in the information conditions were informed we either hypothesized a positive or negative effect, so we define a "correct" guess as a respondent accepting whatever directional justification was offered in the treatment they received. For respondents in the news experiment, for example, this means individuals in the "positive" treatment condition were correct if they guessed the purpose was to show a preference for news from co-partisan sources and individuals in the "negative" treatment condition were correct if they guessed the expectation was to show a preference for news from out-party sources. In these experiments, additional information induced between 10- and 22-percentage-point increases in the probability of guessing the experiment's purpose later in the survey. This means the additional information successfully changed what participants understood as the purpose of the experiments, moving respondent perceptions in the "positive" conditions in a different direction than their counterparts in the "negative" conditions. Relative to the unidirectional treatments in Survey 1, this alternative scheme drives a wider wedge between the perceptions of the two information conditions, amplifying the potential risk for EDEs to alter the treatment effects estimated in these groups relative to the uninformed control group.

While the increases in the rates of correctly guessing a study's hypothesis are detectable, there remain sizable shares of respondents who fail to infer the hypothesis even when it is explicitly stated to them. This suggests that many survey respondents are simply inattentive—one mechanism that may diminish the threat of EDEs. If survey respondents are not engaged enough to recall information presented to them minutes earlier, it is unlikely they will successfully navigate the complex task of inducing EDEs.

We now turn to our key test, which measures the differences in treatment effects between conditions where respondents were given additional information about an experiment's hypothesis, and conditions where they were given no additional information. We again note here that, aside from the previously discussed issues with the resumé experiment included in Survey 2, the treatment effects in the baseline conditions closely align with the effects in the prior studies they replicate (see Online Appendix B). This means these tests examine deviations from familiar baseline estimates of the treatment effects due to the introduction of information about researcher expectations.

This first set of tests uses two samples from Amazon's Mechanical Turk. Figure 2 displays the results of the "Gradation" and "Directional" treatments across the framing, partisan news and resumé experiments. We find no evidence that any of the demand treatments

**FIGURE 2. No Evidence of EDEs When Revealing Hypothesis**




*Note*: The figure displays the differences in treatment effects (difference-in-differences) between conditions where respondents were given information on a researcher's hypothesis, and the control condition in which no additional information was given. Positive estimates indicate changes in the treatment effect in the direction of the stated hypothesis. The results show no evidence that knowledge of the researcher's expectations causes respondents to help confirm a hypothesis. Bars represent 95% confidence intervals.

changed the substantive treatment effects of primary interest in these studies. In general, these treatment effects are statistically indistinguishable from the ones we observe in the control condition (i.e., the effects produced by replicating the published studies without supplying any additional information). The only borderline statistically significant results come from the first partisan news experiment, where revealing the hypothesis made respondents *less* likely to respond in ways that would confirm it. However, this attenuation was not replicated in the second partisan news experiment, raising doubts about the robustness of this finding. Overall, we find no support for the key prediction of the demand effects hypothesis. Although we successfully moved respondent perceptions of the purpose of each experiment, revealing this information did not help to confirm the stated hypotheses.

## ARE SURVEY RESPONDENTS CAPABLE OF INDUCING DEMAND EFFECTS?

Finding no evidence for demand effects in the initial surveys, we conducted additional surveys designed to parse the mechanism behind these null effects by maximizing the risk of EDEs. As theorized above, there are at least two plausible reasons why EDEs may fail to materialize even when respondents are armed with information about an experimenter's hypothesis. First, respondents may be unable, perhaps due to cognitive limitations, to respond in ways that produce EDEs. Alternatively, respondents may be capable of inducing demand effects but simply *not inclined* to do so, as in portrayals of indifferent or negativistic research participants.

To arbitrate between these mechanisms, we implement a third EDE treatment scheme in which respondents encountered no information, an explicit statement of the hypothesis, or an explicit statement paired with the offer of a bonus payment if respondents answered questions in a way that would support the stated hypothesis. Table 3 displays the text of these treatment conditions in the partisan news experiment (see Tables A.1–A.3 in Online Appendix for wording in other experiments). These bonuses were for $0.25. Amounts of similar magnitude have proven sufficient to alter respondent behavior in other contexts. For instance, Bullock et al. (2015, 539) find that the opportunity for bonuses of this scale reduced the size of partisan gaps in factual beliefs by 50%.

If we make the reasonable assumption that survey respondents would rather earn more money for their time than intentionally defy a request from a researcher, offering additional financial incentives for exhibiting demand effects can shed light on the mechanism behind the lack of EDEs in the first two surveys. If EDEs fail to occur even when an additional financial reward is offered, we can infer that inability likely precludes demand effects. If, on the other hand, financial incentives produce EDEs, we can infer that the previous null results were likely due to a lack of desire from respondents to engage in demand-like behavior.

Determining whether respondents are unwilling, or simply unable, to produce EDEs helps inform the

**TABLE 3. Text of Incentive Scheme Treatments in Partisan News Experiment**

| | Incentive scheme |
|---|---|
| Control: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read." |
| Explicit: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read. The researchers conducting this survey expect that individuals are more likely to choose a news story if it is offered by a news outlet with a reputation of being friendly towards their preferred political party." |
| Explicit + incentive: | "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read. The researchers conducting this survey expect that individuals are more likely to choose a news story if it is offered by a news outlet with a reputation of being friendly towards their preferred political party. If your responses support this theory, you will receive a $0.25 bonus payment!" |

external validity of this study. The experiments replicated here are likely candidates for EDEs as they employ fairly straightforward designs, with only one treatment and one control condition, and make minimal effort to disguise researcher intent (i.e., no deception). If we determine that respondents are unable to produce EDEs even in this environment, it is likely that more complex experimental designs not replicated here are even more robust to EDEs.

To test this, we again conduct the framing and partisan news experiments, and also replicate two additional experiments: Tomz and Weeks (2013)—a study of democratic peace theory—and Aarøe and Petersen (2014) which hypothesizes that support for social welfare programs will be greater when welfare recipients are described as unlucky rather than lazy. In all experiments, respondents are either told nothing about the hypotheses, told the hypotheses explicitly, or told the hypotheses explicitly and offered a bonus payment for responding in accordance with these expectations.

Before discussing the main results, we again reference manipulation checks. Figure 3 displays changes in the probability of correctly guessing each experiment's hypothesis relative to the control condition where no information on the hypothesis was provided. As the figure shows, the information treatments again increased the share of respondents aware of each hypothesis, though the effects are much larger in the M-Turk samples than in the Qualtrics samples, a point to which we will return below.

Figure 4 displays the main results of our incentive-based EDE interventions. Once again there is no evidence of demand effects when respondents are explicitly informed of the hypothesis. However, when a bonus payment is offered to the M-Turk samples, the treatment effects increase in the expected direction. In the democratic peace experiment, the effect of describing the hypothetical nation as a democracy increases by 14 percentage points relative to the control condition that did not supply information on the hypothesis. Similarly, in the welfare study the financial incentives induce a borderline statistically significant five-percentage-point increase in the treatment effect compared to the effect in the control condition that received no information about experimenter intent.[7]

However, even with financial incentives, we find no evidence of EDEs in the Qualtrics samples. Since the two survey platforms engage participants that vary on many unobserved dimensions, it is difficult to pinpoint the reasons for these divergent results. However, the manipulation checks in the Qualtrics studies, displayed in Figure 3, suggest that respondents in this more representative pool are less attentive than M-Turkers. This pattern is in line with the intuition in Berinsky, Huber, and Lenz (2012), which warns that the risk of EDEs may be especially pronounced among the experienced survey takers on the M-Turk labor market. The inattentiveness indicated by the small shares of respondents that could be prompted to correctly guess the experiment's hypothesis even when additional financial incentives are offered again highlights an obstacle to EDEs, and also suggests treatment effects recovered in survey experiments are more akin to intention-to-treat effects (ITTs) than average treatment effects (ATEs), since many respondents assigned to treatment remained, in effect, untreated.

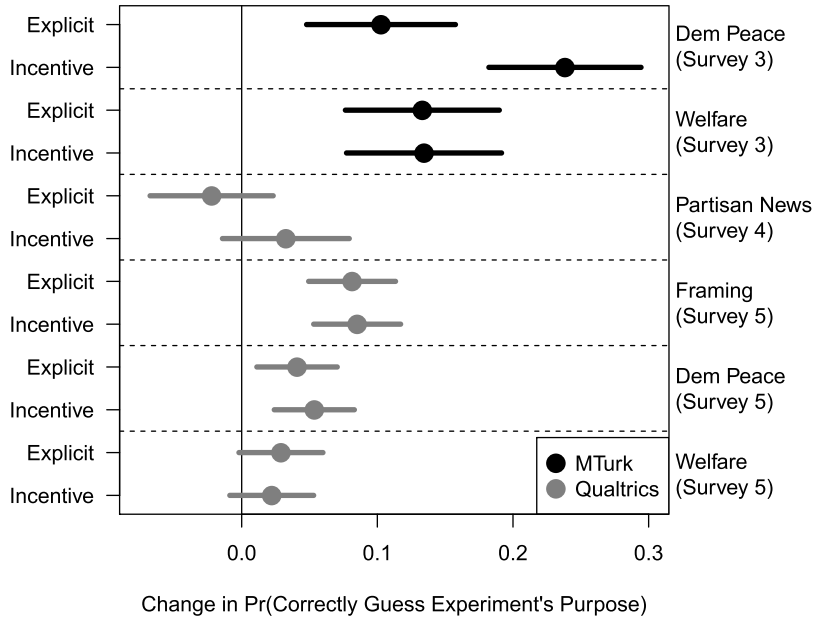
While these additional incentive conditions demonstrate modest evidence that M-Turkers are capable of inducing EDEs in the unusual case where they are offered extra money for doing so, they also show no evidence of EDEs among M-Turkers in the typical scenario where no such incentive is offered. In typical survey experimental settings, we again fail to recover evidence of the presence of EDEs.

## ARE EDEs PRESENT AT BASELINE?

The previous results demonstrate what happens to treatment effects in survey experiments when conditions that are theoretically conducive to EDEs are exacerbated. Contrary to common characterizations of demand effects, we find that providing survey respondents

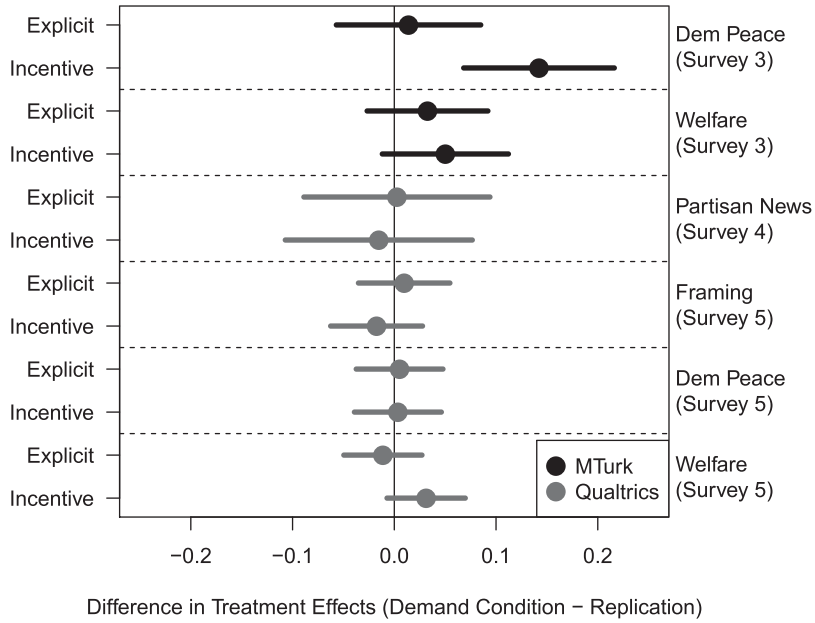
[7] For results pooling across all studies that featured additional incentives see Table B.6.

**FIGURE 3. Manipulation Check: Information and Incentive Treatments Increase Risk of EDEs**




*Note*: The figure displays the effects of revealing information on an experiment's hypothesis on the probability of correctly guessing the experiment's hypothesis from a multiple choice list later in the survey. The results indicate the treatments were effective at increasing the share of respondents aware of the experiment's hypothesis, thereby increasing the theoretical risk of EDEs. Respondents in the Qualtrics samples appear less attentive. Bars represent 95% confidence intervals.

**FIGURE 4. Financial Incentives Can Sometimes Induce EDEs**




*Note*: The figure displays the differences in treatment effects (difference-in-differences) between conditions where respondents were given either information on a researcher's hypothesis, a financial incentive for inducing an EDE—or both—and the control condition in which no additional information was given. Bars represent 95% confidence intervals.

information on the purpose of an experiment generally has no impact on the estimated treatment effects. Even the presence of additional monetary incentives has an inconsistent effect on these results. Still, this evidence cannot rule out EDEs completely. The reason is that some respondents may have inferred the purpose of the

experiment even without the additional information. If they then reacted differently due to this knowledge, it is possible that, even in the control condition where no extra information is provided, treatment effect estimates are inflated by the presence of "clever" respondents acting in line with researcher expectations. The directional manipulations included in Study 2 help in this regard as they move respondent's perceived expectations away from the most prevalent expectations offered by prior studies in those research areas. This section includes an additional test.

To evaluate this possibility, we leverage respondents' participation in multiple experiments in surveys 1–3 and 5 (see Table 1). In these surveys, we identify the respondents most likely to have inferred the experiments' hypotheses on their own: those who correctly guessed the hypothesis of the first experiment they encountered. Conversely, we label respondents as not likely to infer hypotheses on their own if they were unable to correctly guess the hypothesis of the first experiment they encountered.[8] We then compare the treatment effects estimated for these two groups of respondents in the *second* experiment they encountered in the survey. If "clever" respondents inflate treatment effects due to demand-like behavior, we should observe larger effects among them compared to respondents who are less likely to infer an experiment's purpose.[9]

Table 4 displays the results of models comparing treatment effects among those who did and did not correctly guess the first experiment's purpose pooled across all surveys in which multiple experiments were included. The first column is generated using only the sample of respondents who did not receive information about the hypothesis of the first experiment they encountered (those in the baseline, "no information" condition). The second column is generated from data on all respondents who correctly guessed the hypothesis in their first experiment, whether they received additional information or not.[10] In both sets of results, we find no evidence that "clever" respondents exhibit differentially large treatment effects. While the interaction terms in these models—which represent the difference in treatment effects between "clever" respondents and their counterparts—are positive, neither is statistically distinguishable from zero. We reach the same conclusions when breaking out the experiments one by one rather than pooling all the data, but in those cases we suspect our tests are severely underpowered.

**TABLE 4. Pooled Estimates of Treatment Effect Variation by Correct Guess of Hypothesis in Previous Experiment**

| | No demand information | All conditions |
|---|---|---|
| (Intercept) | 0.53* | 0.55* |
| | (0.03) | (0.03) |
| Treatment | 0.18* | 0.19* |
| | (0.06) | (0.06) |
| Correct guess | −0.04 | −0.01 |
| | (0.02) | (0.03) |
| Treatment × Correct guess | 0.06 | 0.01 |
| | (0.04) | (0.04) |
| *N* | 1,232 | 3,750 |

Models include study fixed effects, continuous outcomes rescaled between 0 and 1.
Robust standard errors, clustered by study, in parentheses
*Indicates significance at $p < 0.05$.

Some might wonder whether the positive point estimate on the interaction term in Table 4, Column 1 indicates the presence of EDEs, the large standard error notwithstanding. Suppose we take this estimate of 0.06 (six-percentage-points) to be true, and make the further conservative assumption that this entire effect is due to EDEs, and not due to other sources of differential response between correct and incorrect guessers. Given that correct guessers make up roughly 38% of the sample used to estimate Column 1 in Table 4, this means that we would expect EDEs to inflate an estimated average treatment effect by roughly two- to three-percentage-points, from about 0.18 to 0.21.

How often would this degree of bias alter the inference in a typical political science survey experiment? To gauge this, we reproduced an analysis from Mullinix et al. (2015), which replicated 20 survey experimental designs that received funding through Time Sharing Experiments for the Social Sciences (TESS) on both convenience samples from Amazon's Mechanical Turk and nationally representative samples from Knowledge Networks. These experiments, "address diverse phenomena such as perceptions of mortgage foreclosures, how policy venue impacts public opinion, and how the presentation of school accountability data impacts public satisfaction…" (Mullinix et al. 2015, 118). We transformed all 40 treatment effects which appeared in Figure 2 in Mullinix et al. (2015) into absolute value percentage-point shifts on each study's outcome scale. We then diluted each treatment effect toward zero by three percentage points to mimic the largest EDE our

[8] While those who correctly guessed the hypothesis of the first experiment they encountered without outside information are theoretically the most-likely group to contribute to EDEs, we note that even these respondents faced challenges in diagnosing experimenter intent across multiple studies, a necessary condition for engaging in demand-like behavior. Among those who received no information regarding the hypotheses of either experiment, the correlations between correctly guessing experimenter intent across the six potential experiment pairs in our studies were: −0.02, 0.07, 0.07, 0.08, 0.14, and 0.36. Though mostly positive, these correlations are modest. We take this as further evidence of the obstacles that limit survey respondents' ability to induce EDEs.

[9] Note that by testing for heterogeneity in the effect of a treatment that was randomly assigned after respondents did or did not guess the correct hypothesis in a previous experiment, we avoid the threat of post-treatment bias that would be present if we compared treatment effects between correct and incorrect guessers within an experiment (Angrist and Pischke 2009; Rosenbaum 1984).

[10] In the "All Conditions" analysis, respondents in a directional treatment condition in Study 2 are coded as correctly guessing experimenter intent if they went in the direction consistent with the information provided by the information treatment. Study 2 respondents that did not receive additional information are coded as correct if they chose the hypothesis consistent with the original study in all cases.

paper suggests is likely to be realized (see Appendix Figure B.8 for results). Doing so changed the sign of four out of 40 effects, though all of those results were not statistically significant to begin with, so there would be no change in inference. Two additional effects lost statistical significance using two-standard-error confidence intervals, and the vast bulk of substantive conclusions remained unchanged. Taken together, Table 4, Column 1, under the most conservative assumptions, suggests some risk of EDEs among a subset of respondents, but the effects are not large enough to refute our general claim that EDEs are unlikely to meaningfully bias a survey experimental result except in studies attempting to detect very small treatment effects.

## DISCUSSION AND CONCLUSION

Survey experiments have become a main staple of behavioral research across the social sciences, a trend aided by the increased availability of inexpensive online participant pools. With the expansion of this type of study, scholars have rightly identified a set of concerns related to the validity of survey experimental results. One common concern is that survey respondents—especially ones who frequently take part in social science experiments—have both the means and the incentives to provide responses that artificially confirm a researcher's hypothesis and deviate from their sincere response to an experimental setting. In this study, we provide some of the first empirical evidence regarding the existence and severity of this theoretical vulnerability.

Our results consistently defy the expectations set out by the EDE critique. Rather than assisting researchers in confirming their hypotheses, we find that revealing the purpose of experiments to survey respondents leads to highly similar treatment effects relative to those generated when the purpose of the experiment is not provided. We also provide evidence as to the mechanism that produces these null results. By offering additional financial incentives to survey participants for responding in a way that confirms the stated hypotheses, we show that, with rare exceptions, respondents appear largely unable to engage in demand-like behavior. This suggests that in typical research settings, where such incentives are unavailable, respondents are unlikely to aid researchers in confirming their hypotheses.

These results have important implications for the design and interpretation of survey experiments. While there may be other reasons to obfuscate a study's purpose or misdirect respondents, such as fostering engagement[11] or avoiding social desirability bias, our evidence suggests that the substantial effort and resources researchers expend obfuscating hypotheses to prevent demand-like behavior may be misguided. These tactics include ethically questionable attempts to deceive participants in order to preserve the scientific validity of results. Even in the event that a hypothesis is explicitly stated to the participant, there appears to be little risk it will inflate the observed treatment effects.

In light of our findings, there are several additional questions worthy of pursuit. There may be substantial variation in how respondents react to knowledge of an experiment's hypothesis across substantive areas. Though we have attempted to test for the presence of EDEs across a range of topics by covering all empirical subfields in political science, it remains possible that certain topics may be especially vulnerable to EDEs. There may also be heterogeneity among respondents. Subject pools with varying levels of experience participating in survey experiments may respond differently to the stimuli examined here.

In spite of these limitations, our consistent inability to uncover evidence of hypothesis-confirming behavior across multiple samples, survey platforms, research questions and experimental designs suggests that long-standing concerns over demand effects in survey experiments may be largely exaggerated. In general, knowledge of a researcher's expectations does not alter the behavior of survey participants.

## SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055418000837.

Replication materials can be found on Dataverse at: https://doi.org/10.7910/DVN/HUKSID.

## REFERENCES


Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton: Princeton University Press.

Arceneaux, Kevin. 2008. "Can Partisan Cues Diminish Democratic Accountability?" *Political Behavior* 30 (2): 139–60.

Aarøe, Lene, and Michael Bang Petersen. 2014. "Crowding Out Culture: Scandinavians and Americans Agree on Social Welfare in the Face of Deservingness Cues." *The Journal of Politics* 76 (3): 684–97.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.Com's Mechanical Turk." *Political Analysis* 20 (3): 351–68.

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.

Bortolotti, Lisa, and Matteo Mameli. 2006. "Deception in Psychology: Moral Costs and Benefits of Unsought Self-Knowledge." *Accountability in Research* 13: 259–75.

Bullock, John G. 2011. "Elite Influence on Public Opinion in an Informed Electorate." *American Political Science Review* 105 (3): 496–515.

Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10 (4): 519–78.


[11] For example, McConnell et al. (2018) embeds partisan cues in a paid copyediting task and told respondents that, "the text was from the website of a new software company that we (the employers) hoped to launch soon"(8). Leading respondents to believe they were being paid for actual work was needed to obtain valid measures of some dependent variables, including the number of errors respondents corrected.

Butler, Daniel M., and Margrit Tavits. 2017. "Does the Hijab Increase Representatives' Perceptions of Social Distance?" *The Journal of Politics* 79 (2): 727–31.
Chandler, Jesse, Pam Mueller, and Gabrele Paolacci. 2014. "Nonnaivete Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavior Research Methods* 46 (1): 112–30.
Chandler, Jesse, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate A. Ratfliff. 2015. "Using Nonnaive Participants Can Reduce Effect Sizes." *Psychological Science* 26 (7): 1131–9.
Charness, Gary, Uri Gneezy, and Michael A. Kuhn. 2012. "Experimental Methods: Between-Subject and Within-Subject Design." *Journal of Economic Behavior & Organization* 81 (1): 1–8.
Cook, Thomas D., James R. Bean, Bobby J. Calder, Robert Frey, Martin L. Krovetz, and Stephen R. Resiman. 1970. "Demand Characteristics and Three Conceptions of the Frequently Deceived Subject." *Journal of Personality and Social Psychology* 14 (3): 185–94.
Deming, David J., Noam Yuchtman, Amira Abulafi, Claudia Goldin, and Lawrence F. Katz. 2016. "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study." *American Economic Review* 106 (3): 778–806.
de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* 108 (11): 3266–302.
Dickson, Eric S. 2011. "Economics versus Psychology Experiments." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, Jame H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press, 58–69.
Druckman, James N., and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and its Effects." *American Journal of Political Science* 56 (4): 875–96.
Druckman, James N., and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, Jame H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press, 41–57.
Fowler, Anthony, and Michele Margolis. 2014. "The Political Consequences of Uninformed Voters." *Electoral Studies* 34: 100–10.
Frank, B. L. 1998. "Good News for the Experimenters: Subjects Do Not Care about Your Welfare." *Economics Letters* 61: 171–4.
Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15 (1): 1–20.
Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26 (3): 213–24.
Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1): 1–30.
Hitlin, Paul. 2016. *Research in the Crowdsourcing Age, a Case Study*: Pew Research Center Report. http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/.
Huber, Gregory A., Seth J. Hill, and Gabriel S. Lenz. 2012. "Sources of Bias in Retrospective Decision Making: Experimental Evidence on Voters' Limitations in Controlling Incumbents." *American Political Science Review* 106 (4): 720–41.
Iyengar, Shanto. 2011. "Laboratory Experiments in Political Science." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupa. New York: Cambridge University Press, 73–88.
Iyengar, Shanto, and Kyu S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59 (1): 19–39.
Kam, Cindy D. 2007. "Implicit Attitudes, Explicit Choices: When Subliminal Priming Predicts Candidate Preferences." *Political Behavior* 29: 343–67.
Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research." *Political Behavior* 29 (4): 415–40.
Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1 (1): 59–80.
McConnell, Christopher, Yotam Margalit, Neil Malhotra, and Matthew Levendusky. 2018. "The Economic Consequences of Partisanship in a Polarized Era." *American Journal of Political Science* 62 (1): 5–18.
McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science* 5: 31–61.
Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2 (2): 109–38.
Mummolo, Jonathan. 2016. "News from the Other Side: How Topic Relevance Limits the Prevalence of Partisan Selective Exposure." *The Journal of Politics* 78 (3): 763–73.
Mutz, Diana C. 2011. *Population-Based Survey Experiments*: Princeton, NJ: Princeton University Press.
Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and its Effect on Tolerance." *American Political Science Review* 91 (3): 567–83.
Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist* 17 (11): 776–83.
Orne, Martin T., and Wayne G. Whitehouse. 2000. "Demand Characteristics." In *Encyclopedia of Psychology*, ed. Alan E. Kazdin. Washington, D.C.: American Psychological Association and Oxford Press, 469–70.
Page, Monte M. 1970. "Role of Demand Awareness in the Communicator Credibility Effect." *Journal of Social Psychology* 82: 57–66.
Paolacci, Gabriele, and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23 (3): 184–8.
Piper, Allison I. 1998. "Conducting Social Science Laboratory Experiments on the World Wide Web." *Library & Information Science Research* 20 (1): 5–21.
Rosenbaum, Paul R. 1984. "The Consquences of Adjustment for a Concomitant Variable that Has Been Affected by the Treatment." *Journal of the Royal Statistical Society. Series A (General)* 147: 656–66.
Rosenthal, Robert. 1976. *Experimenter Effects in Behavioral Research*. New York: Irvington Publishers.
Rosnow, Ralph, and Robert Rosenthal. 1997. *People Studying People: Artifacts and Ethics in Behavioral Research*: New York: Freeman.
Sawyer, Alan G. 1975. "Demand Artifacts in Laboratory Experiments in Consumer Research." *Journal of Consumer Research* 1 (4): 20–30.
Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51: 515–30.
Sherman, Susan R. 1967. "Demand Characteristics in an Experiment on Attitude Change." *Sociometry* 30 (3): 246–60.
Siah, Cha Yeow. 2005. "All that Glitters Is Not Gold: Examining the Perils and Obstacles in Collecting Data on the Internet." *International Negotiation* 10 (1): 115–30.
Sniderman, Paul M. 2011. "The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, Jame H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press, 102–15.
Tomz, Michael R., and Jessica L. P. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107 (4): 849–65.
Weber, Stephen J., and Thomas D. Cook. 1972. "Subject Effects in Laboratory Research: An Examination of Subject Roles, Demand Characteristics, and Valid Inference." *Psychological Bulletin* 77 (4): 273–95.
White, Ariel, Anton Strezhnev, Christopher Lucas, Dominika Kruszewska, and Connor Huff. 2018. "Investigator Characteristics and Respondent Behavior in Online Surveys." *Journal of Experimental Political Science* 5 (1): 56–67.
Zizzo, Daniel John. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics* 13 (1): 75–98.

# Online Appendix for "Demand Effects in Survey Experiments: An Empirical Assessment"

Jonathan Mummolo and Erik Peterson

# Appendix A: Experimental Materials

## Examples of Experimental Treatments

Figure A1: One version of the framing experiment treatment, in which the article raises concerns about free speech.

**Aryan Nation Tests Denver's Commitment to Free Speech**
Susan Peterson, NY Times

How far is Denver, Colo. prepared to go to protect freedom of speech? The Aryan Nation (AN, a white supremacist organization) has requested a permit to conduct a speech and rally in Denver during the spring of 2017. Numerous courts have ruled that the U.S. Constitution ensures that AN has the right to speak and hold rallies on public grounds, and that individuals have the right to hear AN's message if they are interested. Officials will decide whether to approve or deny the request in January.

Opinion about the speech and rally is mixed. Many community members worry about the rally, but support the group's right to speak. Clifford Strong, a Northwestern University law professor, remarked, "I hate the Aryan Nation, but they have the right to speak, and people have the right to hear them if they want to. We may have some concerns about the rally, but the right to speak and hear what you want takes precedence over our fears about what could happen."

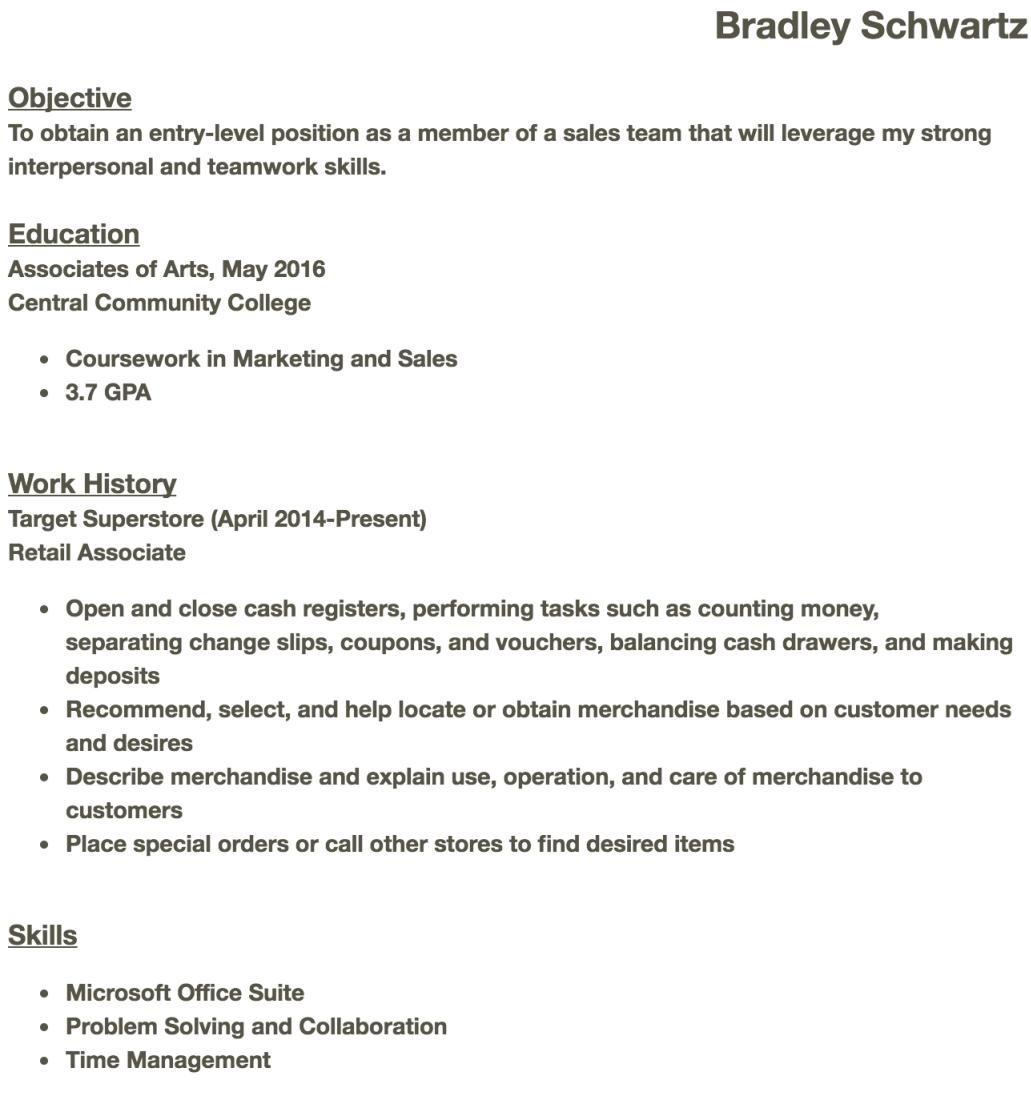
Figure A2: A sample news selection task.

**Which news item would you prefer to read?**

| | News Item A | News Item B |
|---|---|---|
| **Source:** | **Fox News** | **USA Today** |
| **Headline:** | **Trump Revives Keystone Pipeline Rejected by Obama** | **Boy, 17, Charged With Attempted Murder in School Shooting** |

News Item A ○ News Item B ○

Figure A3: One version of the resumé treatment, in which the applicant's name indicates he is white.



**Bradley Schwartz**

**Objective**

**To obtain an entry-level position as a member of a sales team that will leverage my strong interpersonal and teamwork skills.**

**Education**

**Associates of Arts, May 2016**
**Central Community College**

- **Coursework in Marketing and Sales**
- **3.7 GPA**

**Work History**

**Target Superstore (April 2014-Present)**
**Retail Associate**

- **Open and close cash registers, performing tasks such as counting money, separating change slips, coupons, and vouchers, balancing cash drawers, and making deposits**
- **Recommend, select, and help locate or obtain merchandise based on customer needs and desires**
- **Describe merchandise and explain use, operation, and care of merchandise to customers**
- **Place special orders or call other stores to find desired items**

**Skills**

- **Microsoft Office Suite**
- **Problem Solving and Collaboration**
- **Time Management**

Figure A4: One version of the democratic peace experiment, in which the hypothetical country is described as not a democracy.

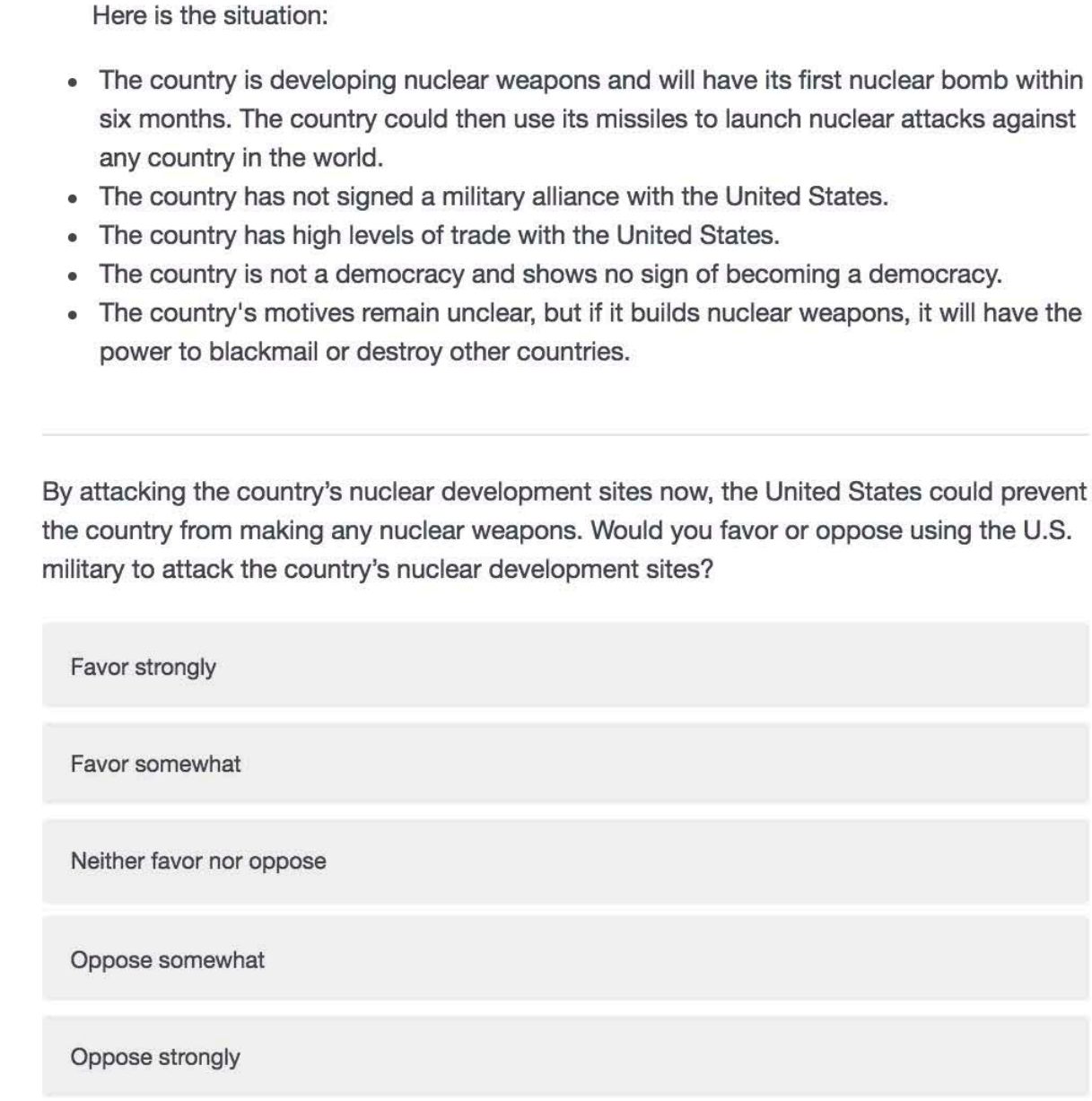
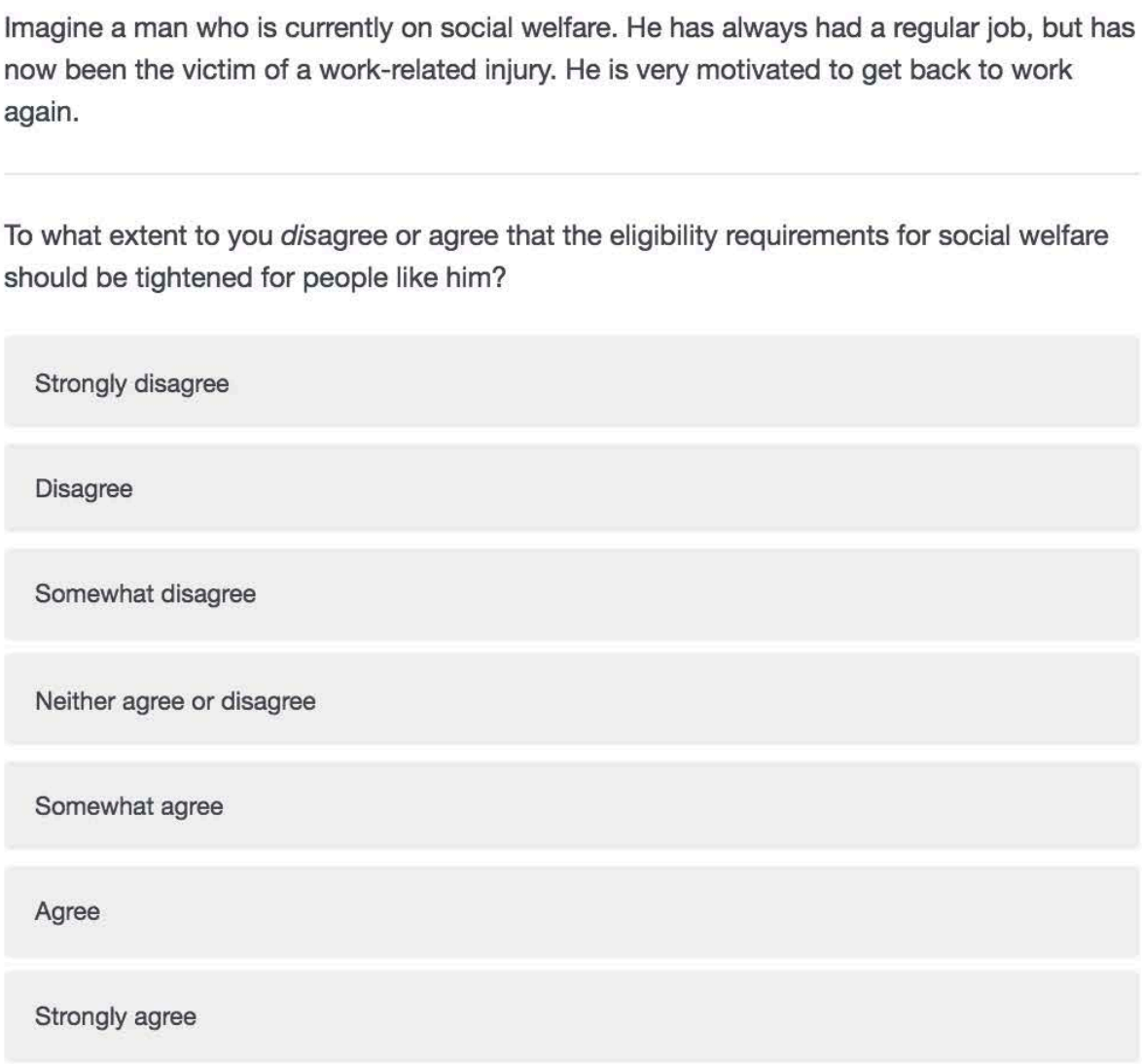


Figure A5: One version of the welfare experiment, in which the welfare recipient is described as unlucky.



## Manipulation Checks

Figure A6: The multiple choice question given to respondents after participating in the free speech framing experiment in Survey 1.

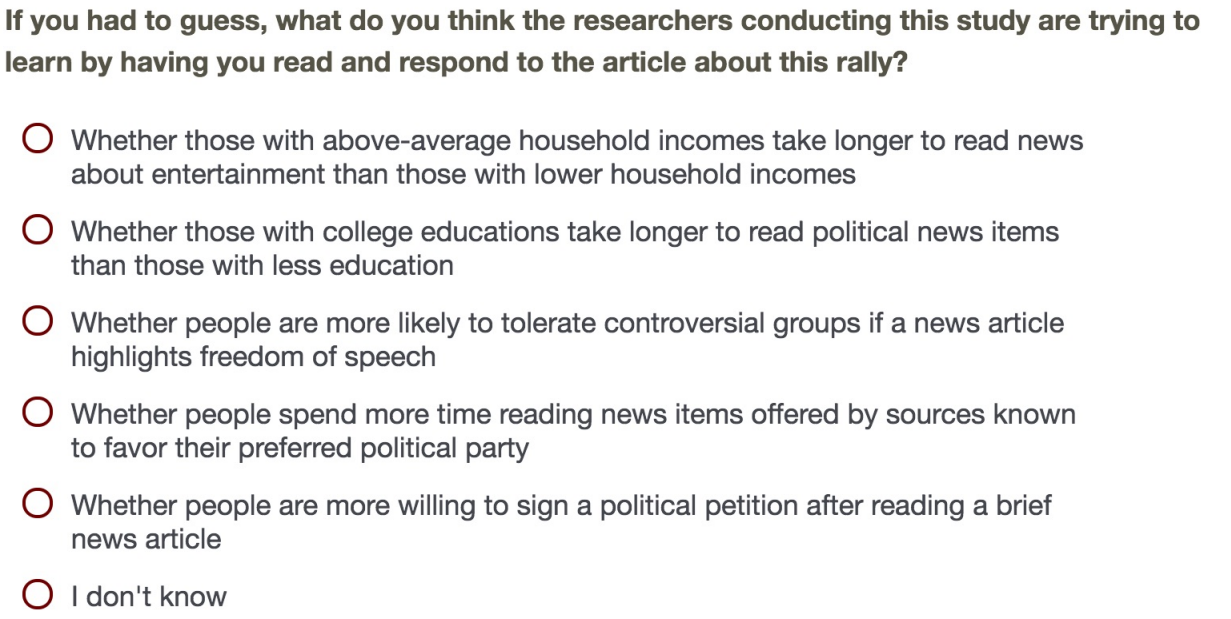


Figure A7: The multiple choice question given to respondents after participating in the partisan selective exposure experiment in Survey 1.

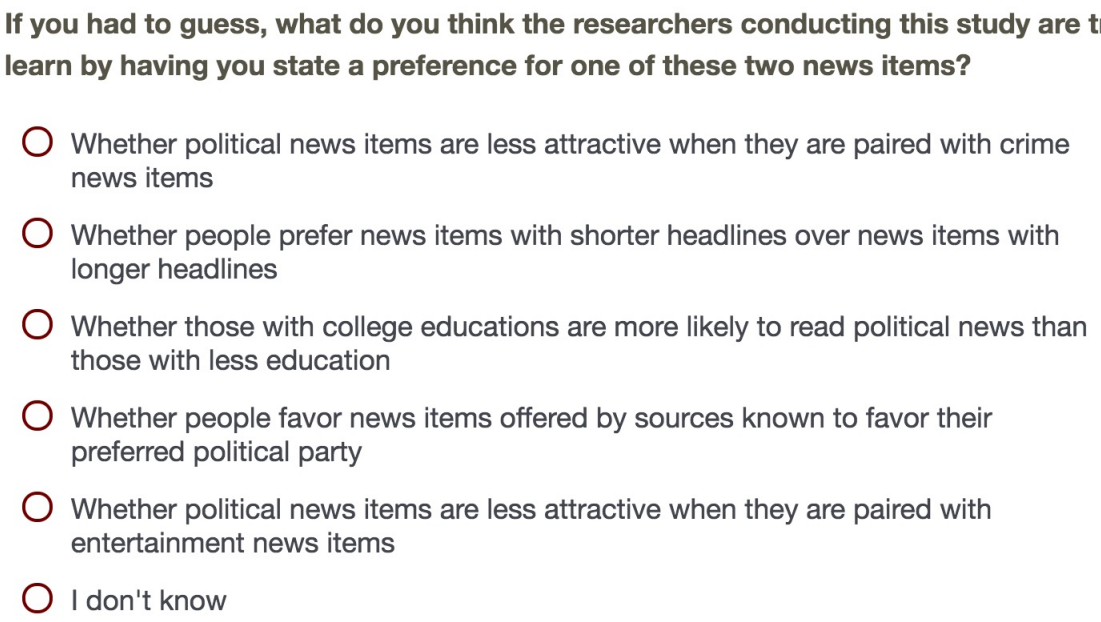



Figure A8: The multiple choice question given to respondents after participating in the free speech resumé experiment in Survey 2.

Figure A9: The multiple choice question given to respondents after participating in the partisan selective exposure experiment in Survey 2.

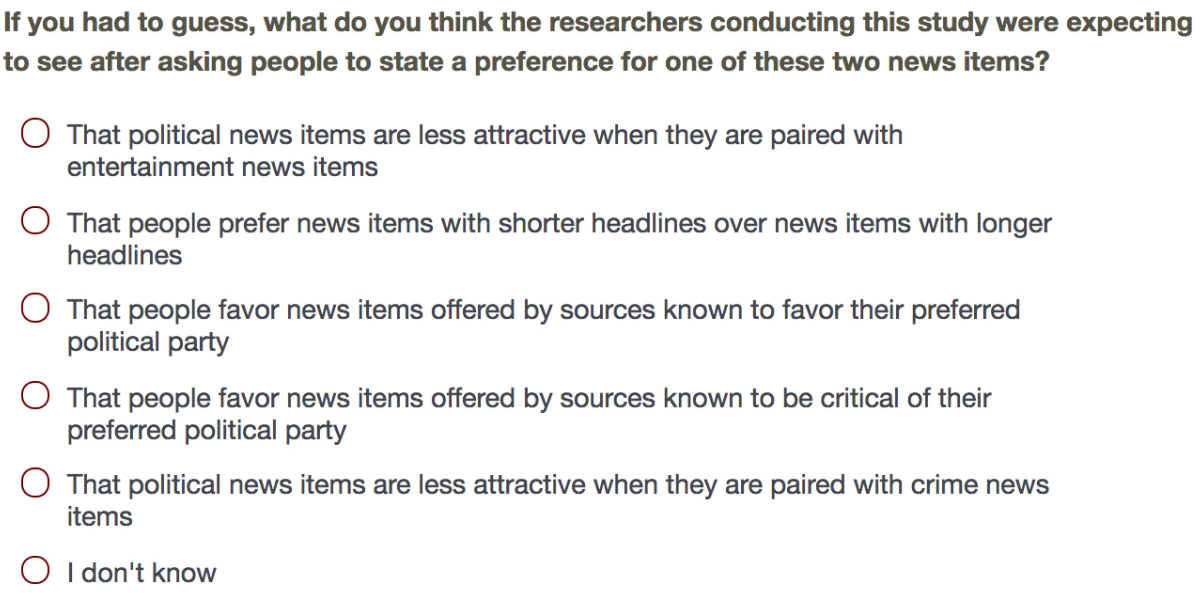


Figure A10: The multiple choice question given to respondents after participating in the democratic peace experiment in Survey 3.

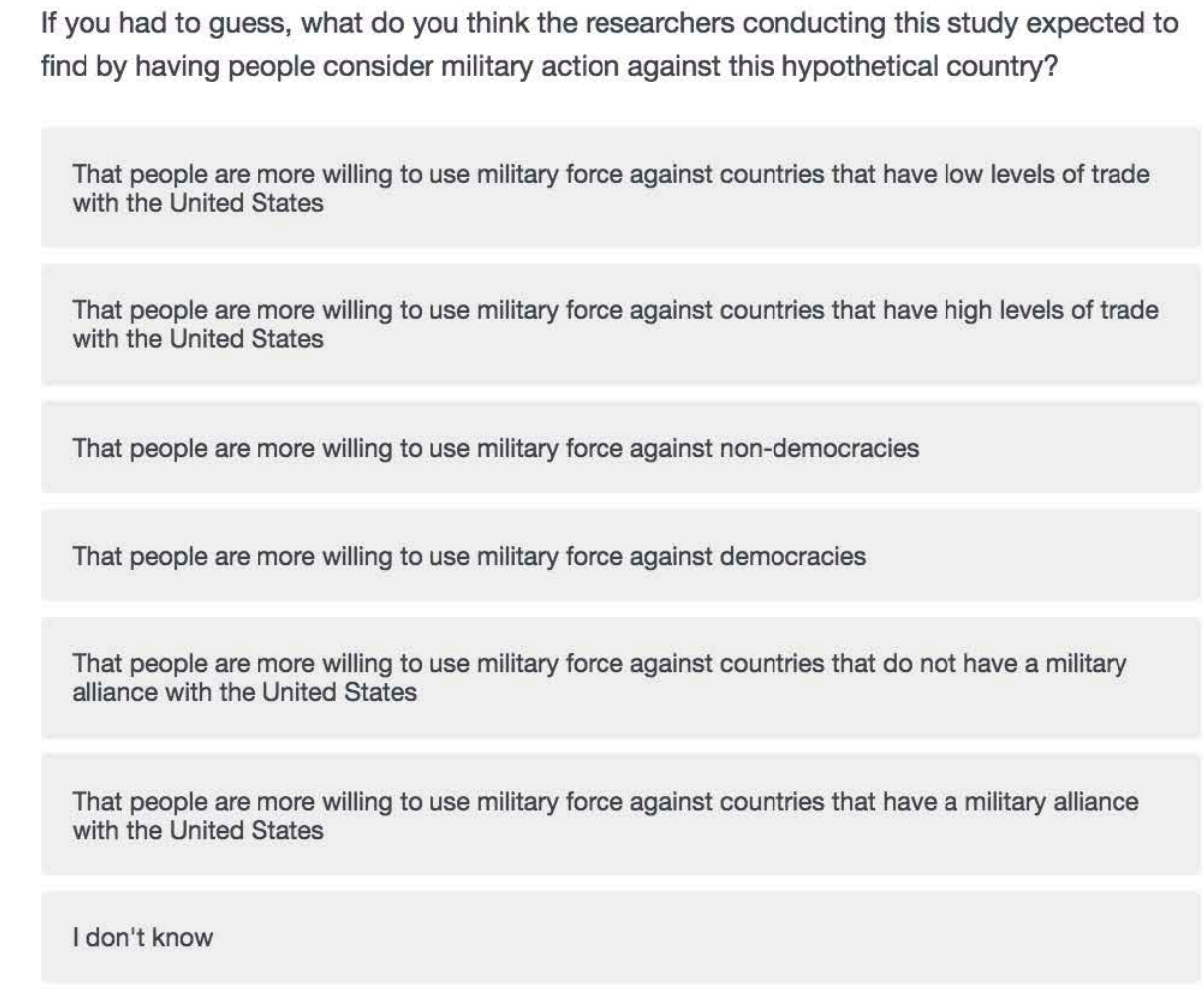
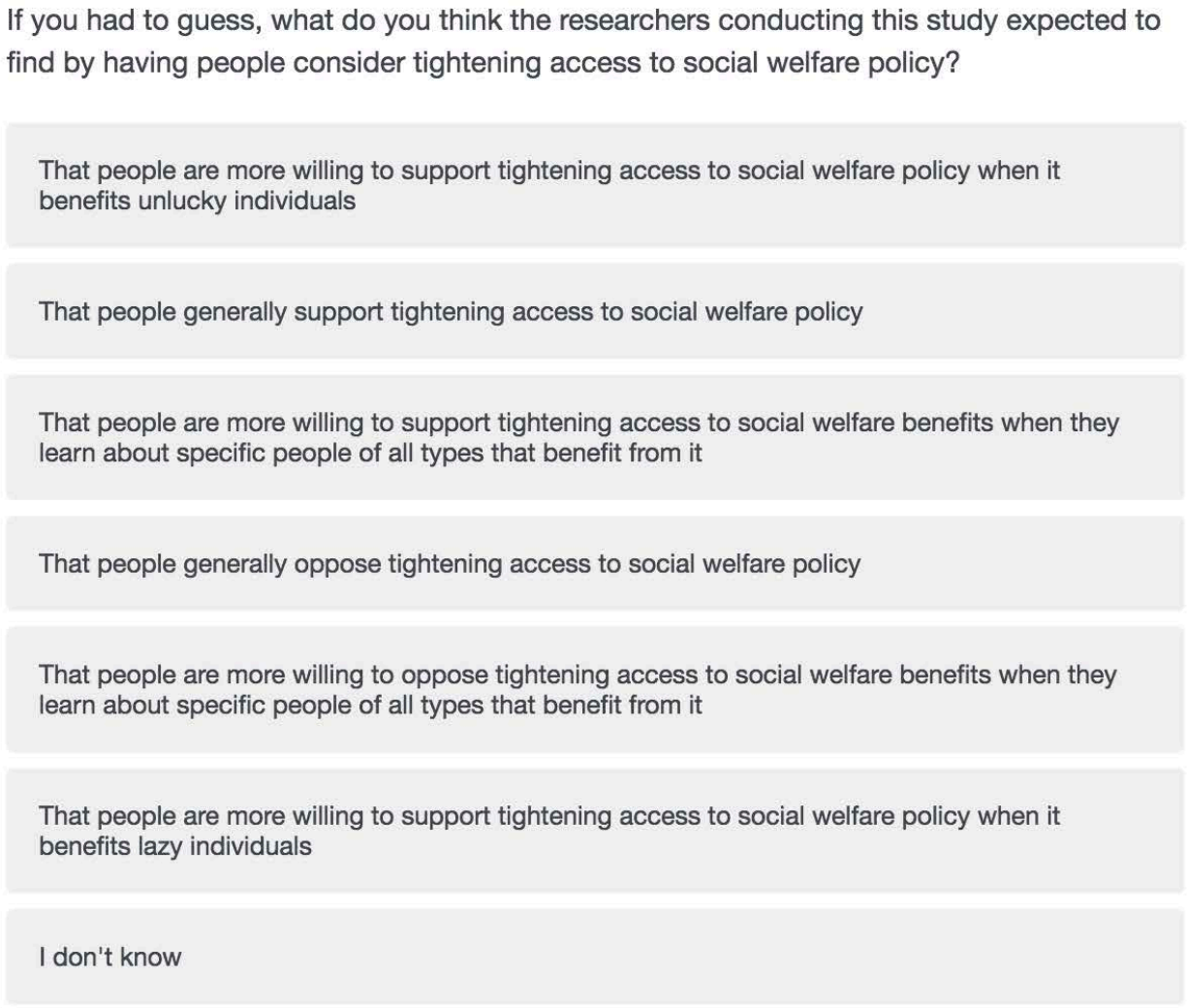


Figure A11: The multiple choice question given to respondents after participating in the welfare experiment in Survey 3.



## Additional EDE Treatments

Table A1: EDE Treatments in Resumé Experiment in Survey 2

| **Treatment Condition** | **Resumé Experiment** |
|---|---|
| Control | "Think of yourself as a Human Resources officer tasked with determining which applicants should receive interviews for an entry-level sales position at a large corporation. On the following screen, you will see a hypothetical (not real) resumé and be asked to answer the questions that follow |
| Hypothesis 1 | "Think of yourself as a Human Resources officer tasked with determining which applicants should receive interviews for an entry-level sales position at a large corporation. On the following screen, you will see a hypothetical (not real) resumé and be asked to answer the questions that follow. The purpose of this exercise is so we can measure whether the race of a job applicant affects how likely people are to receive an interview callback. We expect that job candidates with names indicating they are white will be more likely to receive an interview because of the historical advantages this group has had on the job market." |
| Hypothesis 2 | "Think of yourself as a Human Resources officer tasked with determining which applicants should receive interviews for an entry-level sales position at a large corporation. On the following screen, you will see a hypothetical (not real) resumé and be asked to answer the questions that follow. The purpose of this exercise is so we can measure whether the race of a job applicant affects how likely people are to receive an interview callback. We expect that job candidates with names indicating they are African American will be more likely to receive an interview because corporations are increasingly looking to diversify their workforces." |

Table A2: EDE Treatments in Framing Experiment in Surveys 1 and 5

| Gradation Scheme | | Incentive Scheme | |
|---|---|---|---|
| **Control:** | "Please read the article on the following screen below about a hypothetical (not real) situation." | **Control:** | "Please read the article on the following screen below about a hypothetical (not real) situation." |
| **Hint:** | "Please read the article on the following screen below about a hypothetical (not real) situation." The purpose of this is so we can measure whether the content of the article affects people's attitudes toward controversial groups in society." | **Explicit:** | "Please read the article on the following screen below about a hypothetical (not real) situation." The researchers conducting this study expect that highlighting freedom of speech will make people more tolerant of controversial groups in society. |
| **Explicit:** | "Please read the article on the following screen below about a hypothetical (not real) situation." The purpose of this is so we can measure whether highlighting freedom of speech makes people more tolerant of controversial groups in society." | **Explicit + Incentive:** | "Please read the article on the following screen below about a hypothetical (not real) situation." The researchers conducting this study expect that highlighting freedom of speech will make people more tolerant of controversial groups in society. If your responses support this theory, you will receive a $0.25 bonus payment! |

Table A3: EDE Treatments in Democratic Peace and Welfare Experiments in Surveys 4 and 5

| Incentive Scheme | | Incentive Scheme | |
|---|---|---|---|
| **Control:** | "There is much concern these days about the spread of nuclear weapons. We are going to describe a situation the U.S. could face in the future. For scientific validity the situation is general, and is not about a specific country in the news today. Some parts of the description may strike you as important; other parts may seem unimportant. After describing the situation, we will ask your opinion about a policy option." | **Control:** | "We are interested in how people evaluate social welfare policy. After describing a situation, we will ask your opinion about a policy option." |
| **Explicit:** | "There is much concern these days about the spread of nuclear weapons. We are going to describe a situation the U.S. could face in the future. For scientific validity the situation is general, and is not about a specific country in the news today. Some parts of the description may strike you as important; other parts may seem unimportant. After describing the situation, we will ask your opinion about a policy option. The researchers conducting this survey expect that individuals are less likely to support military action against democratic countries than non-democratic countries." | **Explicit:** | "We are interested in how people evaluate After describing a situation, we will ask your opinion about a policy option. The researchers conducting this survey expect that individuals will support tightening welfare policy when welfare recipients are described as lazy and oppose tightening welfare policy when welfare recipients are described as unlucky." |
| **Explicit + Incentive:** | "There is much concern these days about the spread of nuclear weapons. We are going to describe a situation the U.S. could face in the future. For scientific validity the situation is general, and is not about a specific country in the news today. Some parts of the description may strike you as important; other parts may seem unimportant. After describing the situation, we will ask your opinion about a policy option. The researchers conducting this survey expect that individuals are less likely to support military action against democratic countries than non-democratic countries. If your responses support this theory, you will receive a $0.25 bonus payment!" | **Explicit + Incentive:** | "We are interested in how people evaluate After describing a situation, we will ask your opinion about a policy option. The researchers conducting this survey expect that individuals will support tightening welfare policy when welfare recipients are described as lazy and oppose tightening welfare policy when welfare recipients are described as unlucky. If your responses support this theory, you will receive a $0.25 bonus payment!" |

## Appendix B: Additional Experimental Results

### Descriptive Statistics

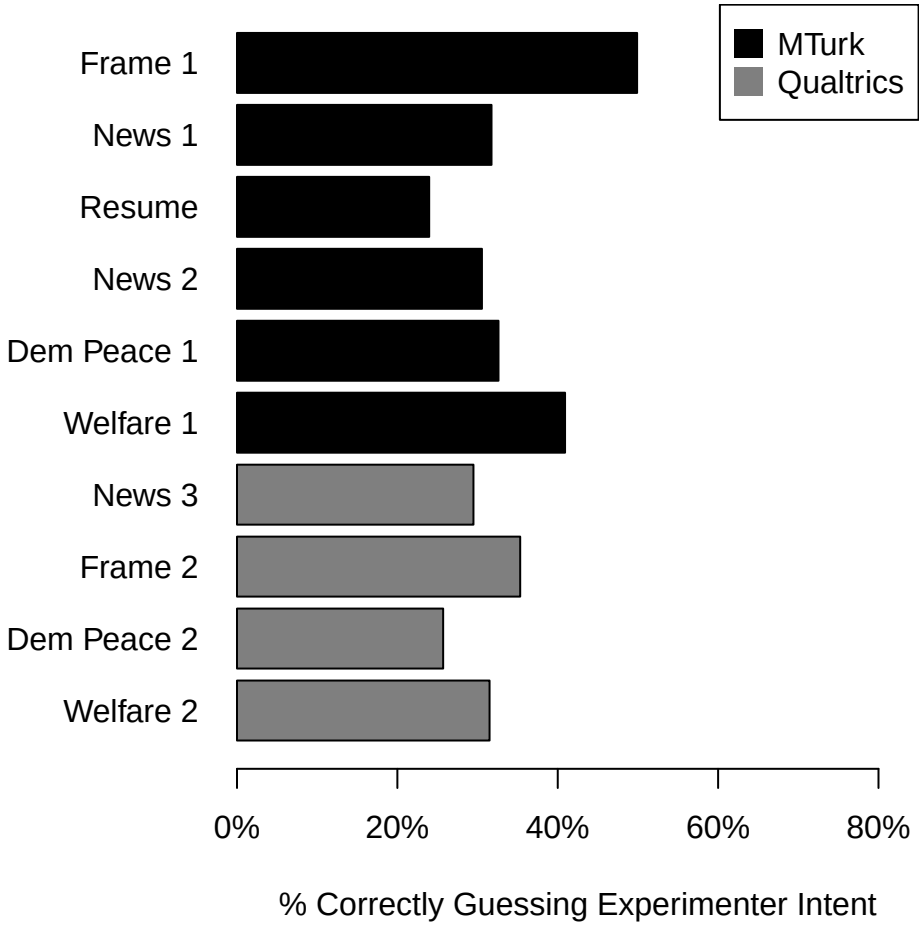
The table below displays descriptive statistics across these samples. The first three studies were convenience samples from Mechanical Turk. The last two studies were samples from an online survey vendor, Qualtrics. The values for age and income represent the mean values for each survey. The other categories represent the proportion of survey respondents in that category. For partisanship, we pool "leaners" together with the parties they lean toward. Information on gender was not collected in Surveys 1 and 2.

Table B1: Survey Demographics

| | Survey 1 | Survey 2 | Survey 3 | Survey 4 | Survey 5 |
|---|---|---|---|---|---|
| Black | 0.09 | 0.08 | 0.09 | 0.08 | 0.10 |
| Hispanic | 0.08 | 0.08 | 0.07 | 0.05 | 0.10 |
| White | 0.74 | 0.74 | 0.73 | 0.80 | 0.73 |
| Other Race | 0.09 | 0.10 | 0.09 | 0.07 | 0.07 |
| College or More | 0.52 | 0.52 | 0.51 | 0.64 | 0.65 |
| Female | | | 0.55 | 0.51 | 0.50 |
| Age | 37.40 | 36.76 | 41.47 | 47.85 | 47.12 |
| Income ($) | 55,608 | 59,470 | 59,929 | 73,584 | 75,342 |
| Democrat | 0.57 | 0.59 | 0.57 | 0.52 | 0.48 |
| Republican | 0.30 | 0.31 | 0.30 | 0.48 | 0.45 |
| Independent | 0.13 | 0.11 | 0.12 | 0.00 | 0.07 |
| Sample Size | 1,395 | 1,635 | 1,874 | 2,374 | 5,550 |

An important descriptive quantity that emerges from these studies is the share of respondents who are able to ascertain a survey experiment's purpose in the conditions where they were not provided any additional information. The figure displays this separately for each of the studies used here. Each bar represents the share of individuals able to pick the correct hypothesis from a closed-choice list after participating in the study in the conditions where they were not offered any additional information.

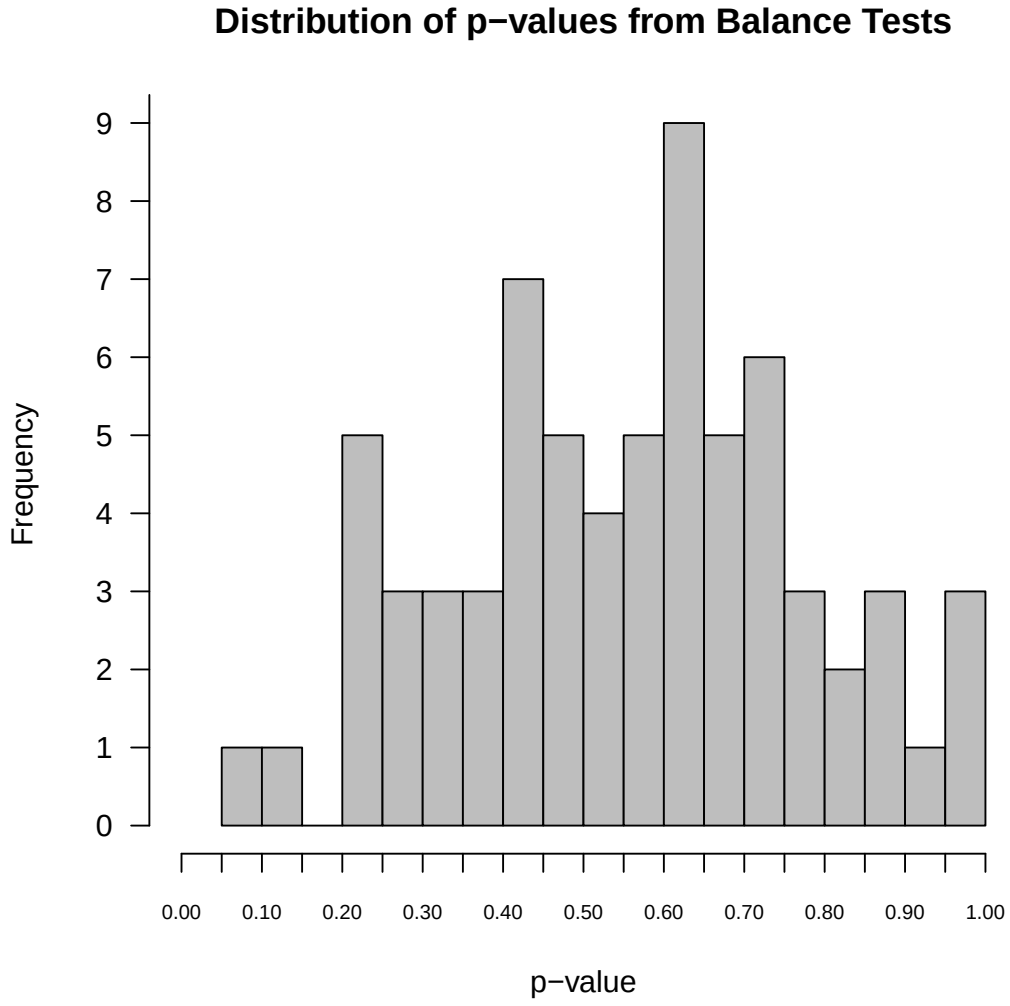
Figure B1: **Rate of Guessing Intent In Baseline Conditions.** The figure displays the proportion of respondents who guessed each experiment's intent in the baseline conditions that did not provide any additional information about the researcher's hypothesis.




In general sizable majorities are unable to infer experimenter intent when they are not provided with additional information. This occurs even when they are provided with a closed list of options and have just participated in the study.

## Balance Tests

Figure B2: The histogram displays the distribution of $p$-values generated by $F$ tests to assess balance on observables across treatment conditions in all experiments. Indicators for being in a single treatment arm of the experiment were regressed on measures of race, gender, partisanship, education, income and age. The $F$ tests assess the null hypothesis that the coefficients on these covariates are jointly zero, which should be the case if randomization achieved adequate balance. The results indicate adequate balance on observables.

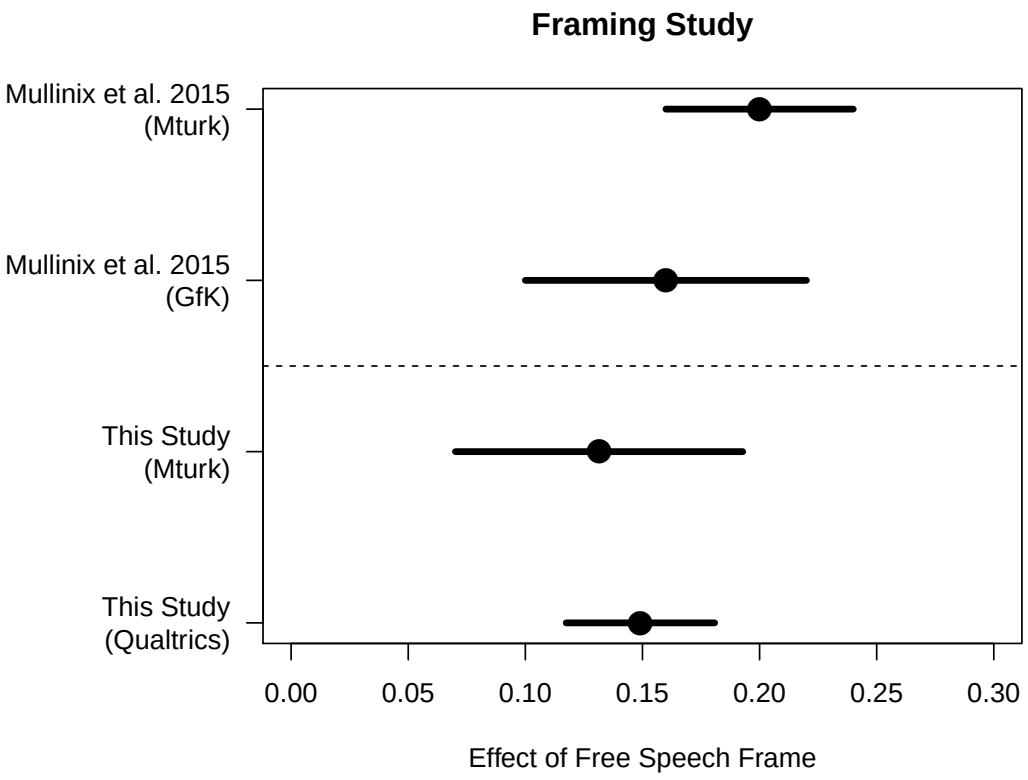


**Replication of Original Experiments**

The next set of figures compare the treatment effects in the experimental conditions without additional information to the earlier experiments they replicate. The results in our studies closely follow those from prior work, with the exception of the resumé experiment, which was originally conducted as a field experiment on actual employers. This offers greater confidence that the overall experimental context for our study and the "baseline" conditions to which we compare the various demand effect conditions are typical of online experimental settings.

First, we compare the effect of a free speech frame on support for permitting a hate-group rally in our studies to estimates from Mullinix et al. (2015), which conducts the same experiment on a convenience sample from Mturk and a nationally-representative sample from GfK/Knowledge Networks. The effects from both samples used in this prior study are both close to the estimates obtained in this study in our Survey 2 (conducted on an Mturk sample) and Survey 5 (conducted on a sample from Qualtrics).

Figure B3

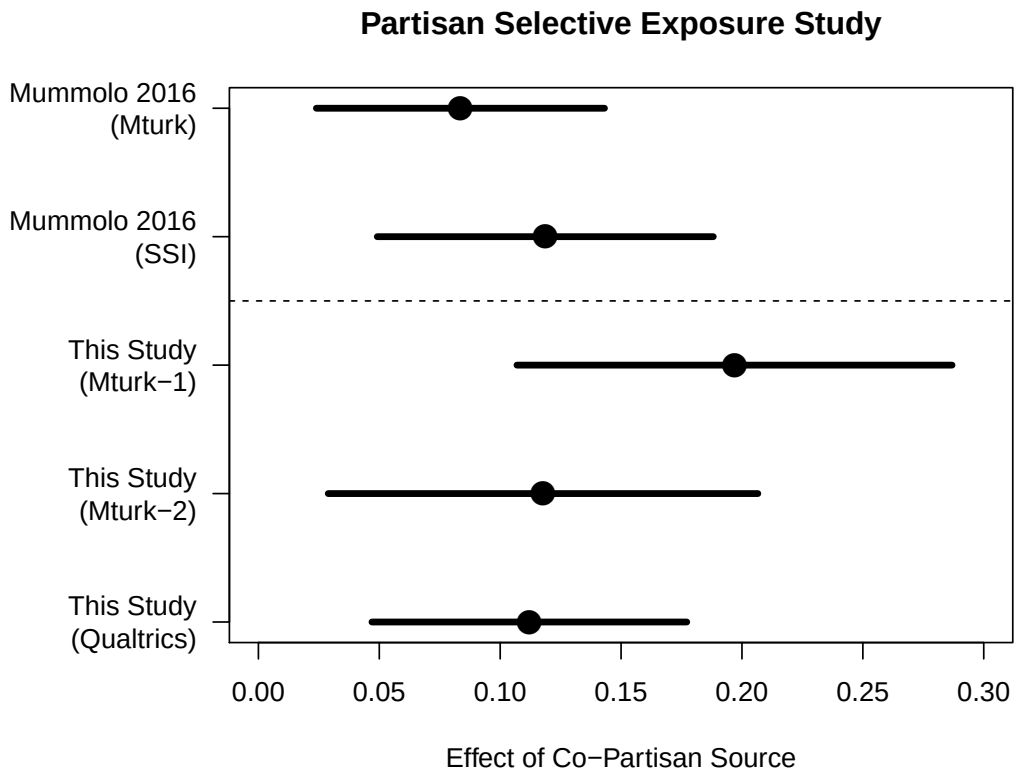



We next examine the effect of a co-partisan news source on the probability an individual chooses to read an article from that source. This experiment was substantively inspired by Iyengar and Hahn (2009) and based on the conjoint design in Mummolo (2016), which conducts the experiment on samples from Mturk and SSI. In the present study, we conduct similar experiments in Surveys 1 and 2, conducted on Mturk samples, and Survey 4, conducted on a Qualtrics sample. For the closest comparison between these sets of studies we make two adjustments to this replication data. First, because Mummolo (2016) uses full randomization of news sources, there are many profiles where individuals select between two pieces of content from the same source. In contrast, our experiments used randomization from a list of three news sources without replacement, to ensure the content always came from different outlets. For this reason we remove all the same-outlet conjoint pairs from

the Mummolo (2016) data when comparing results. Second, the experiments presented here involved only one round of news selection whereas Mummolo (2016) asked individuals to evaluate multiple rounds, with a modest decline in the effects of co-partisanship in later rounds of the experiment. We focus on comparing our experiments to the first conjoint round from Mummolo (2016) to offer the closest correspondence between the two sets of studies.

After making these adjustments there is close correspondence between the sets of results. The magnitude and direction of these treatments is similar in the new set of studies to this earlier work.

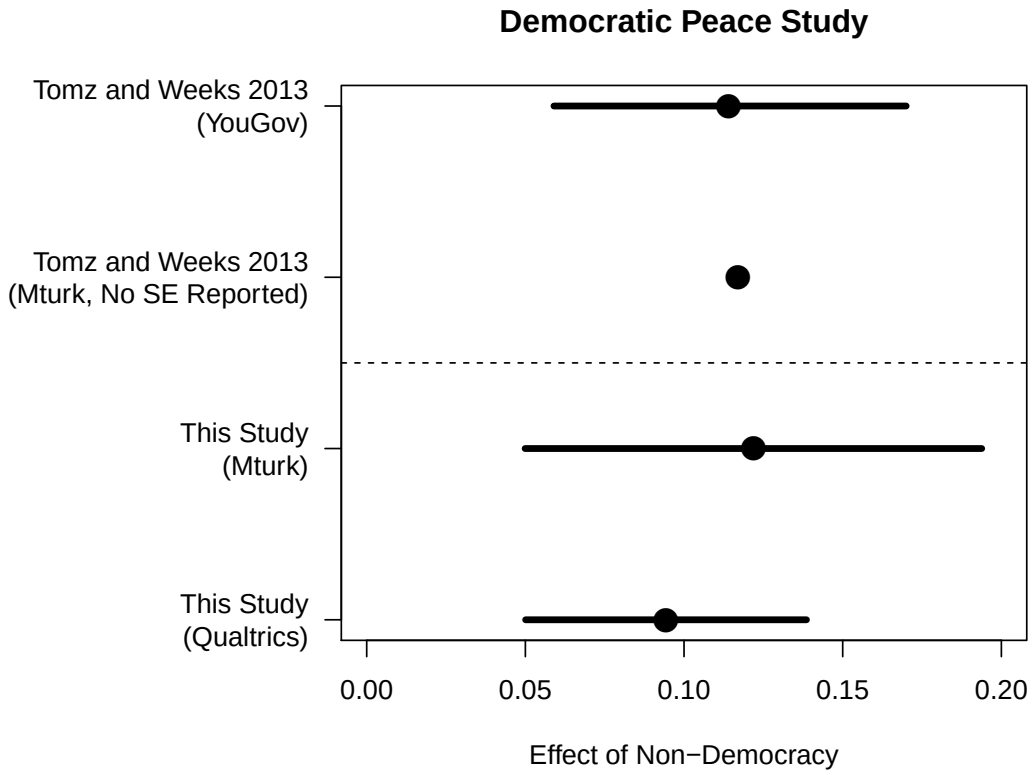
Figure B4




We compare the effect of a country being a non-democracy on the probability that survey

respondents are willing to support an attack on a proposed nuclear facility to the effect recovered in the original study, Tomz and Weeks (2013). In the original study, the authors employ a sample from YouGov and also conduct a replication study on MTurk, although they do not report a standard error for this second test. We adjust the coding of our main outcome variable to align with the binary coding used in the original study and compare these effects to estimates from Study 3, conducted on an Mturk sample, and Study 5, conducted on a sample from Qualtrics. We observe close correspondence in the effect estimates across these different studies.

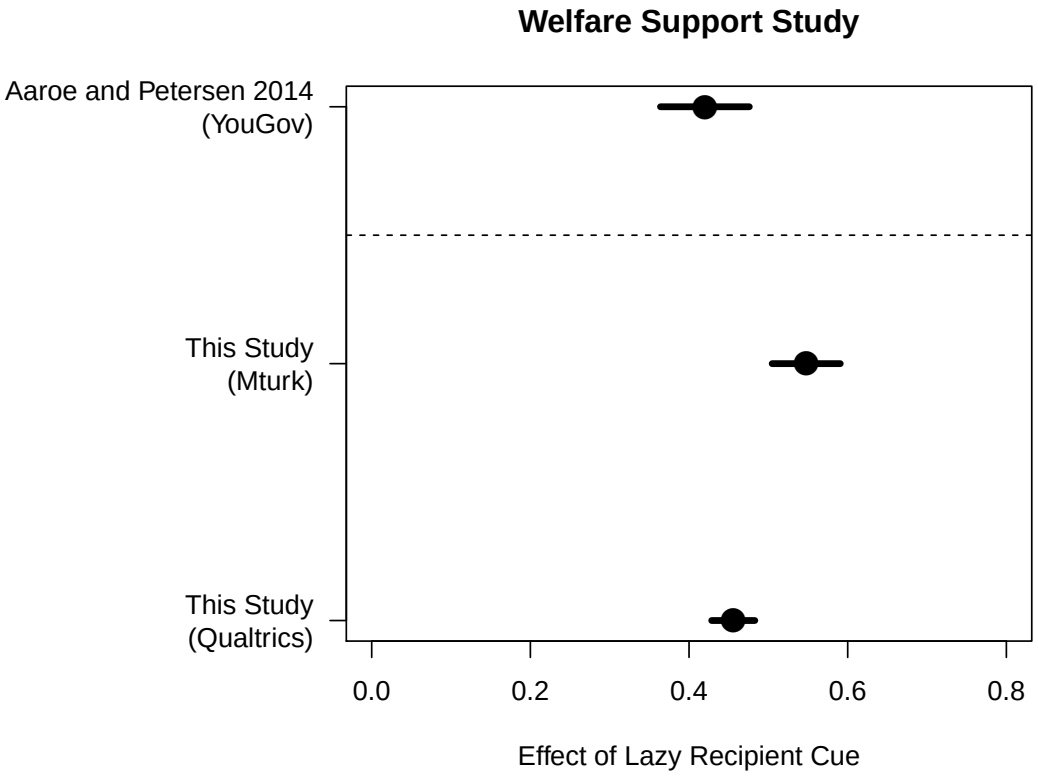
Figure B5




We next compare the effect of a welfare recipient being described as lazy, relative to unlucky, on support for making access to welfare more restrictive. In the original study Aarøe and Petersen (2014) use a two-country sample. Here we focus on comparing our

results to what the original study produced among respondents from the United States in a sample drawn from a YouGov panel. Our results are drawn from Survey 3 (an Mturk sample) and Survey 5 (a Qualtrics sample). Once again the direction and magnitude of these effect estimates are similar to those obtained in the original study across both replications.

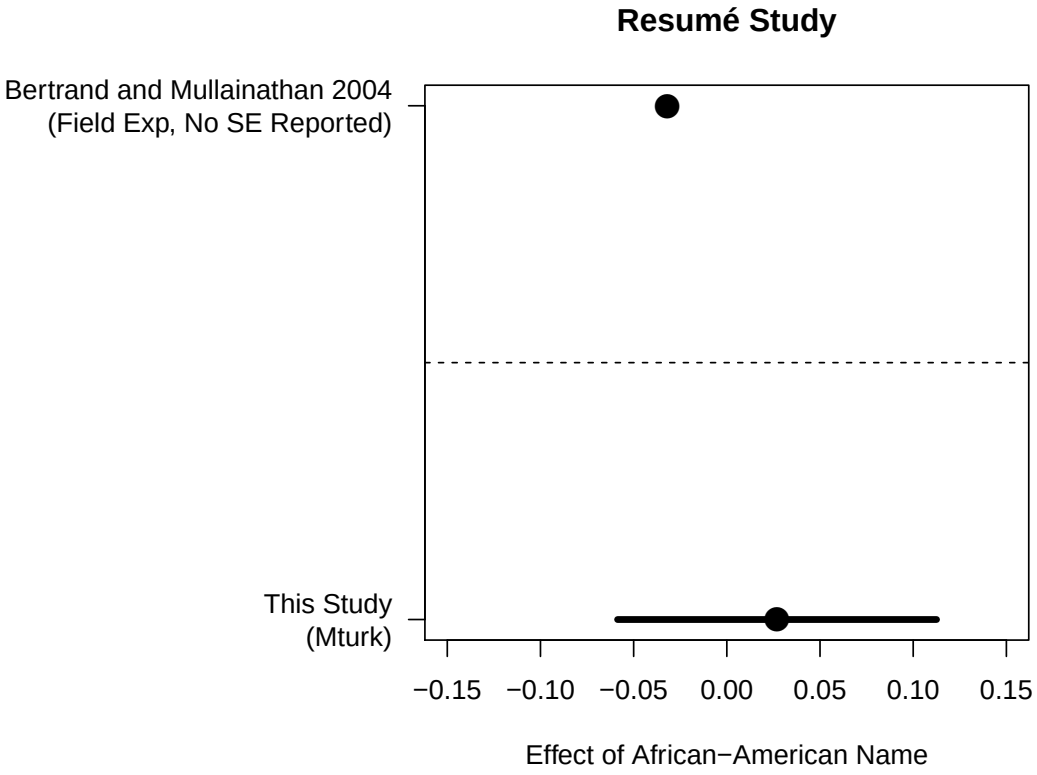
Figure B6




One study that does display a discrepancy with prior work is the resumé experiment. In the original study, a field experiment, Bertrand and Mullainathan (2004) show that African-American names reduce callback rates for a job application. Our Survey 2 included a version of this study asking individuals to evaluate a resumé on Mechanical Turk. We recode the 5-pt scale from the survey experimental outcome into a binary measure to more closely mirror the outcome in the original study. In this instance we find a small, positive point estimate for the effect of an African-American name on support for offering the applicant

an interview. This estimate does not reach statistical significance, but is in the opposite direction of the original study which finds a statistically significant decrease on the outcome. We suspect this disparity is due to the fact that Bertrand and Mullainathan (2004) was a field experiment conducted on actual employers, not a survey experiment conducted on the mass public, (though a recent labor market field experiment (Deming et al. 2016) also failed to find consistent race effects).

Figure B7




### Additional Analyses

This table separately displays the treatment effect estimates for each demand condition across all the studies discussed here.

Table B2: Treatment Effects By Demand Condition - All Studies

| | Survey | Experiment | Demand Condition | Effect | SE | Lower CI | Upper CI |
|---|---|---|---|---|---|---|---|
| 1 | Survey 1 | Framing 1 | Replication | 0.13 | 0.03 | 0.07 | 0.19 |
| 2 | Survey 1 | Framing 1 | Replication+Hint | 0.18 | 0.03 | 0.11 | 0.24 |
| 3 | Survey 1 | Framing 1 | Replication+Explicit | 0.09 | 0.03 | 0.03 | 0.16 |
| 4 | Survey 1 | News 1 | Replication | 0.20 | 0.04 | 0.11 | 0.29 |
| 5 | Survey 1 | News 1 | Replication+Hint | 0.07 | 0.05 | -0.02 | 0.16 |
| 6 | Survey 1 | News 1 | Replication+Explicit | 0.07 | 0.05 | -0.02 | 0.16 |
| 7 | Survey 2 | Resume | Replication | 0.04 | 0.02 | -0.01 | 0.08 |
| 8 | Survey 2 | Resume | Replication+Negative Effect | 0.07 | 0.02 | 0.03 | 0.11 |
| 9 | Survey 2 | Resume | Replication+Positive Effect | 0.06 | 0.02 | 0.02 | 0.10 |
| 10 | Survey 2 | News 2 | Replication | 0.12 | 0.04 | 0.03 | 0.21 |
| 11 | Survey 2 | News 2 | Replication+Negative Effect | 0.10 | 0.04 | 0.01 | 0.19 |
| 12 | Survey 2 | News 2 | Replication+Positive Effect | 0.16 | 0.04 | 0.08 | 0.25 |
| 13 | Survey 3 | Dem Peace | Replication | 0.09 | 0.03 | 0.03 | 0.14 |
| 14 | Survey 3 | Dem Peace | Replication+Explicit | 0.10 | 0.02 | 0.05 | 0.15 |
| 15 | Survey 3 | Dem Peace | Replication+Incentive | 0.23 | 0.03 | 0.18 | 0.28 |
| 16 | Survey 3 | Welfare | Replication | 0.55 | 0.02 | 0.50 | 0.59 |
| 17 | Survey 3 | Welfare | Replication+Explicit | 0.58 | 0.02 | 0.54 | 0.62 |
| 18 | Survey 3 | Welfare | Replication+Incentive | 0.60 | 0.02 | 0.55 | 0.64 |
| 19 | Survey 4 | News 3 | Replication | 0.11 | 0.03 | 0.05 | 0.18 |
| 20 | Survey 4 | News 3 | Replication+Explicit | 0.11 | 0.03 | 0.05 | 0.18 |
| 21 | Survey 4 | News 3 | Replication+Incentive | 0.10 | 0.03 | 0.03 | 0.16 |
| 22 | Survey 5 | Framing 2 | Replication | 0.15 | 0.02 | 0.12 | 0.18 |
| 23 | Survey 5 | Framing 2 | Replication+Explicit | 0.16 | 0.02 | 0.13 | 0.19 |
| 24 | Survey 5 | Framing 2 | Replication+Incentive | 0.13 | 0.02 | 0.10 | 0.16 |
| 25 | Survey 5 | Dem Peace 2 | Replication | 0.08 | 0.02 | 0.05 | 0.11 |
| 26 | Survey 5 | Dem Peace 2 | Replication+Explicit | 0.09 | 0.02 | 0.06 | 0.12 |
| 27 | Survey 5 | Dem Peace 2 | Replication+Incentive | 0.09 | 0.02 | 0.06 | 0.12 |
| 28 | Survey 5 | Welfare 2 | Replication | 0.46 | 0.01 | 0.43 | 0.48 |
| 29 | Survey 5 | Welfare 2 | Replication+Explicit | 0.44 | 0.01 | 0.42 | 0.47 |
| 30 | Survey 5 | Welfare 2 | Replication+Incentive | 0.49 | 0.01 | 0.46 | 0.51 |

Table B3: Treatment Effects Conditional on Correct Guess - All

| | Framing 1 | News 1 | Resume | News 2 | Dem Peace 1 | Welfare 1 | Dem Peace 2 | Welfare 2 | Framing 2 |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 0.54* | 0.46* | 0.68* | 0.45* | 0.33* | 0.21* | 0.42* | 0.35* | 0.44* |
| | (0.04) | (0.03) | (0.03) | (0.02) | (0.03) | (0.03) | (0.02) | (0.02) | (0.02) |
| Treatment | 0.19* | 0.12 | 0.07 | 0.16* | 0.16* | 0.57* | 0.10* | 0.37* | 0.16* |
| | (0.05) | (0.09) | (0.04) | (0.08) | (0.05) | (0.04) | (0.03) | (0.03) | (0.03) |
| Guess First | 0.16* | 0.02 | −0.04 | −0.06 | −0.00 | −0.00 | −0.04 | −0.08* | 0.13* |
| | (0.07) | (0.04) | (0.06) | (0.04) | (0.05) | (0.05) | (0.04) | (0.04) | (0.04) |
| Treatment*Guess First | −0.23* | −0.06 | 0.05 | 0.18 | −0.10 | 0.06 | −0.02 | 0.15* | −0.07 |
| | (0.09) | (0.13) | (0.07) | (0.13) | (0.08) | (0.06) | (0.06) | (0.05) | (0.06) |
| $N$ | 227 | 418 | 235 | 416 | 275 | 295 | 624 | 638 | 622 |

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

The table displays the subsequent treatment effects among those who did and did not correctly guess the hypothesis of the first experiment they participated in. This analysis includes those who guessed correctly in the first experiment regardless of whether they were given additional information about the hypothesis or not.

Table B4: Treatment Effects Conditional on Correct Guess - Baseline

| | Framing 1 | News 1 | Resume | News 2 | Dem Peace 1 | Welfare 1 | Dem Peace 2 | Welfare 2 | Framing 2 |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 0.55* | 0.41* | 0.62* | 0.48* | 0.37* | 0.27* | 0.41* | 0.32* | 0.46* |
| | (0.07) | (0.05) | (0.07) | (0.05) | (0.07) | (0.06) | (0.04) | (0.04) | (0.04) |
| Treatment | 0.16 | 0.27 | 0.03 | 0.06 | 0.12 | 0.48* | 0.08 | 0.39* | 0.15* |
| | (0.09) | (0.15) | (0.09) | (0.14) | (0.09) | (0.08) | (0.06) | (0.05) | (0.06) |
| Guess First | 0.09 | 0.03 | 0.05 | −0.09 | −0.03 | 0.00 | −0.11 | −0.06 | 0.03 |
| | (0.12) | (0.08) | (0.11) | (0.07) | (0.11) | (0.10) | (0.07) | (0.06) | (0.07) |
| Treatment*Guess First | −0.25 | −0.12 | 0.00 | 0.32 | −0.12 | 0.06 | 0.08 | 0.12 | 0.03 |
| | (0.17) | (0.21) | (0.13) | (0.22) | (0.14) | (0.13) | (0.10) | (0.08) | (0.10) |
| $N$ | 76 | 138 | 81 | 126 | 84 | 97 | 202 | 227 | 201 |

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

The table displays the subsequent treatment effects among those who did and did not correctly guess the hypothesis of the first experiment they participated in. This analysis is limited to those who were assigned to the control condition (no additional information on the hypothesis) in the first experiment.

The next two tables pool together results from the various studies for additional precision. The table below examines whether, across all the studies used here, exposure to any of the demand conditions (i.e., hint, explicit, directional or incentive) produced any detectable variation in the experimental treatment effects relative to the baseline conditions that received no additional information. Whether combining all of these studies together or separating the studies on Mechanical Turk and Qualtrics there is no detectable shift in the treatment effects estimated in these studies based on the demand conditions.

Table B5: Pooled Estimates of Treatment Effect Variation by Demand Condition

| | All | Mturk Studies | Qualtrics Studies |
|---|---|---|---|
| (Intercept) | 0.49* | 0.50* | 0.43* |
| | (0.03) | (0.04) | (0.05) |
| Treatment | 0.21* | 0.19* | 0.23* |
| | (0.07) | (0.07) | (0.11) |
| Demand Condition | 0.01 | 0.00 | 0.02 |
| | (0.01) | (0.01) | (0.02) |
| Treatment × Demand Condition | −0.02 | 0.01 | −0.03 |
| | (0.02) | (0.03) | (0.04) |
| $N$ | 33027 | 11631 | 21396 |

Models include study fixed effects, continuous outcomes rescaled between 0-1

Robust standard errors, clustered by study, in parentheses

* indicates significance at $p < 0.05$

The table below examines this same result this time for the set of studies included in surveys 3, 4 and 5 that shared a similar incentive scheme where respondents were in a baseline condition, received information about experimenter intent or received information about experimenter intent and an incentive to respond in a manner consistent with these expectations. This separates out the effects of the information and incentive treatments.

In the pooled analysis there are no detectable changes in the treatment effect based on the availability of either incentives or information in the demand conditions. This also holds when subsetting to just those participants in the Qualtrics surveys (surveys 4 and 5). The one set of results that does offer evidence of changes in these treatment effects occurs when focusing

on the Mechanical Turk respondents who encountered these conditions (Survey 3). Here there is a small upward shift in the treatment effect based on the availability of information and a much larger shift based on the availability of incentives. Across the set of ten studies conducted on five different surveys this is the lone instance where we observe a shift in respondents due to the demand effect treatments, and this is concentrated in the conditions that heighten the incentives and information necessary to comply with experimenter demand to levels that are not present in typical survey experimental environments.

Table B6: Pooled Estimates of Treatment Effect Variation by Demand Condition Type

| | All | Mturk Studies | Qualtrics Studies |
|---|---|---|---|
| (Intercept) | 0.27* | 0.21* | 0.45* |
| | (0.04) | (0.10) | (0.01) |
| Treatment | 0.22* | 0.32 | 0.20* |
| | (0.08) | (0.23) | (0.01) |
| Demand Condition-Information | −0.00 | −0.01 | 0.00 |
| | (0.00) | (0.02) | (0.01) |
| Demand Condition-Incentive | −0.00 | −0.03* | 0.00 |
| | (0.01) | (0.00) | (0.01) |
| Treatment × Demand Condition-Information | 0.01 | 0.02* | 0.00 |
| | (0.01) | (0.01) | (0.01) |
| Treatment × Demand Condition-Incentive | 0.02 | 0.10* | 0.00 |
| | (0.02) | (0.05) | (0.01) |
| $N$ | 25057 | 3661 | 21396 |

Models include study fixed effects, continuous outcomes rescaled between 0-1

Includes studies from Surveys 3, 4 and 5 with both Information and Incentive conditions

Robust standard errors, clustered by study, in parentheses

* indicates significance at $p < 0.05$

Figure B8: **Three-Point EDEs Would Have Few Consequences for Inference.** The figure shows estimated treatment effects from 20 survey experiments fielded on M-Turk and through Time-Sharing Experiments for the Social Sciences (TESS) from Mullinix et al. (2015). Revised effects are plotted in absolute value, and red and black points denote effects that change in terms of either sign or significance after a three-point demand effect is imposed (i.e., after subtracting three points from positive effects, and adding three points to negative effects). Diluting treatment effects by three percentage points would change the sign of four out of 40 effects, though all of those were not statistically significant to begin with, so there would be no change in inference. Two additional effects lose statistical significance using ± two-Std. Error confidence intervals. The vast bulk of substantive conclusions from these studies would remain unchanged given a three-point demand effect.