Demand Effects in Survey Experiments: An Empirical Assessment

Jonathan Mummolo and Erik Peterson[*]
Stanford University
Dept. of Political Science
Encina Hall West, Suite 100
Stanford, CA 94305

April 21, 2017

Abstract

Survey experiments are ubiquitous in the behavioral social sciences. A frequent critique of this type of study is that evidence which supports a researcher's expectations occurs due to Experimenter Demand Effects (EDEs)—a form of bias in which participants infer the purpose of an experiment and respond so as to help confirm a researcher's hypothesis. In this paper, we argue that traditional survey experimental designs possess several features that make them robust to these concerns. We then explicitly test for the presence of EDEs using a series of experiments which randomly assign participants to receive varying levels of information about each experiment's hypothesis. Replicating three widely used experimental designs, we find that informing participants of an experiment's purpose either has no detectable effect on the observed treatment effects, or severely attenuates them. Even when informed of a study's purpose, participants do not appear inclined to assist researchers. These findings have important implications for the design and interpretation of survey experiments.

[*] Jonathan Mummolo and Erik Peterson are Ph.D. candidates in Stanford's Department of Political Science. They can be reached at jmummolo@stanford.edu and epete1@stanford.edu.

A long-standing critique of social science experiments is that evidence which supports researcher expectations is an artifact produced by "experimenter demand effects" (EDEs), (Orne 1962, Sears 1986, Zizzo 2010, Iyengar 2011). 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 efforts researchers expend to guard against them, ranging from subtle efforts to disguise experimental treatments and outcome measures, to deception intended to mask a study's intent.[2]

While the concept originated to critique laboratory experiments (Orne 1962), concerns about EDEs are, in theory, highly relevant for survey experiments. A particular concern is that survey experiments frequently utilize online subject pools, like Amazon's Mechanical Turk, where experienced experimental participants have incentives to be especially attentive to researcher expectations (Goodman et al. 2013, Krupnikov and Levine 2014). In a highly influential study,[3] Berinsky et al. (2012, 366) recommend researchers avoid revealing their intentions in online survey experiments due to concerns about EDEs:

> MTurk 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

[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] These design features also sometimes serve additional purposes beyond alleviating EDEs.

[3] As of January 2017, Berinsky et al. 2012 had over 1,100 citations on Google Scholar.

> 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 MTurk respondent pool . . .

If present, EDEs could render experimental results in an array of major literatures unreliable. Yet there is little evidence demonstrating (1) the existence of EDEs in survey experiments or (2) the magnitude of these effects. Replicating several highly cited experimental designs, we assess the severity and consequences of demand effects by randomly assigning participants to receive varying levels of information about the purpose of each experiment before participating. While this meta-randomization increased the share of respondents aware of each experiment's hypothesis—and thus, the theoretical risk of EDEs—we find no evidence that this knowledge led participants to assist in confirming the stated hypotheses. In some cases, alerting participants to the study's purpose produced no detectable difference in treatment effects. In others, this information led to a severe attenuation of treatment effects.

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. Even when made aware of the experiment's goal, respondents did not appear inclined to assist researchers. Counter to the traditional EDE critique, but consistent with research on experiments conducted on non-naive samples (Chandler et al. 2015), we also find that treatment effects were sometimes smaller among the participants with knowledge of an experiment's purpose. These results have important implications for the design, implementation and interpretation of survey experiments, and suggest that efforts to obfuscate the aim of experimental studies may at best be unnecessary, and at worst unintentionally "stack the deck" in favor of the hypotheses being tested.

Concerns About Experimenter Demand Effects

Orne (1962) raises a fundamental concern for the practice of experimental social science research. Orne argues that, 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 to be 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 Zisso 2010, Bortolotti and Mameli 2006, Rosnow and Rosenthal 1997, Weber and Cook 1972).

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 experimental 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 et al. 2012), survey vendors (Mutz 2011), local advertisements (Kam et al. 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 Sniderman 2011, Mutz 2011, Gaines et al. 2007 for discussions of the growth of survey experiments in political science).

Experimenter Demand Effects in Survey Experiments

Widespread concerns about the problems that EDEs pose for survey experiments are reflected in the efforts researchers make to circumvent them when designing experiments. These countermeasures stem from a shared assumption that EDEs can be limited by obfuscating an experimenter's intentions from participants.

In one approach, researchers disguise experimental treatments and primary outcome measures from participants. Fowler and Margolis (2014, 103) embed information about the issue positions of political parties inside 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 et al. (2014, 27) advocate the use of "conjoint" experiments, in which respondents choose between two alternatives (e.g., political candidates) based on 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 employ cover stories to misdirect participants about the goal of the experiment (e.g., McDermott 2002, Bortolotti and Mameli, 2006; Dickson 2011). 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 a set of 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. Moreover, as we review in this section, there are also 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 blind, between-subject designs with participants unaware of the experimental cell in which they have been placed. In these studies, experimental treatments are embedded inside a broader survey instrument, blurring the line between the experimental sections of the study and non-randomized material that all respondents encounter.

This creates a complicated pathway for participants to infer experimenter intentions. Not only must they parse the experimental and non-experimental portions of the survey instrument, but having done so, they need to reason out the broader experimental design and determine the experimenter-intended behaviors, even as they only encounter the contents of a single cell of the broader experimental design. If an error occurs in this process, even would-be "helpful" subjects will 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. In laboratory experiments, the primary source of demand effects 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 in experimental research at the outset of this type of research (Piper 1998; McDermott 2002, 34; Siah 2005, 122-123)

A second set of potential limitations is that, even if participants correctly infer experimenter intent, 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 go in the opposite direction of 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 represent the on-average inclination of a set of subjects, it would mean that the presence of demand characteristics works *against* a researcher's hypotheses. While there is limited empirical evidence on the distribution of these groups in various subject pools, prior research offers 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 et al. 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 an account where they attempt to validate researcher hypotheses.

Despite the widespread application of the EDE concept to survey experiments, our discussion highlights several elements that may limit demand effects in survey experiments. However, there is limited empirical 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 three well-known experimental designs while also randomizing the degree to which the purpose of the experiment is revealed to participants. Our data come from adult volunteers recruited on Amazon's Mechanical Turk, which hosts an experienced pool of survey respondents (see e.g., Berinsky et al. 2012; Hitlin 2016). While this pool of participants may present disadvantages for many studies, we view it as an ideal source in this context. If EDEs are present, we are likely to observe them in this group, since experienced survey takers are more likely to comprehend the purpose and structure of survey-based experimental designs. Prior research also 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 et al. 2012).

We conducted two surveys in early 2017, each of which contained two experiments. In the first study (N=1,373), we test for EDEs in the context of a classic framing study, a substantive area where concerns over demand effects have been expressed in laboratory experimental contexts (e.g., Page 1970, Sherman 1967). In this experiment, respondents were asked to read a hypothetical news article about a white supremacist group attempting to hold a rally in a U.S. city (Nelson et al. 1997, Mullinix et al. 2015). In the control condition, respondents saw an article which merely describes the group's request to hold the rally. In the treatment condition, respondents saw a version of the article that highlights the group's first amendment right to hold the rally. Following the article, both groups were asked how willing they would be to allow the rally. They were also asked whether they would be willing to sign a petition in support of the group's right to hold the rally, a "decoy" dependent variable intended to help make the purpose of the experiment less obvious ex ante. The hypothesis, based on prior findings, was that those exposed to the free speech frame will 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 2x2 table (see Figure 4 in the Appendix), each with randomized headlines and sources, and asked them to state a preference for one or the other. The sources were 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 et al. 2014), in which each of the two news items offered to each respondent was treated as a separate observation.[4]

Prior to participating in these experiments, participants were randomly assigned to one of three conditions. In the control condition, respondents were simply told that they would be asked to read a news article/select a preferred news item. In the "hint" condition, participants learned the topic of the experiment (i.e., whether the news outlet influences which news items are preferred). The "explicit" condition informed participants of the researcher's hypothesis, including a description of the treatment and the direction of the expected treatment response. Table 1 displays the wording of these conditions for each

[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 8 in the Appendix displays the results of tests for balance on observables for all experiments.

experiment.

Table 1: Text of Treatments Conveying Information on Purpose of Experiment (Study 1)

Treatment Condition	**Free Speech Experiment**	**Partisan News Experiment**
Control	“Please read the article on the following screen below about a hypothetical (not real) situation.”	“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	“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.”	“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.”
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.”	“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.”

We sought to replicate and extend our results in the second study (N=1,512). We again conducted two experiments, this time assigning respondents to either receive no information about the study’s intent, to be told the effect of the treatment would be positive, or to be told the effect of the treatment would be negative. We again included the news conjoint experiment, telling treated respondents that we expected them to choose the politically friendly or unfriendly sources, along with brief rationales to make these messages more believable (see Table 2 for exact wording). The second experiment in this study was a classic resumé experiment, in which we randomly assigned participants to see a resumé for a job applicant who had either a historically white name (Bradley Schwatz) or historically African American name (DeAndre Jefferson) (Names drawn from Butler and Homola N.D.; See Figure 3 in the Appendix). We then told respondents to imagine they were the human resources employee and asked them to indicate how likely they would be to call the applicant for an interview. As with the news experiment, we assigned each participant to either encounter no information about the hypothesis, to be told we expected white candidates to be preferred, or to be told we expected black candidates to be preferred. In each case the hypothesis was accompanied by a brief rationale.

Table 2: Text of Treatments Conveying Information on Purpose of Experiment (Study 2)

Treatment Condition	**Resumé Experiment**	**Partisan News 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	“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.”
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.”	“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.”
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.”	“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 second study allows us to determine whether demand effects are more likely to occur when the hypothesis that participants believe is being tested runs in a particular direction. This distinction is important. In Study 1 we may detect no demand effects if the hypothesis we shared with respondents was already in line with the one they would have inferred in the absence of our revealing it. Study 2 thus provides another avenue to detect demand effects.

The quantity of interest in both experiments is a difference in differences. Specifically, we seek to estimate the differences in treatment effects conditional on the information about the experiment’s purpose revealed to participants. This quantity is represented by the following expression:

$$\begin{aligned}&(E[\text{response}|\text{Treatment \& Information}] - E[\text{response}|\text{Control \& Information}])\\ &-(E[\text{response}|\text{Treatment \& No information}] - E[\text{response}|\text{Control \& No information}])\end{aligned}$$

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 to participants should lead them to assist this hypothesis and we should expect treatment effects in the presence of additional information about the experiment's aim to be larger than in the absence of such information. As manipulation checks, we also measure the degree to which the purpose of each experiment was known by respondents by asking them to choose from a multiple choice menu of five possible hypotheses following each experiment. See Figures 4 and 5 in the Appendix for the wording of these items.

Results

A first-order concern is whether respondents grasped the information these treatments revealed about the purpose of the experiments. Table 3 displays the results of OLS regressions of indicators for guessing the purpose of the experiment from a list of five possible hypotheses on indicators for the information treatment conditions. In the first study, the results show that in the free speech experiment, those who were given no information on the experiment's purpose correctly guessed the purpose of the experiment 50% of the time. In the hint condition, respondents were correct 56% of the time, while in the explicit condition, respondents were correct 64% of the time. We see similar results with the partisan news experiment. While those given no information on the experiment's purpose correctly guessed the purpose 32% of the time, those in the hint and explicit conditions were correct 37% and 51% of the time.

The treatments in the second study were also effective, and indicate that we were able to convince substantial proportions of the sample of hypotheses which implied effects that ran in opposing directions. In the resumé experiment, telling respondents that we expected white (black) job applicants to be preferred led to a 20 (22) percentage point increase in the share of respondents which later said this was the study's hypothesis. In the news experiment, the information treatments yielded 10 and 12 percentage point boosts in the share of respondents guessing that the hypothesis was that a copartisan or out-partisan news source would be preferred, respectively. Thus, this information successfully manipulated the degree to which participants understood the purpose of the experiments. This increased the potential for EDEs to occur in the information conditions of each study relative to the no information group.

Table 3: Manipulation Checks: Models of Guessing Purpose of Experiment

	Study 1		Study 2			
	Framing	News	Resumé	Resumé	News	News
(Intercept)	0.50 *	0.32 *	0.24 *	0.05 *	0.31 *	0.06 *
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
Hint	0.06	0.05				
	(0.03)	(0.03)				
Explicit	0.14 *	0.19 *				
	(0.03)	(0.03)				
White Applicant Preferred			0.20 *	0.02		
			(0.03)	(0.02)		
Black Applicant Preferred			-0.01	0.22 *		
			(0.03)	(0.02)		
Copartisan Source Preferred					-0.03	0.12 *
					(0.03)	(0.02)
Out-partisan Source Preferred					0.10 *	0.03
					(0.03)	(0.02)
N	1,373	1,373	1,512	1,512	1,512	1,512

Robust standard errors in parentheses. The outcomes in columns 3-6 are indicators for guessing that the hypothesis was that the white applicant was preferred, that the black applicant was preferred, that the copartisan news source was preferred and that the out-partisan news source was preferred, respectively.
* indicates significance at $p < 0.05$

We next evaluated the treatment effects in each experiment among those *not* told the purpose of the experiment. Table 4 displays the results of models which interact the treatments in each study with the meta treatments indicating how much information was revealed to respondents. In the first study, the coefficient on "Treatment" in Table 4 shows that framing the white supremacist rally in terms of free speech increases the probability of stating that the rally should be allowed by 13 percentage points among those not told the purpose of the study.[5] In the partisan news experiment, offering a news outlet that is known to favor the partisan identity of the survey respondent increases the probability of selection by 20 percentage points. In the second study, we find that the hypothetical job candidate with the historically Black name led to a 4 point boost in willingness to call the applicant for an interview relative to the white applicant, though the result is only borderline statistically significant in this subset of the data.[6] Finally, consistent with the results in Study 1, providing a source that is known to favor a respondent's political party led to a 10-point boost in the probability of selecting a news item.

We now turn to the main results of tests designed to gauge whether and how knowledge of

[5] The item measuring willingness to allow the rally was asked on a 7-point scale, responses were then rescaled to range between 0 and 1 so that treatment effects could be interpreted in terms of percentage points.

[6] We note that this result is at odds with field experimental studies which find that Black applicants are less likely to receive a call (Bertrand and Mullainathan 2004). We therefore suspect that the sign on this coefficient is an artifact of the survey environment in which the experiment was deployed, potentially due to social desirability bias. But since the purpose of the present study is to determine whether treatment effects vary when information about the experiment's purpose is supplied, we still regard the resumé experiment as useful. If the Black applicant is more or less preferred when information about the experiment is provided, that will convey evidence of a demand effect, regardless of whether the estimated effect of providing a historically Black name is biased.

an experiment's purpose alters observed treatment effects in the experiment. The interaction terms in Table 4 represent the differences in treatment effects between various levels of information about the purpose of the experiment and the no information conditions. Figure 1 displays the mean response in each condition for all experiments.

Table 4: Tests of Whether Information on Experiment's Purpose Alters Treatment Effects

	Study 1		Study 2	
	Framing	News	Resumé	News
(Intercept)	0.55 *	0.43 *	0.66 *	0.47 *
	(0.02)	(0.02)	(0.02)	(0.02)
Treatment	0.13 *	0.20 *	0.04	0.10 *
	(0.03)	(0.04)	(0.02)	(0.05)
Hint	-0.06	0.04 *		
	(0.03)	(0.02)		
Explicit	-0.01	0.04 *		
	(0.03)	(0.02)		
Treatment * Hint	0.05	-0.13 *		
	(0.04)	(0.06)		
Treatment * Explicit	-0.04	-0.12		
	(0.04)	(0.06)		
White Candidate Preferred			0.03	
			(0.02)	
Black Candidate Preferred			0.01	
			(0.02)	
Treatment * White Candidate Preferred			0.04	
			(0.03)	
Treatment * Black Candidate Preferred			0.02	
			(0.03)	
Unfriendly Source Preferred				0.00
				(0.02)
Friendly Source Preferred				-0.02
				(0.02)
Treatment * Unfriendly Source Preferred				-0.01
				(0.07)
Treatment * Friendly Source Preferred				0.06
				(0.06)
N	1,373	2,386	1,513	2,338

Robust standard errors in parentheses. "Treatment" indicates the free speech, copartisan news source and historically Black name conditions in the framing, partisan news and resumé experiments, respectively.

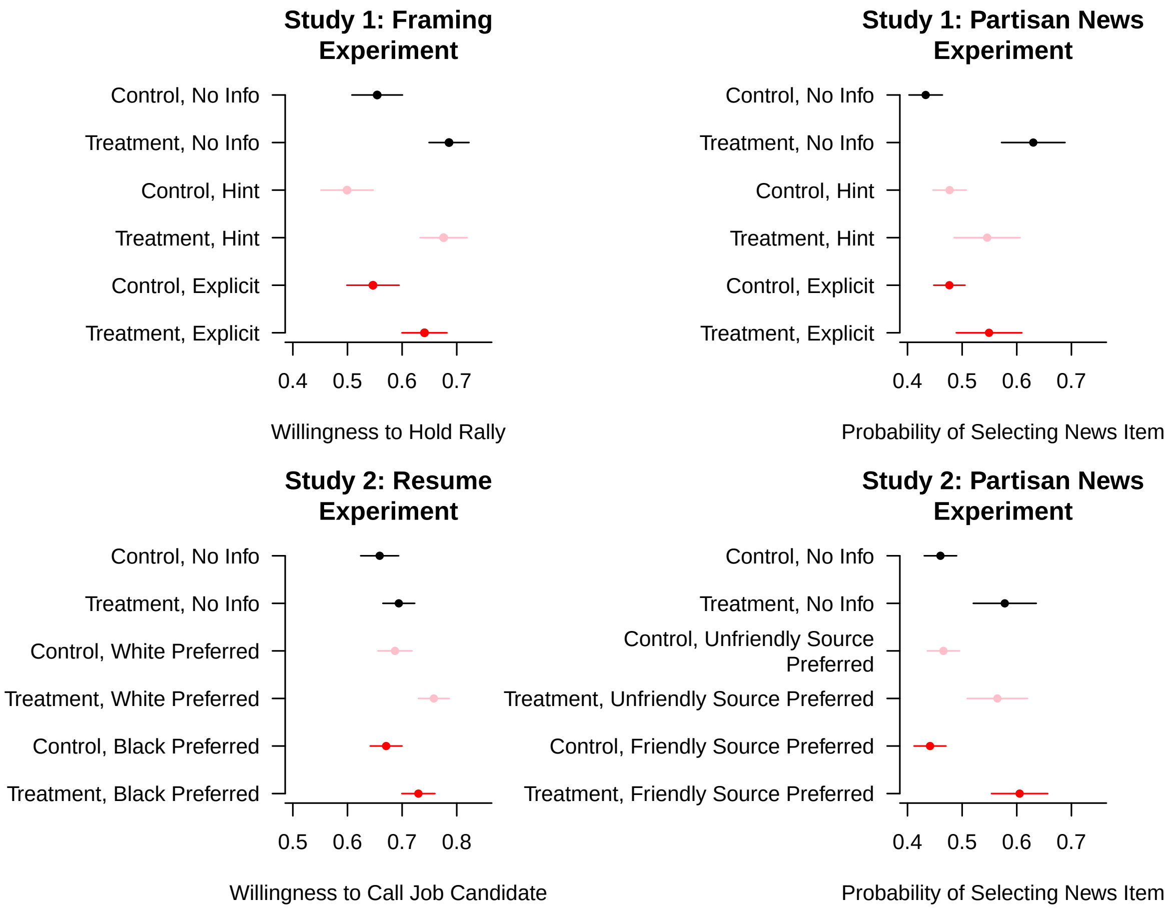
* indicates significance at $p < 0.05$

As the table shows, treatment effects in the free speech experiment among those who received a hint or explicit explanation as to the experiment's purpose were not discernibly different from those observed among those who received no added information about the experiment (i.e., we cannot reject the null hypothesis that the coefficients on the interaction terms are zero). One concern is that our sample size may be too small to detect these differences in differences. However, the fact that the coefficients on "Treatment*Hint" and "Treatment*Explicit" have opposing signs further supports the conclusion that providing information on the experiment's hypothesis produces no consistent change to observed treatment effects in this study.

We observe a different pattern of results in the first partisan news experiment. While respondents who were not given information on the experiment show a treatment effect of 20 percentage points, those in the hint and explicit conditions show effects that are 13 and 12 percentage points *smaller*, respectively. That is, when respondents were told researchers were testing whether a source's reputation influenced article selection, they behaved in a way that *muted* the effect of news source reputations.

One plausible explanation for the differing set of results centers on survey experimental design. In the framing experiment, respondents are assigned to *either* treatment or control, while in the conjoint partisan news experiment, respondents may be exposed to *both* treatment and control conditions, since they see two news items at a time, each of which serves as a separate observation in analysis. When exposed to both conditions and equipped with the knowledge of the researcher's intent, participants may find it much easier to respond strategically. These results suggest that conjoint experiments, (and perhaps all within-subjects designs), are more likely to produce attenuated treatment effects than cross-sectional designs. However, this significant interaction was not replicated in the second study, which raises doubts about whether the attenuation was real. The interaction terms in the resumé study were also indistinguishable from zero. Overall, we find no support for the key pre-

Figure 1: The figure displays the mean responses in each of the condition. All dependent variables range from 0 to 1. "Treatment" refers to the free speech frame, copartisan news source and historically Black name conditions in the framing, partisan news and resumé experiments, respectively.



diction of the demand effects hypothesis. Although we were successfully able to convince respondents as to the purpose of each experiment, revealing this information did not help to confirm the stated hypotheses.

Are "clever" respondents more prone to produce EDEs?

The previous results demonstrate what happens to treatment effects in survey experiments when conditions that are theoretically conducive to EDEs are exacerbated. Contrary to traditional characterizations of demand effects, we find that providing survey respondents information on the purpose of an experiment either has no effect on estimated treatment effects, or attenuates those effects. Still, the evidence we have presented thus far cannot rule out EDEs completely. The reason is that some people in the sample may have inferred the purpose of the experiment even without being provided additional information. If such participants respond differently when equipped with this knowledge than the participants who only knew the experiments' hypotheses because additional information was provided, than they may be capable of contributing to artificially inflated treatment effects nonetheless. In other words, it is possible that even in the control condition, where no extra information is provided, estimated treatment effects may be shaped by demand effects due to the presence of especially clever respondents.

To evaluate this possibility, we first identified respondents in the sample who were most likely to have inferred the experiments' hypotheses on their own: those who received no additional information about the first experiment they encountered in our survey, but who correctly guessed the hypothesis of that experiment nonetheless, (recall that the order of the two experiments was randomized across respondents). Conversely, we labeled respondents as not likely to infer hypotheses on their own if they were in the no-information condition in the first experiment and guessed incorrectly. We then compared the treatment effects between these two groups of respondents in the *second* experiment. If "clever" respondents are inflating treatment effects then we should observe larger effects among them than among those respondents who were unlikely to divine the purpose of the experiment on their own.

Table 5: Treatment Effects in Second Experiment Conditional on Correctly Guessing Purpose of First Experiment

	Study 1				Study 2			
	Framing	Framing	News	News	Resumé	Resumé	News	News
	(all)	(control)	(all)	(control)	(all)	(control)	(all)	(control)
(Intercept)	0.57 *	0.50 *	0.46 *	0.41 *	0.68 *	0.62 *	0.46 *	0.49 *
	(0.04)	(0.08)	(0.03)	(0.05)	(0.07)	(0.07)	(0.03)	(0.05)
Treatment	0.11 *	0.19 *	0.12	0.27	0.07	0.05	0.12	0.03
	(0.05)	(0.10)	(0.09)	(0.15)	(0.09)	(0.11)	(0.08)	(0.16)
Guess First Experiment	0.11	0.13	0.02	0.03	-0.04	0.03	-0.08	-0.10
	(0.07)	(0.14)	(0.04)	(0.08)	(0.11)	(0.09)	(0.04)	(0.07)
Treatment * Guess First Experiment	-0.11	-0.26	-0.06	-0.12	0.05	0.00	0.27 *	0.40
	(0.10)	(0.19)	(0.13)	(0.21)	(0.13)	(0.13)	(0.13)	(0.22)
N	227	76	418	138	235	81	358	110

Robust standard errors in parentheses. "Treatment" indicates the free speech, copartisan news source and historically Black name conditions in the framing, partisan news and resumé experiments, respectively.

* indicates significance at $p < 0.05$

Table 5 displays the results of models comparing treatment effects among those who did and did not correctly guess the first experiment's purpose when no additional information was provided.[7] The columns labeled "all" display the results when using all respondents in Experiment 2 regardless of whether they received extra information about that experiment, while the columns labeled "control" include only those respondents who received no additional information in both experiments. In Study 1, all interaction terms—which represent the differences in treatment effects between those who did and did not correctly guess the purpose of the first experiment—are negative, suggesting that respondents capable of inferring hypotheses are not inclined to respond in ways favorable to researchers' hypotheses. If anything, these "clever" respondents exhibit drastically smaller treatment effects than their counterparts. In the second experiment, we see positive point estimates, but only one statistically significant result, indicating that respondents who correctly guessed the purpose of the resumé experiment went on to exhibit larger treatment effects in the news experiment

[7] For Study 2, we coded respondents as correctly guessing the purpose of experiment 1 if they chose either of the two hypotheses revealed to treated respondents.

than their counterparts. However, the fact that a nearly identical experiment did not produce a similar result in Study 1 suggests this statistically significant result may be purely due to chance.

To be sure, most of these results contain a lot of uncertainty due to the drastic reduction in sample size necessary to isolate these two groups of respondents. But based on these point estimates, there is little indication that a sample with a high proportion of respondents capable of divining the purpose of experiments would bias results in favor of the researcher's hypothesis.

Discussion and Conclusion

Survey experiments have become a main staple of behavioral research across the social sciences, a trend that has been aided by the increased availability of inexpensive online participant pools. With the expansion of this class of studies, 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 capable of artificially confirming a researcher's hypothesis and which deviate from the responses that would result from sincere participation. 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 either similar or attenuated treatment effects relative to those generated when the purpose of the experiment is not provided. Further, we find that individuals who are most capable of divining an experiment's hypothesis produce smaller treatment effects than their counterparts in many cases.

These results have important implications for the design an interpretation of survey experiments moving forward. First, our results suggest that the substantial effort and resources researchers have expended attempting to obfuscate the hypothesis being tested may be misguided. If anything, our results suggest that such tactics will most likely make it easier for researchers to confirm their hypotheses. Our findings also suggest that the results from within-subjects designs such as conjoint experiments, which often expose subjects to both treatment and control conditions, are more likely to hinge on the degree to which the hypothesis being tested is known.

In light of our findings, there are several additional questions worthy of pursuit. While we are able to demonstrate how knowledge of a hypothesis changes experimental results, we are unable to comment on which set of results exhibits more bias. The reason is that the true values of the treatment effects we seek to estimate are unknown. The mechanisms behind the attenuation effects we observe are also unclear. For example, respondents aware of a study's purpose may be motivated to sabotage the study, or may seek to provide normatively desirable responses, both of which could lead to the pattern of results we witness above. In addition, there may be substantial variation in how respondents react to knowledge of an experiment's hypothesis across substantive areas. Finally, respondent pools with varying levels of experience participating in survey experiments may respond differently to the stimuli we examine here.

References

Arceneaux, Kevin. 2008. “Can Partisan Cues Diminish Democratic Accountability?” *Political Behavior* 30(2), 139-160.

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), 351-368.

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.

Butler, Daniel M. and Jonathan Homola. N.D. “An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments.” *Political Analysis*

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), 112-130.

Chandler, Jesse, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate A. Ratfliff. 2015. “Using Nonnaive Participants Can Reduce Effect Sizes.” *Psychological Science* 26(7), 1131-1139.

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), 185-194.

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), 875-896.

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, 100-110.

Frank, B. L. 1998. "Good news for the experimenters: Subjects do not care about your welfare." *Economics Letters*, 61, 171-174.

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), 213-224.

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/

Iyengar, Shanto. 2011. "Laboratory Experiments in Political Science." *Cambridge Handbook of Experimental Political Science* (Eds. James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupa). Pgs. 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. "Implicit Attitudes, Explicit Choices: When Subliminal Priming Predicts Candidate Preferences." *Political Behavior* 29, 343-367.

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), 415-440.

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

Leeper, Thomas J. and Emily Thorson. 2015. "Minimum Sponsorship-Induced Bias in Web Survey Data." Working Paper. https://dl.dropboxusercontent.com/u/414906/SurveySponsorship.pdf

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), 109-138.

Mummolo, Jonathan. 2016. "News from the Other Side: How Topic Relevance Limits the Prevalence of Partisan Selective Exposure." *The Journal of Politics*, 78(3), 763-773.

Mutz, Diana C. 2011. *Population-Based Survey Experiments.* 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), 567-583.

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), 776-783.

Orne, Martin T. and Wayne G. Whitehouse. 2000. "Demand Characteristics." in *Encyclopedia of Psychology*, ed, A.E. Kazdin. Washington, D.C.: American Psychological Association and Oxford Press, 469-470.

Page, Monte M. 1970. "Role of Demand Awareness in the Communicator Credibility Effect." *Journal of Social Psychology* 82, 57-66.

Piper, Allison I. 1998. "Conducting Social Science Laboratory Experiments on the World Wide Web." *Library & Information Science Research* 20(1), 5-21.

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.* Freeman.

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, 515-530.

Sherman, Susan R. 1967. "Demand Characteristics in an Experiment on Attitude Change." *Sociometry* 30(3), 246-260.

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), 115-130.

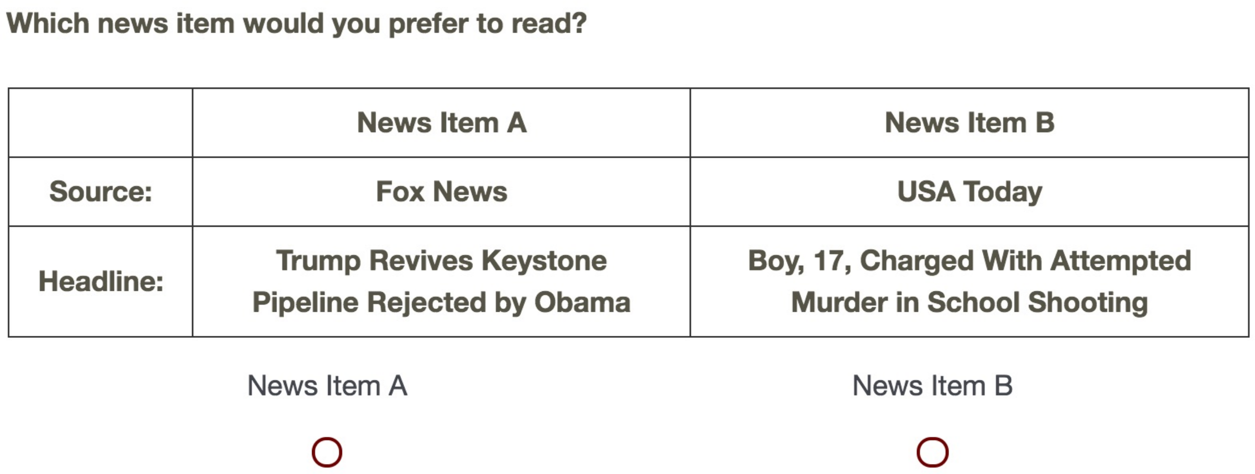
Sniderman, Paul M. 2011. "The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, Jame H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press, 102-115.

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), 273-295.

Zizzo, Daniel John. 2010. "Experimenter demand effects in economic experiments." *Experimental Economics*, 13(1): 75-98.

Appendix

Figure 2: A sample news selection task.



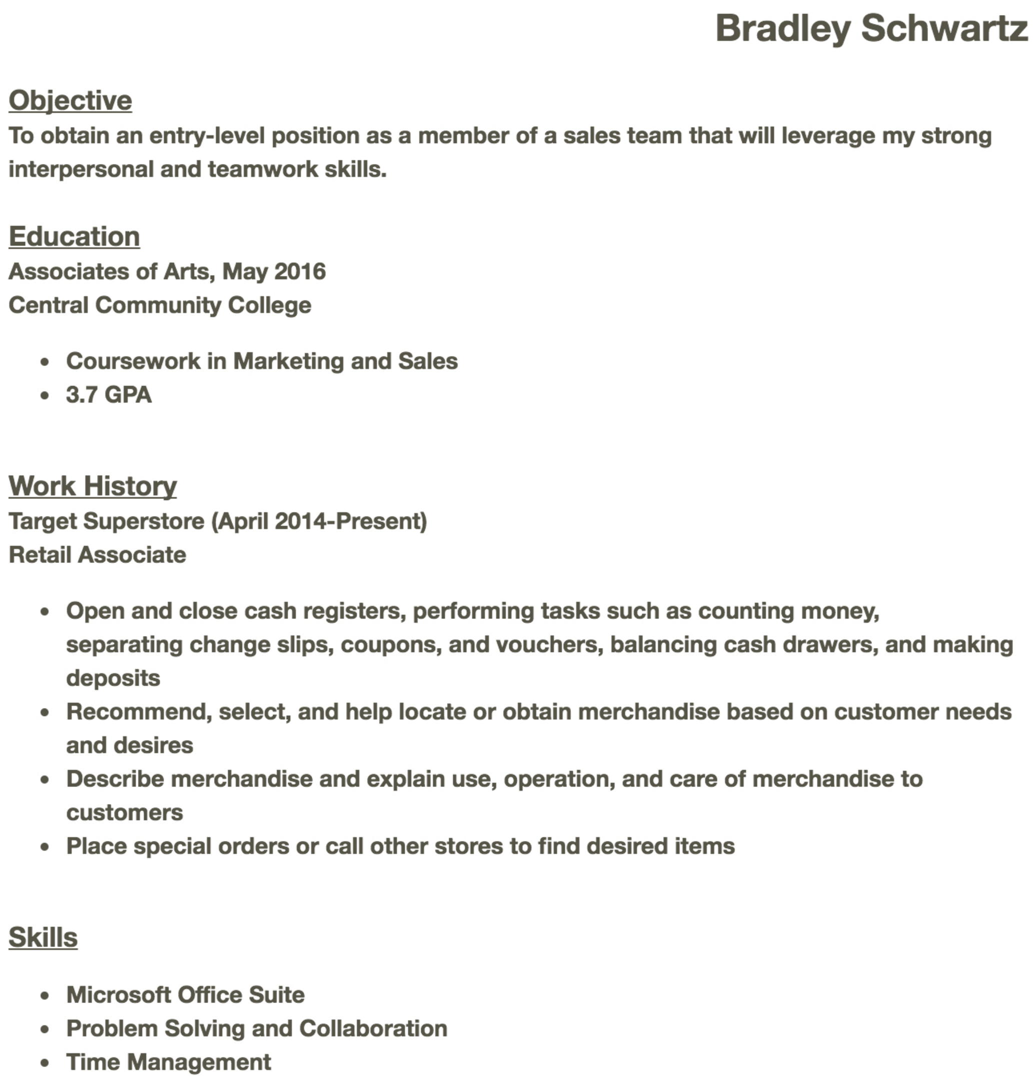
Which news item would you prefer to read?

	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 3: 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

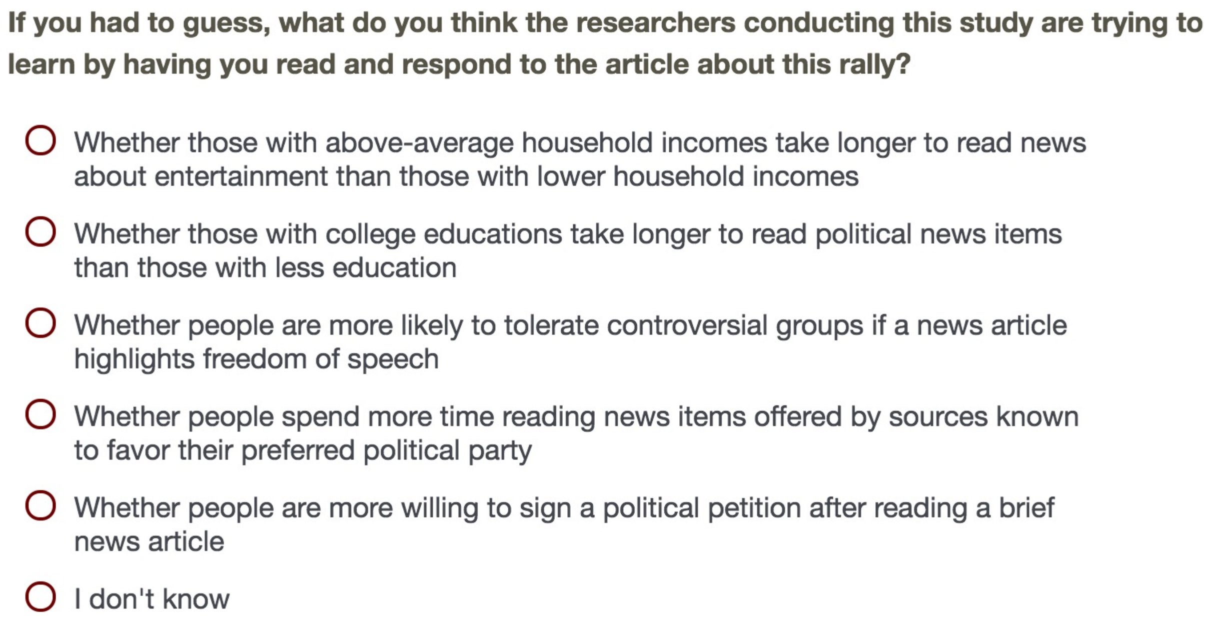
Target Superstore (April 2014-Present)
Retail Associate

- 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 4: The multiple choice question given to respondents after participating in the free speech framing experiment in Study 1.



If you had to guess, what do you think the researchers conducting this study are trying to learn by having you read and respond to the article about this rally?

○ Whether those with above-average household incomes take longer to read news about entertainment than those with lower household incomes

○ Whether those with college educations take longer to read political news items than those with less education

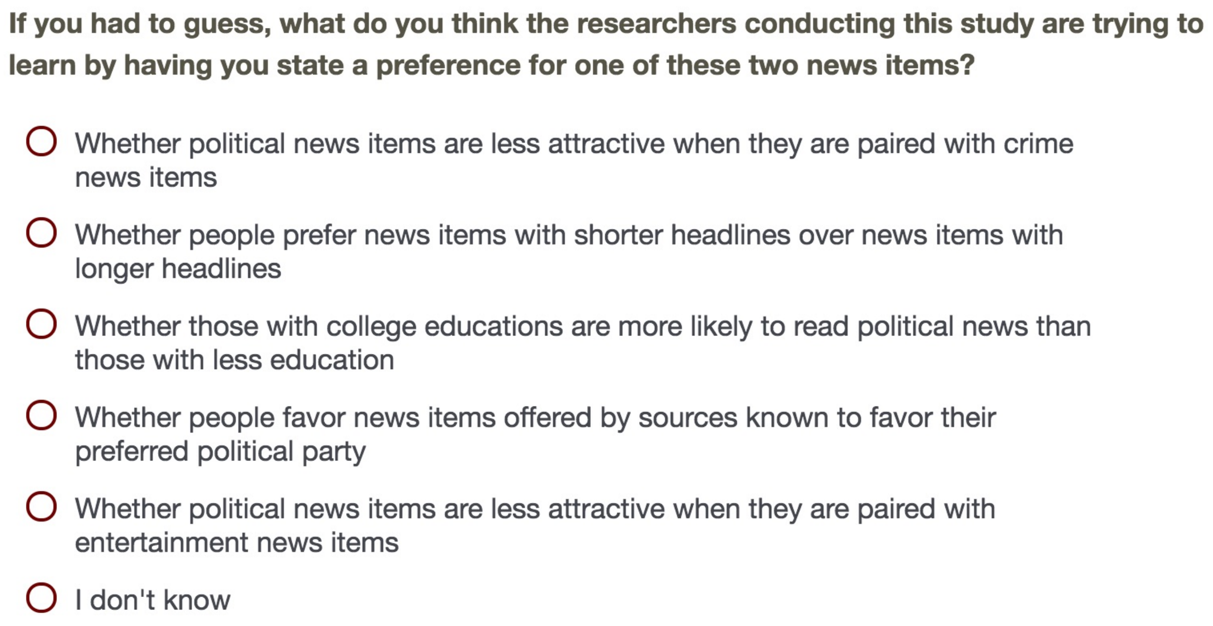
○ Whether people are more likely to tolerate controversial groups if a news article highlights freedom of speech

○ Whether people spend more time reading news items offered by sources known to favor their preferred political party

○ Whether people are more willing to sign a political petition after reading a brief news article

○ I don't know

Figure 5: The multiple choice question given to respondents after participating in the partisan selective exposure experiment in Study 1.



If you had to guess, what do you think the researchers conducting this study are trying to learn by having you state a preference for one of these two news items?

○ Whether political news items are less attractive when they are paired with crime news items

○ Whether people prefer news items with shorter headlines over news items with longer headlines

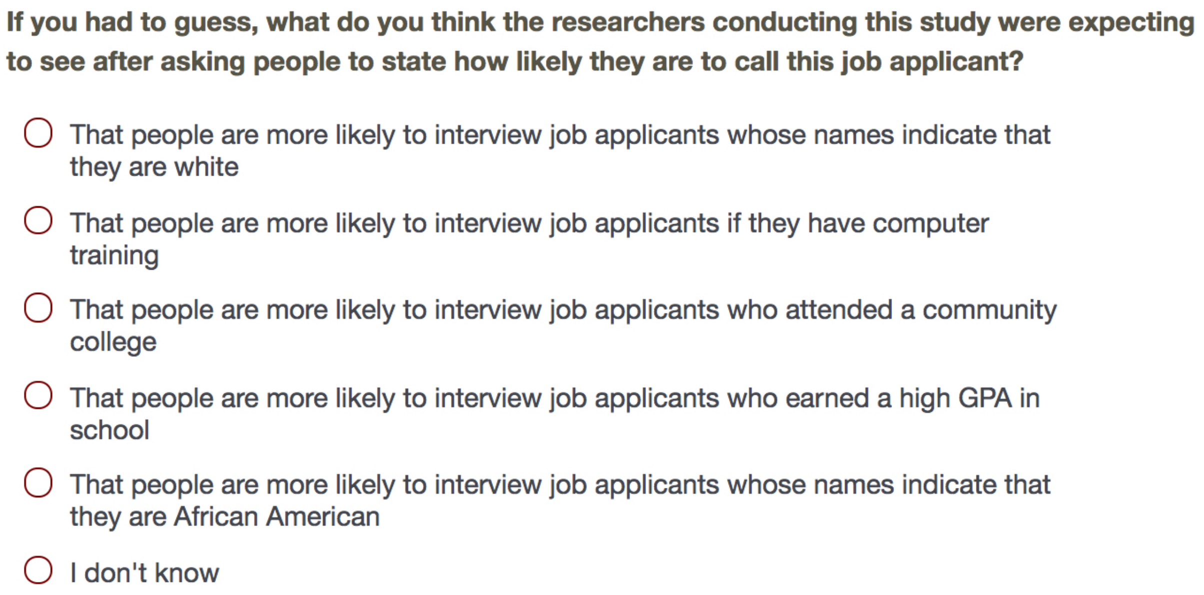
○ Whether those with college educations are more likely to read political news than those with less education

○ Whether people favor news items offered by sources known to favor their preferred political party

○ Whether political news items are less attractive when they are paired with entertainment news items

○ I don't know

Figure 6: The multiple choice question given to respondents after participating in the free speech resumé experiment in Study 2.



If you had to guess, what do you think the researchers conducting this study were expecting to see after asking people to state how likely they are to call this job applicant?

○ That people are more likely to interview job applicants whose names indicate that they are white

○ That people are more likely to interview job applicants if they have computer training

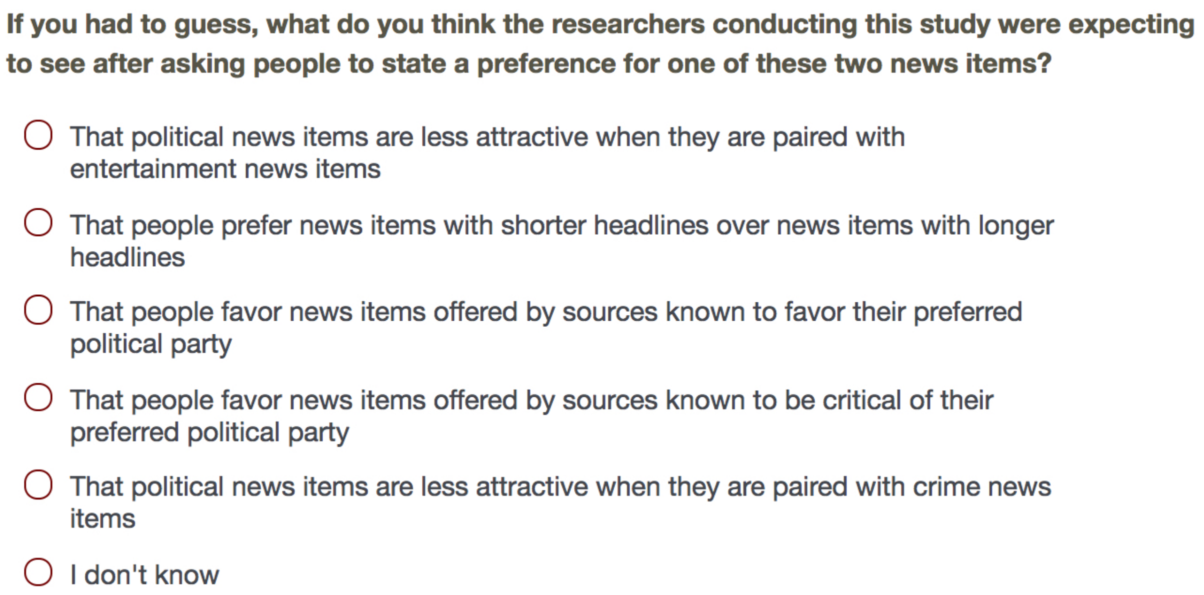
○ That people are more likely to interview job applicants who attended a community college

○ That people are more likely to interview job applicants who earned a high GPA in school

○ That people are more likely to interview job applicants whose names indicate that they are African American

○ I don't know

Figure 7: The multiple choice question given to respondents after participating in the partisan selective exposure experiment in Study 2.



If you had to guess, what do you think the researchers conducting this study were expecting to see after asking people to state a preference for one of these two news items?

○ That political news items are less attractive when they are paired with entertainment news items

○ That people prefer news items with shorter headlines over news items with longer headlines

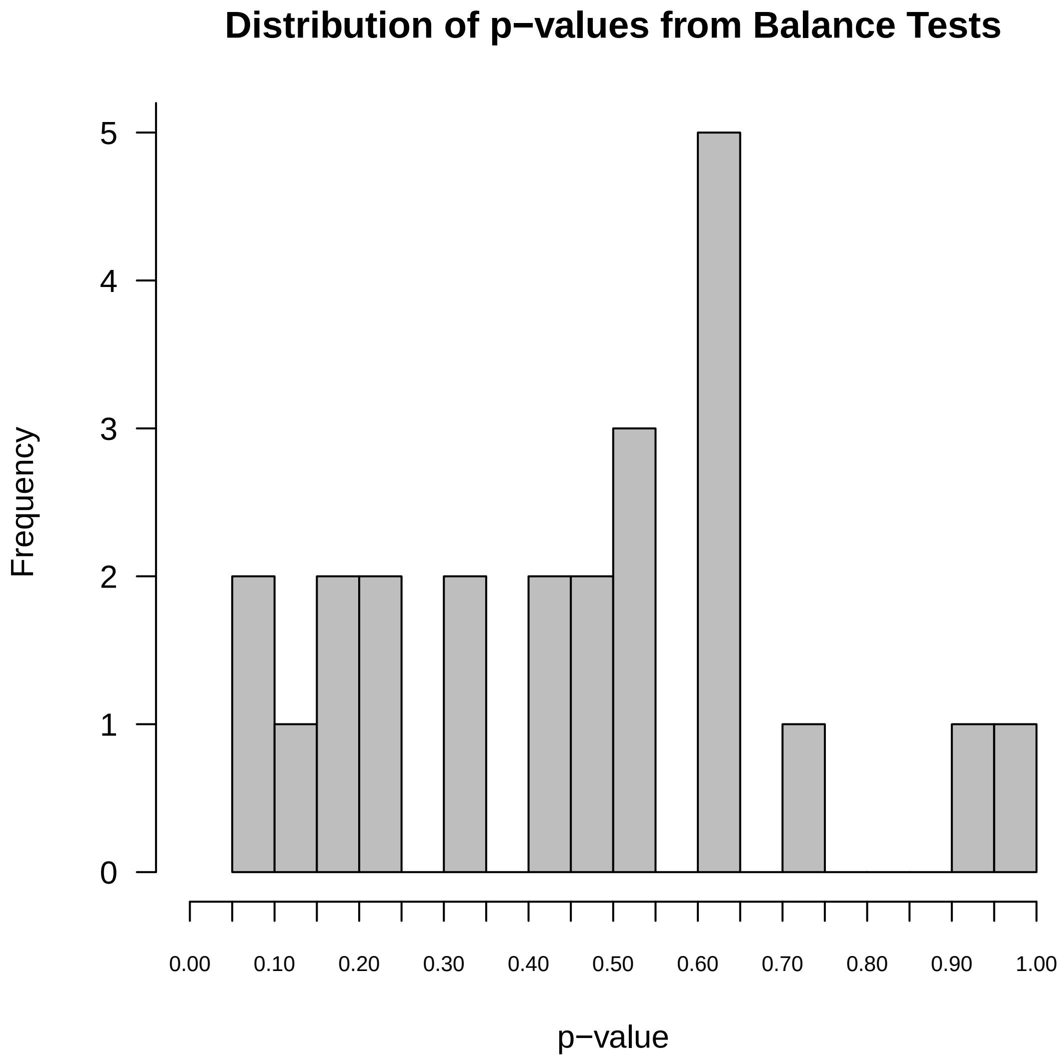
○ That people favor news items offered by sources known to favor their preferred political party

○ That people favor news items offered by sources known to be critical of their preferred political party

○ That political news items are less attractive when they are paired with crime news items

○ I don't know

Figure 8: 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, ethnicity, self-reported turnout in 2016, an indicator of having a BA, 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. As the figure shows, none of these tests generated p-values ≤ 0.05, which is consistent with good balance.


Distribution of p−values from Balance Tests
Frequency
p−value