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."

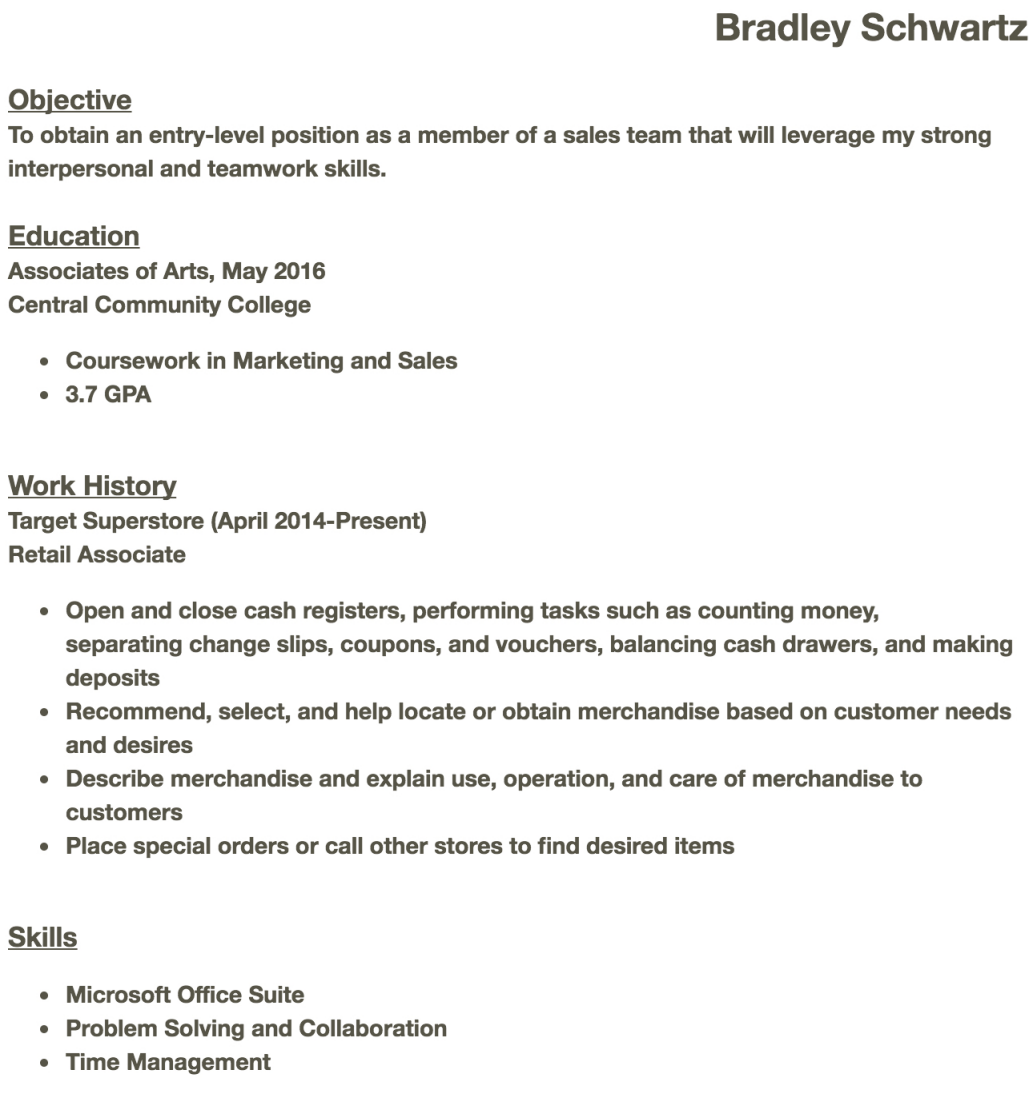
Figure A2: A sample news selection task.

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.

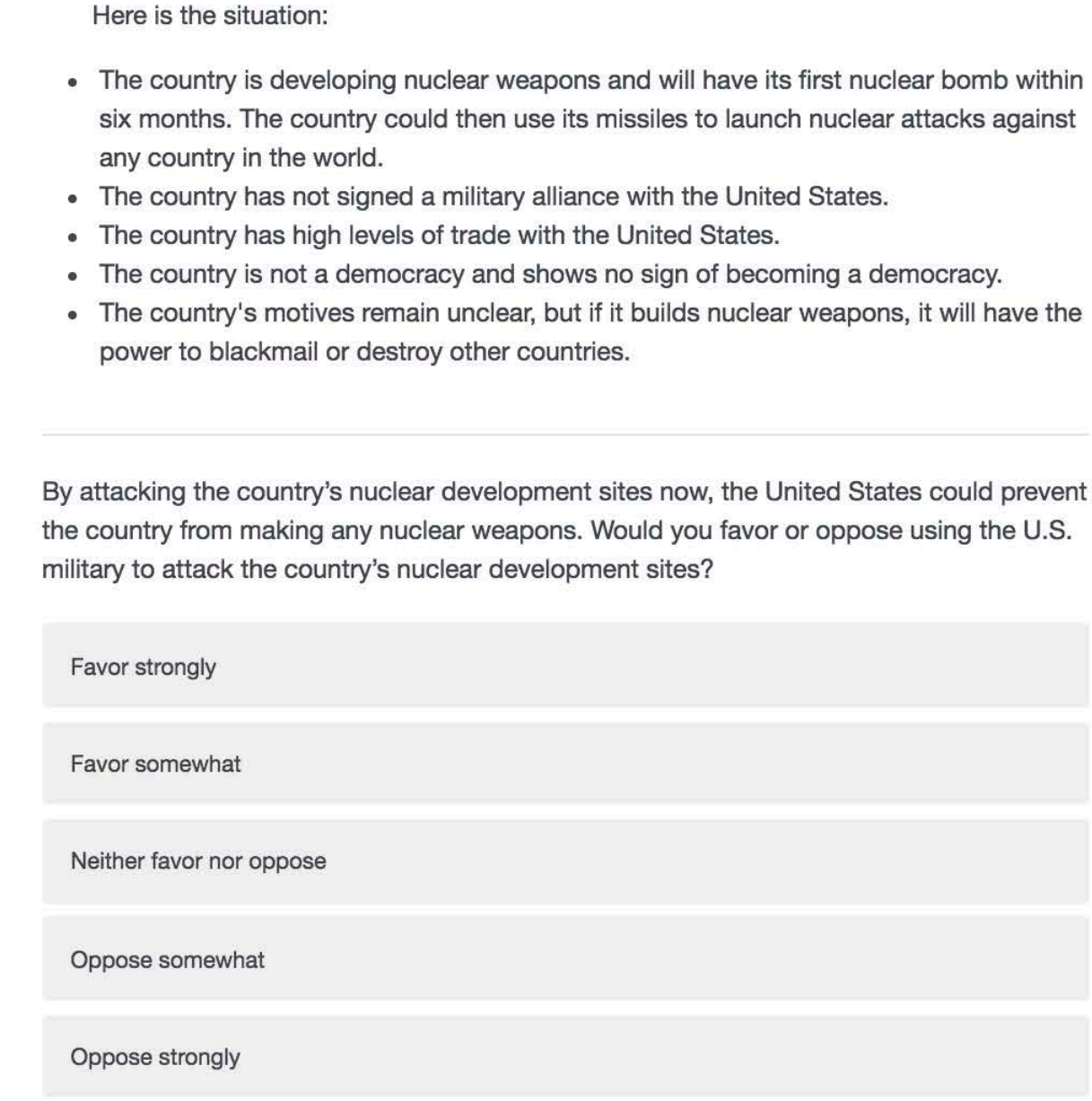
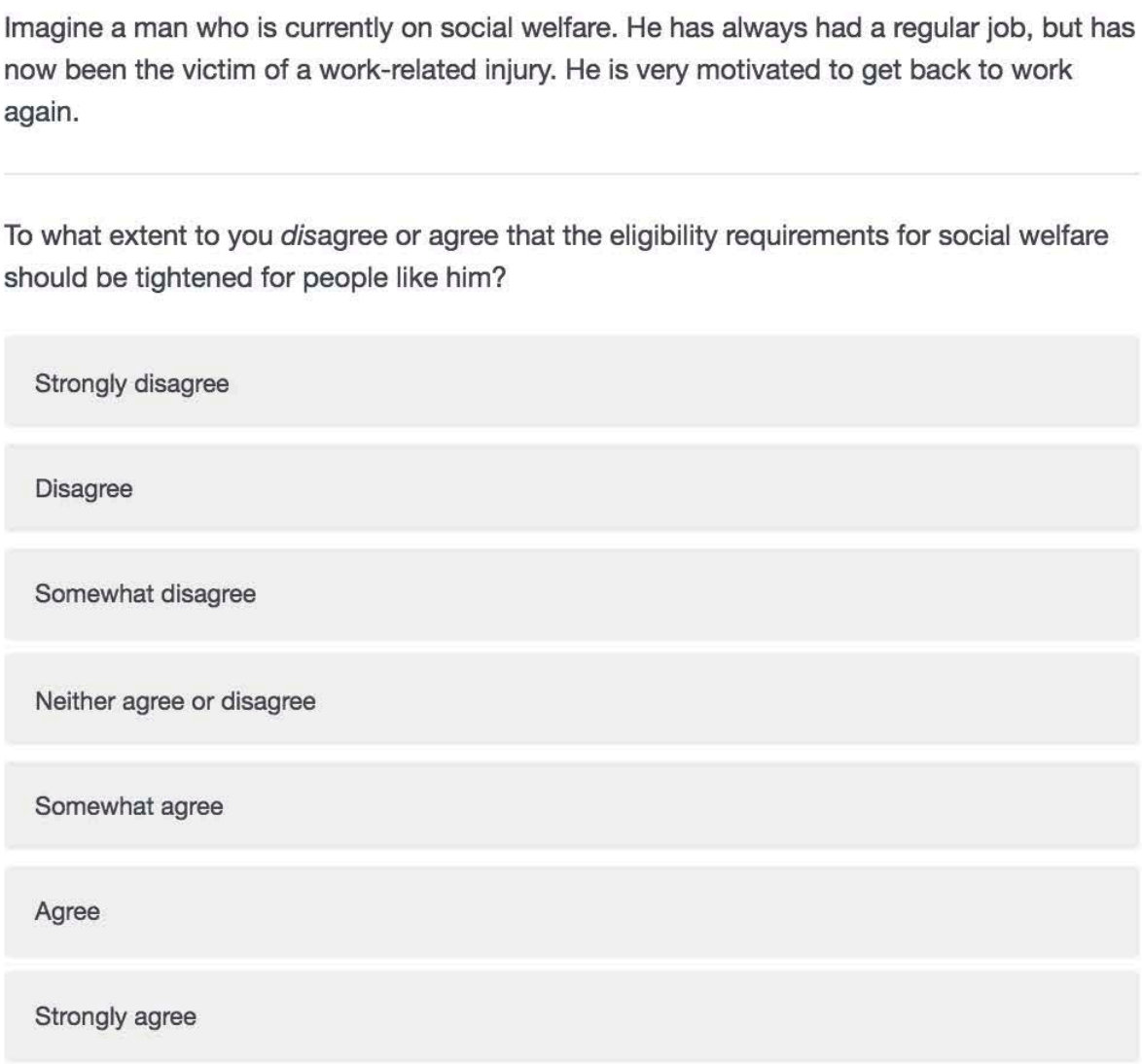


Figure A5: One version of the welfare experiment, in which the welfare recipient is described as unlucky.



Manipulation Checks

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

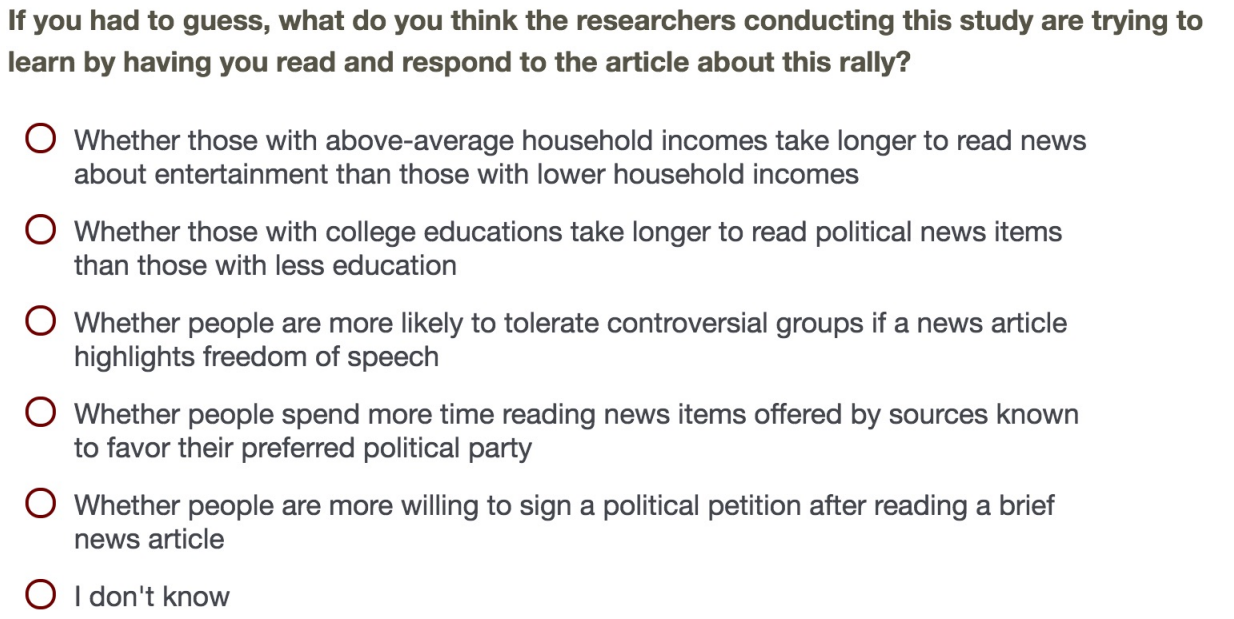


Figure A7: The multiple choice question given to respondents after participating in the partisan selective exposure experiment in Survey 1.

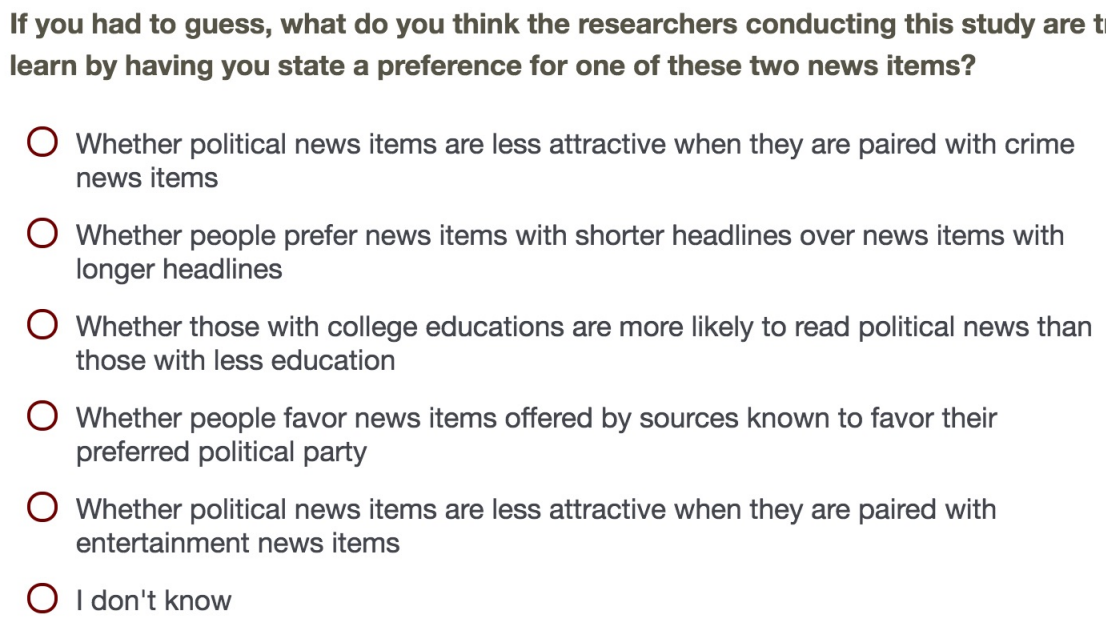


Figure A8: The multiple choice question given to respondents after participating in the free speech resumé experiment in Survey 2.

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

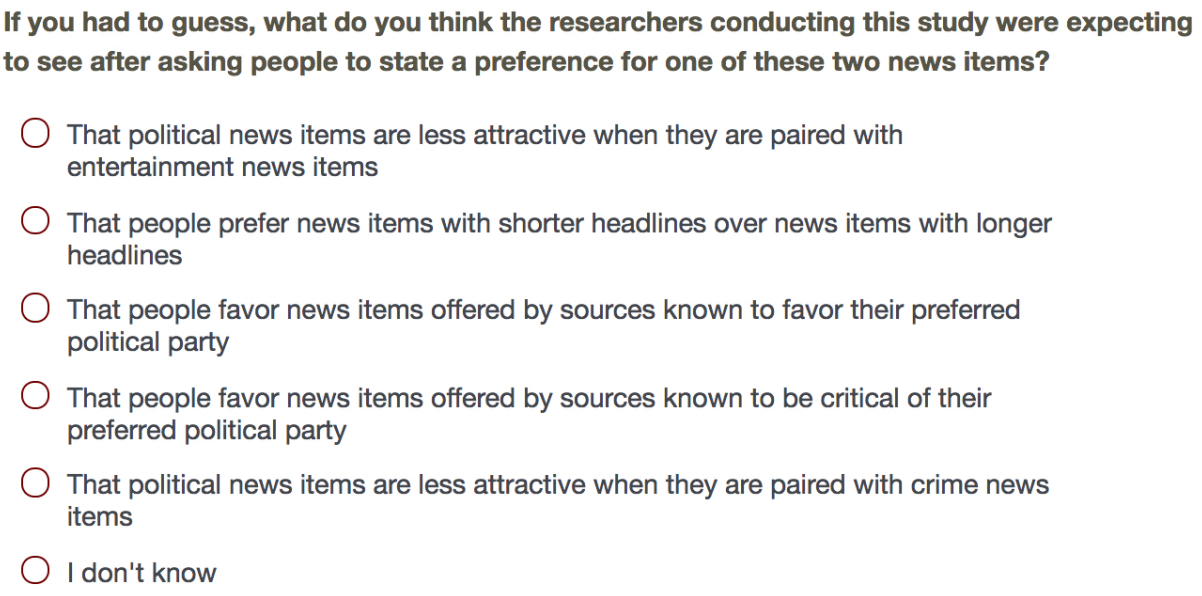


Figure A10: The multiple choice question given to respondents after participating in the democratic peace experiment in Survey 3.

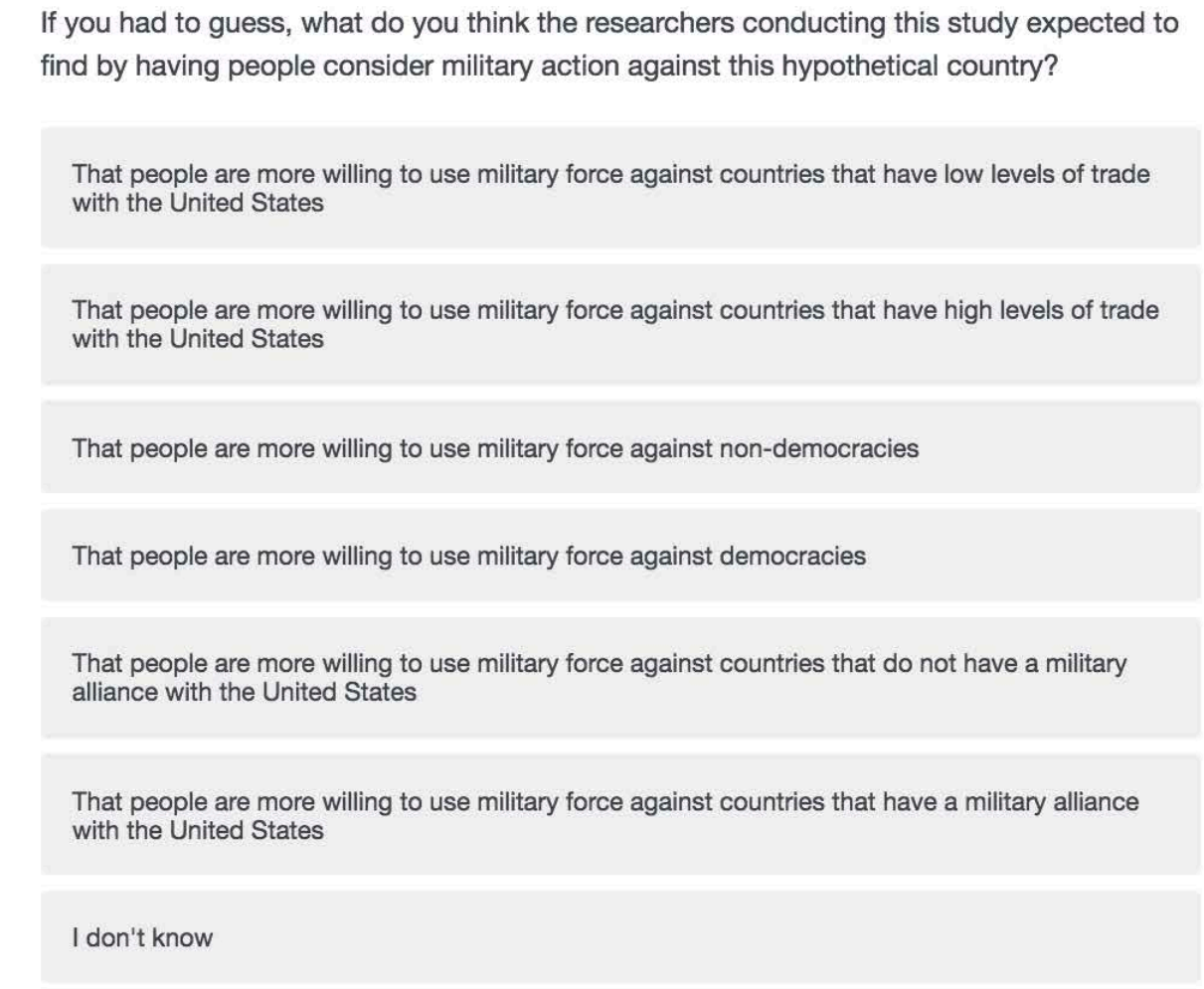
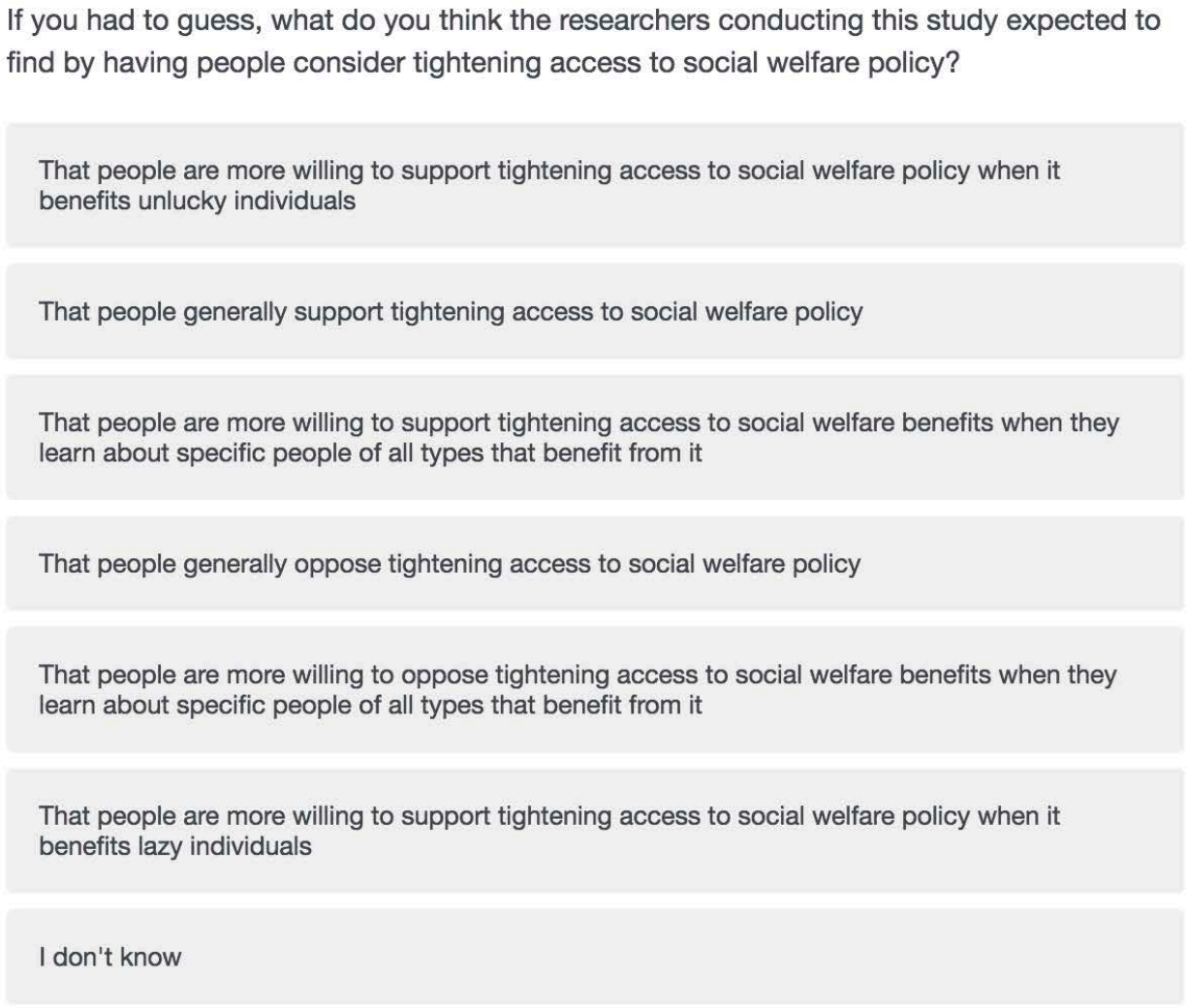


Figure A11: The multiple choice question given to respondents after participating in the welfare experiment in Survey 3.



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

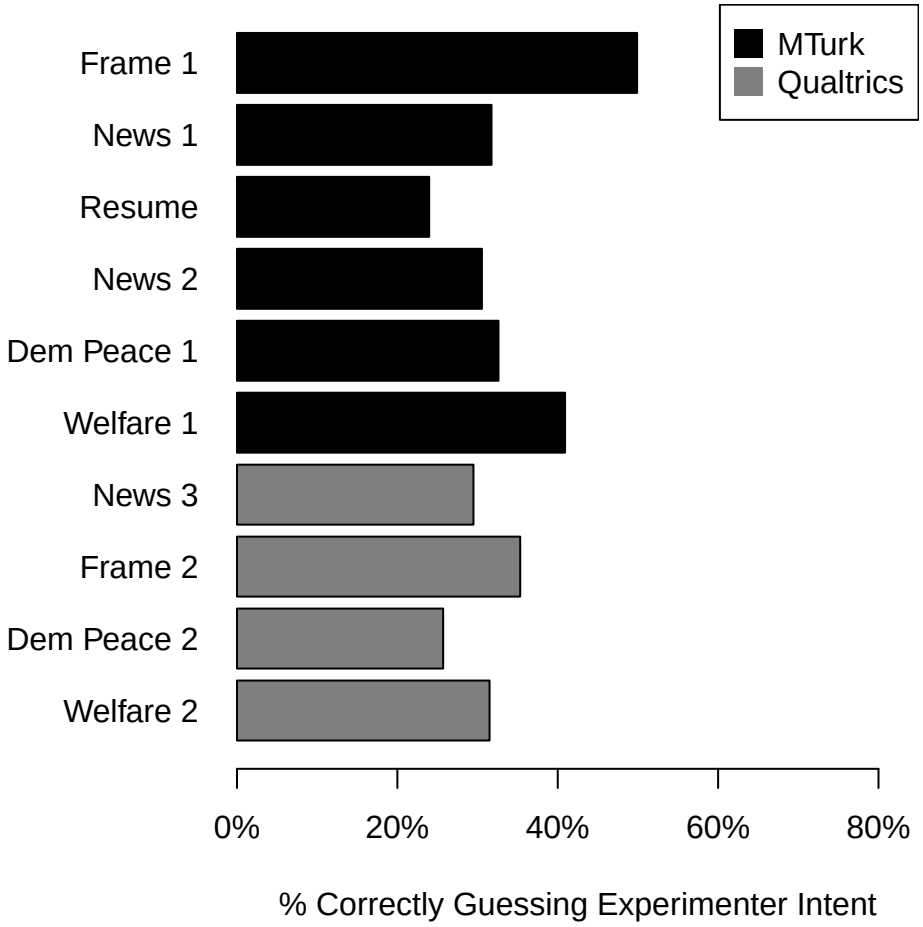
The table below displays descriptive statistics across these samples. The first three studies were convenience samples from Mechanical Turk. The last two studies were samples from an online survey vendor, Qualtrics. The values for age and income represent the mean values for each survey. The other categories 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.

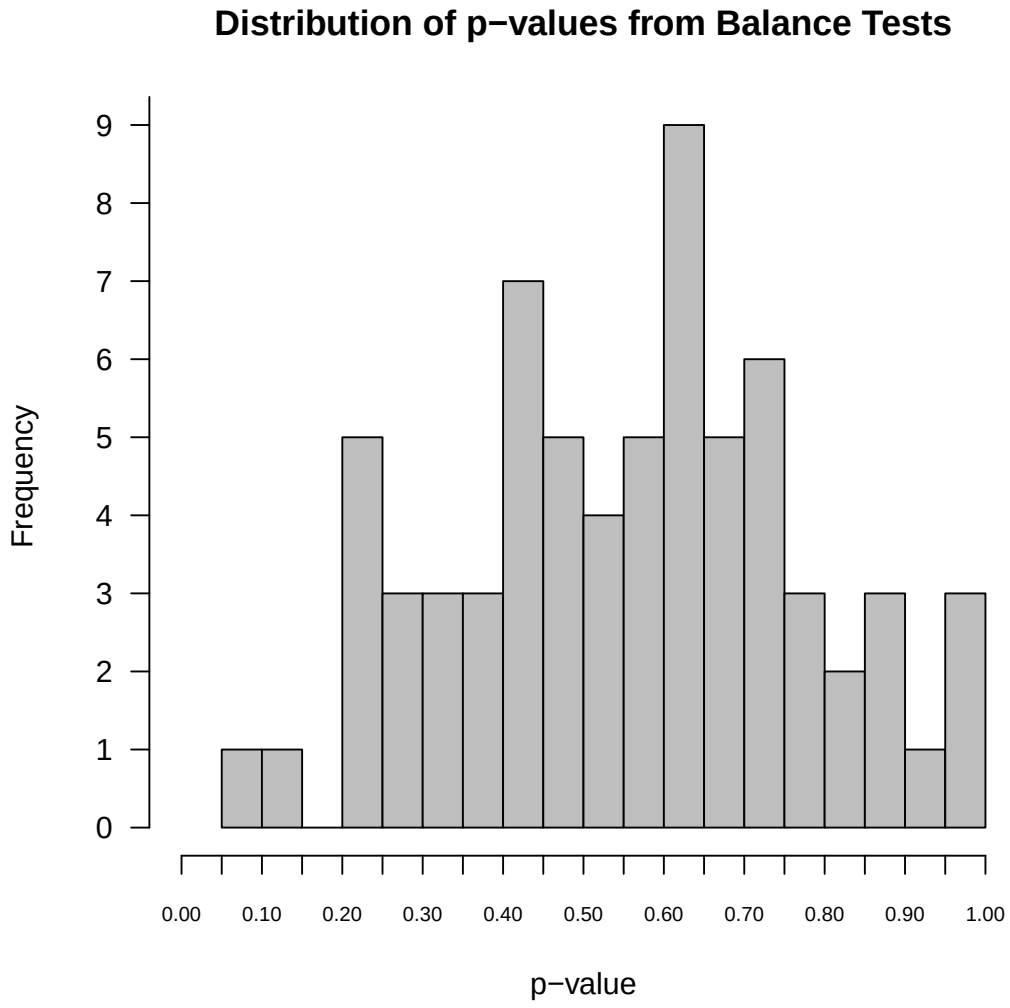
Figure B1: **Rate of Guessing Intent In Baseline Conditions.** The figure displays the proportion of respondents who guessed each experiment's intent in the baseline conditions that did not provide any additional information about the researcher's hypothesis.



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.

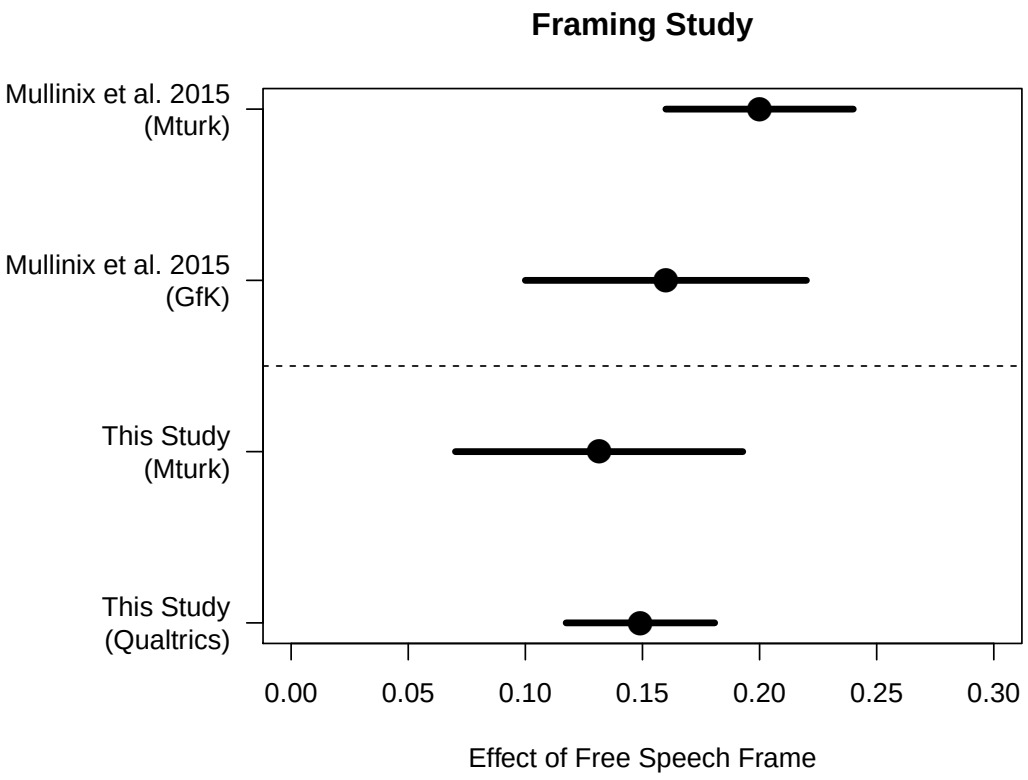


Replication of Original Experiments

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

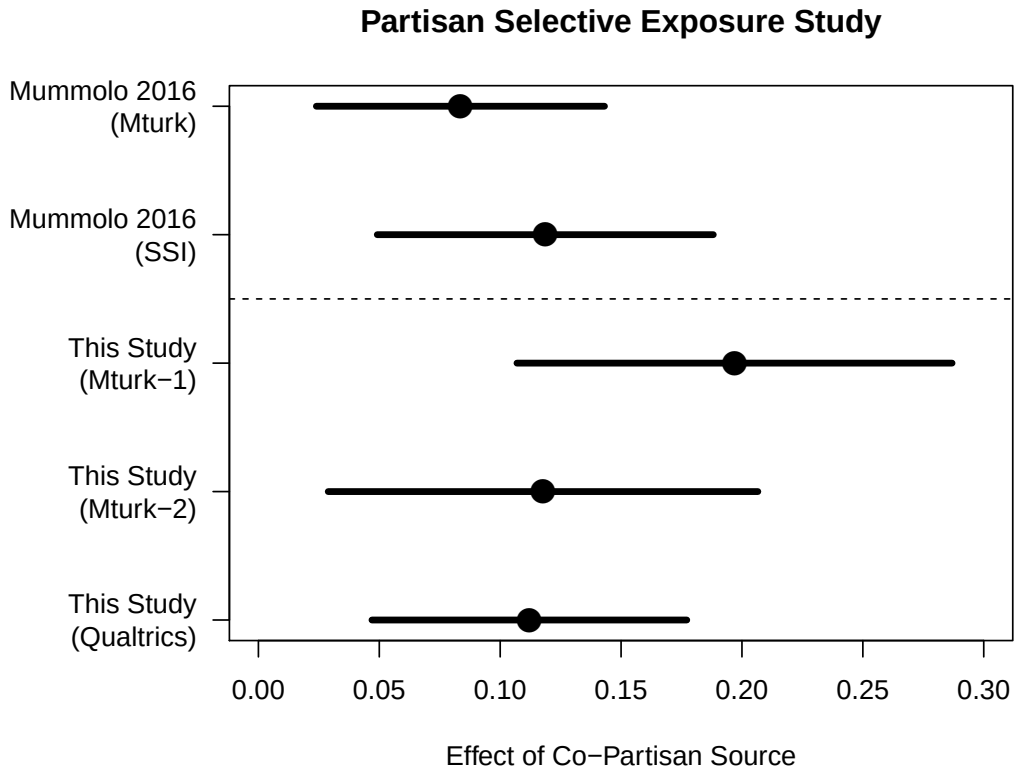


We next examine the effect of a co-partisan news source on the probability an individual chooses to read an article from that source. This experiment was substantively inspired by Iyengar and Hahn (2009) and based on the conjoint design in Mummolo (2016), which conducts the experiment on samples from 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.

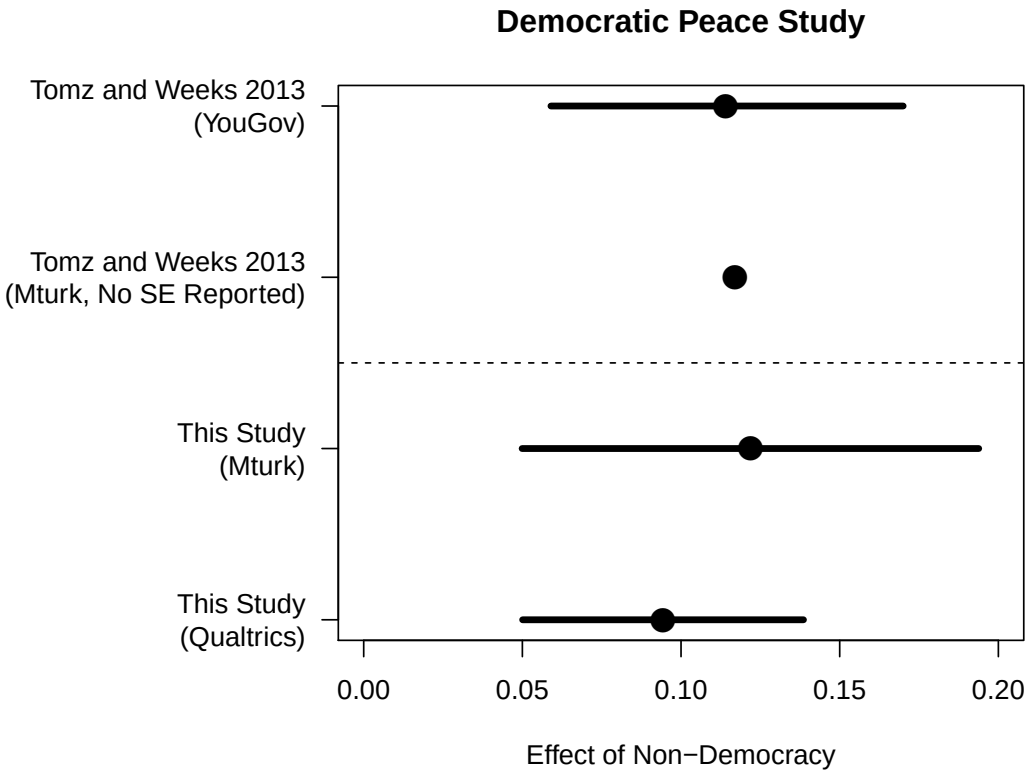
Figure B4



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.

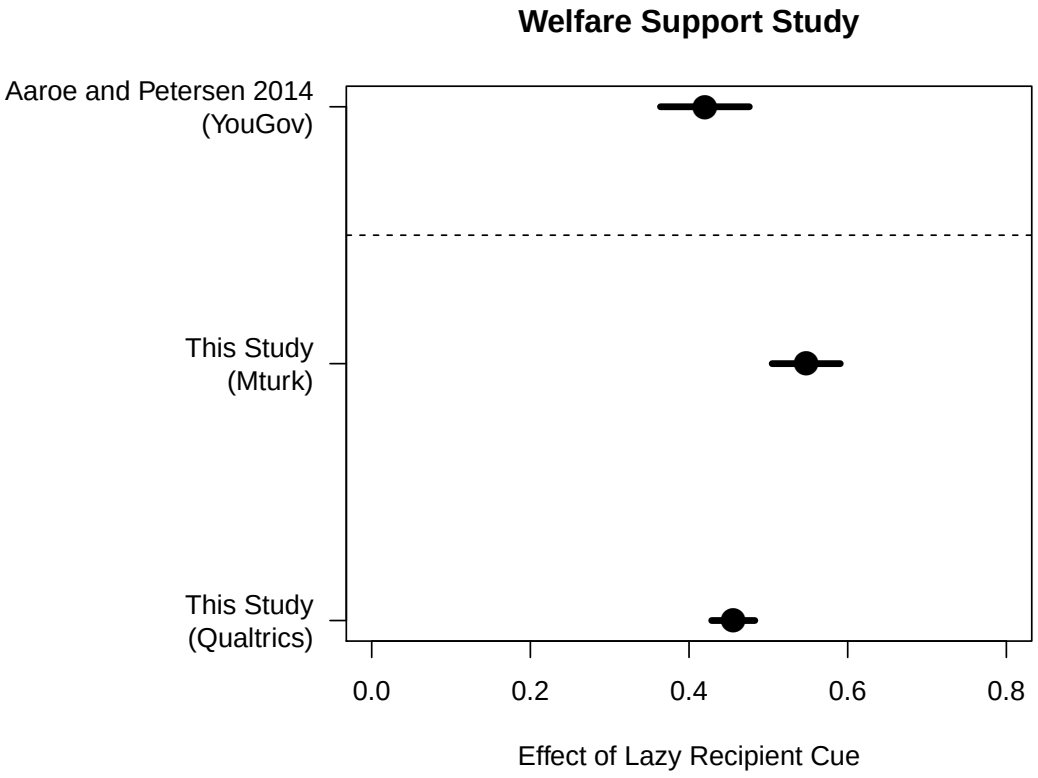
Figure B5



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.

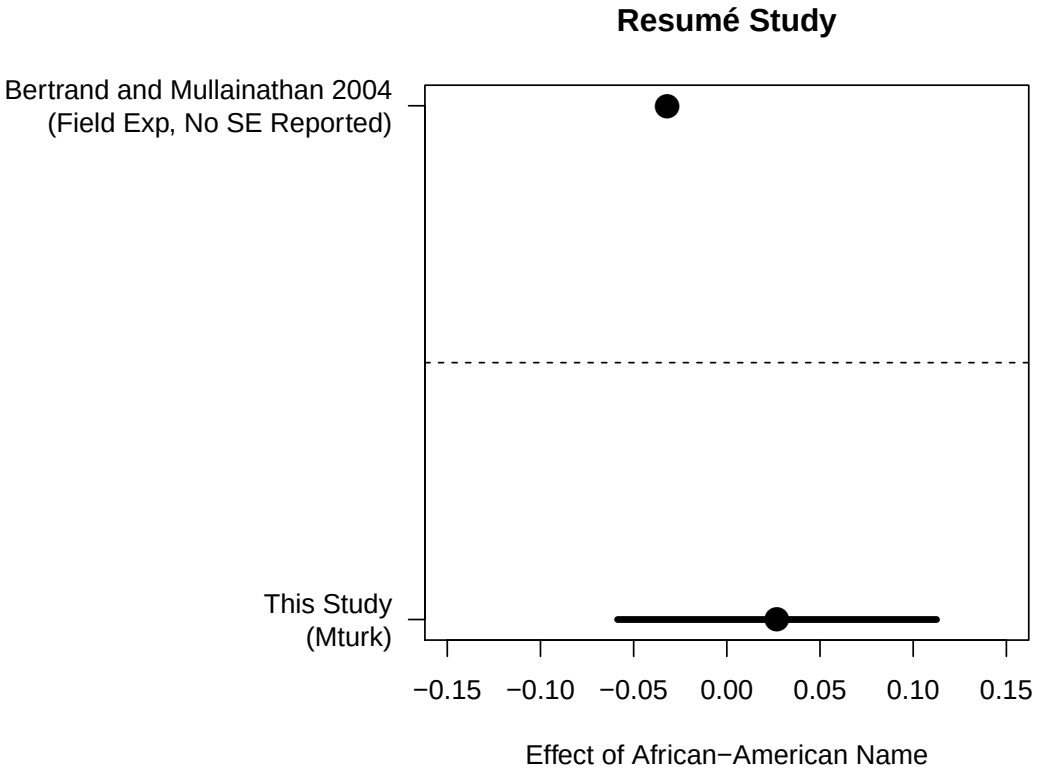
Figure B6



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.