Online Appendix for: The Generalizability of Online Experiments Conducted During The COVID-19 Pandemic

Kyle Peyton, Gregory A. Huber, and Alexander Coppock

November 28, 2020

Contents

List of Figures

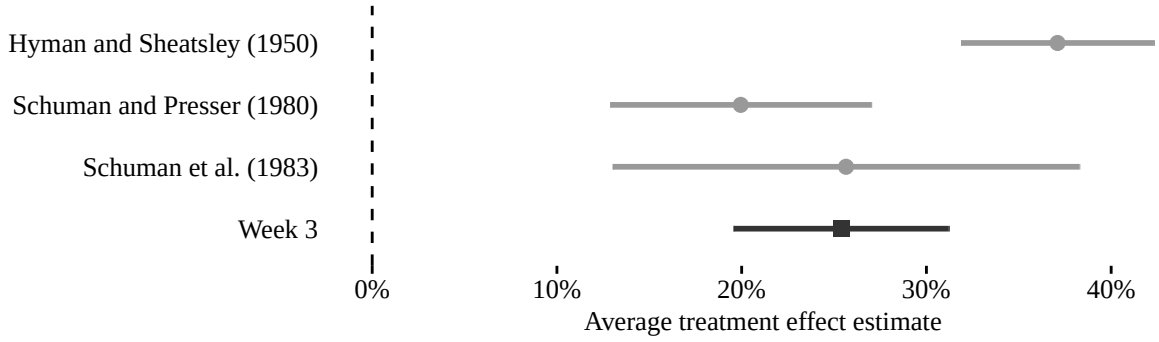
A Individual Studies

A.1 Russian reporters and American news

This classic study by Hyman and Sheatsley (1950) shows that American subjects express more tolerance for Russian journalists to "come in here and send back to their papers the news as they see it" if they are first asked whether *American* journalists should be allowed to operate in Russia. The operating principle seems to be one of reciprocal fairness – after affirming that American journalists should be allowed to work in Russia, subjects appear to feel constrained to allow Russian journalists to work in America.

The original effect estimate is a 36.6 percentage point increase in support for Russian journalists. Our effect estimate of 25.5 points is smaller, but clearly in line with the two earlier replications reported in Schuman and Presser (1996). The baseline levels of support for Russian journalists in the control condition among Americans in 1950 (36%) and 1983 (44%) are quite similar to COVID-era Lucid respondents (45%).

Figure A.1: Effect of question ordering on support for Russian journalists in U.S.



A.2 Gain versus loss framing

In this classic framing experiment by Tversky and Kahneman (1981, Study 1), undergraduates were instructed to imagine the U.S. was preparing for the outbreak of an unusual "Asian disease", which was expected to kill 600 people. In the "gain" framing condition,

participants were asked to select between two courses of action to combat the disease: if Program A is adopted, 200 people will be saved; if Program B is adopted, there is a 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. In the "loss" framing condition, participants were asked to select between two different formulations: if Program A is adopted, 400 people will die; if Program B is adopted, there is a 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

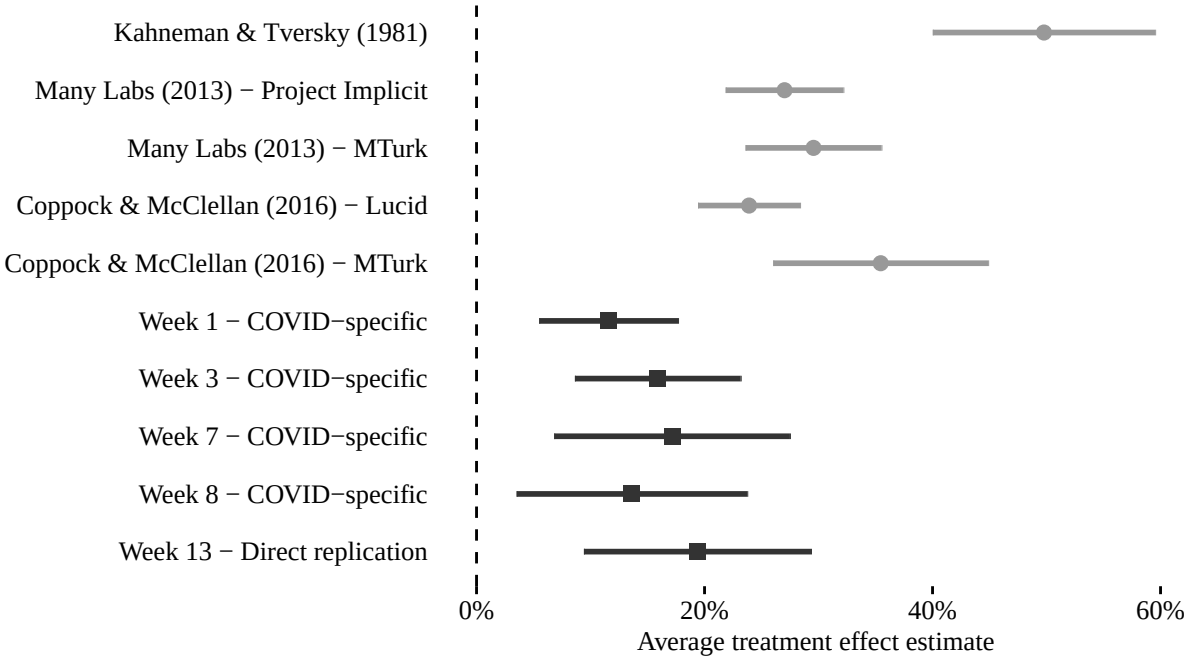
In both framing conditions, the expected number of deaths is 400 for both Program A and Program B. In the original study, 72% selected Program A in the gain frame, whereas 22% selected Program A in the loss frame, for an estimated treatment effect of 50 percentage points. According to Tversky and Kahneman (1981), the observed preference reversal illustrates how individuals' choices involving gains are risk averse whereas choices involving losses are risk seeking.

This experiment has been widely replicated across time in both student samples and online samples. Figure A.2 plots estimated treatment effects from the original study (student sample) alongside those obtained from pre-COVID studies (online samples) in the Many Labs replication project (Klein et al., 2014) in 2013, in 2013 on MTurk (Berinsky, Huber and Lenz, 2012), in 2016 on Lucid (Coppock and McClellan, 2018), and in five of our replications. The first four of our replications are COVID-specific versions of the original. Participants were instead asked to imagine that "the Mayor of a U.S. city is preparing for another outbreak of the novel coronavirus in the Spring of 2021, which is expected to kill 600 city residents." The fifth replication is a direct replication of the pre-COVID experiments using the original wording.

The summary effect size for our five replications is 0.15 (SE = 0.02, $P < 0.01$), approximately 56% the size of the summary effect size for the four pre-COVID experiments (summary effect size: 0.27, SE = 0.01, $P < 0.01$). Although the replications estimates are, on average, smaller than those from pre-COVID experiments, all replication estimates are

statistically distinguishable from zero, and in the expected direction. We therefore conclude that the replications were successful, regardless of whether COVID-specific language was used in the scenario description.

Figure A.2: Effect of gain vs. loss frame in "Asian disease" problem



A.3 Effect of framing on decision making

In another classic framing experiment by Tversky and Kahneman (1981, Study 10), undergraduates were instructed to imagine a scenario in which they were buying two items, one for $15 and another for $125. Participants in the "cheap" condition read the following prompt, with those in the "expensive" condition seeing the prices in parentheses: "Imagine that you are about to purchase a jacket for $125 ($15), and a calculator for $15 ($125). The salesman informs you that the calculator you wish to buy is on sale for $10 ($120) at the other branch of the store, located 20 minutes drive away. Would you make the trip to the other store?"

Although the total cost of both items was $140 in each condition, with a potential of $5 in savings for traveling, 68% of participants said they would travel when they could save $5

on the $15 item, whereas 29% said they would travel when they could save $5 on the $125 item. According to Tversky and Kahneman (1981), this difference of 39 percentage points illustrates how individuals' assess the potential gains and losses of outcomes in relative, rather than absolute, terms. When paying $15 for an item, a $5 discount seems substantial, whereas a $5 discount on a $125 item seems negligible.

This experiment has been replicated numerous times in both student and online samples. We use a slightly modified version of the original study from Klein et al. (2018) as our pre-COVID benchmark for online samples. In this study, participants were presented with the following prompt, with those in the "expensive" condition seeing the prices in parentheses: "Imagine that you are about to purchase a ceramic vase for $250 ($30), and a wall hanging for $30 ($250). The salesman informs you that the wall hanging you wish to buy is on sale for $20 ($240) at the other branch of the store, located 20 minutes drive away. Would you make the trip to the other store?"

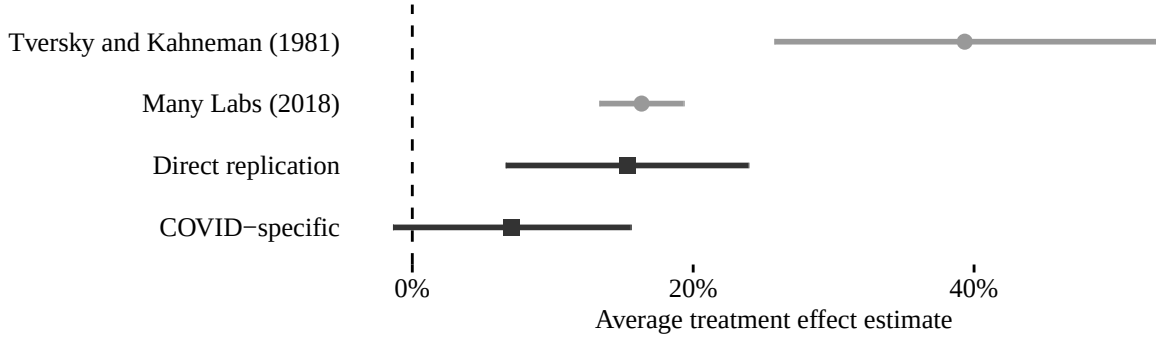
In the replication by Klein et al. (2018), 49% of participants said they would travel to save $10 on the "cheap" wall-hanging whereas 32% said they would travel to save $10 on the "expensive" wall-hanging. In our Week 7 replication study, half the participants were assigned to an experiment using this same scenario (wall hanging and ceramic vase). The other half were assigned to a COVID-specific scenario, where "ceramic vase" was replaced with "a box of Clorox disinfecting wipes," and "wall hanging" was replaced with "a box of N-95 respirator masks". See Figures C.1-C.4 for a full description of each condition.

Figure A.3 plots estimated treatment effects from the original study (student sample) alongside the pre-COVID benchmark study (online sample) and our replications. The estimated effect in the direct replication of 15 percentage points was indistinguishable from the pre-COVID benchmark (16 percentage points). For the COVID-specific experiment, the estimated treatment effect of 7 percentage points was indistinguishable from zero, and smaller than both the pre-COVID benchmark (difference of 9 percentage points, SE = 0.05,

$P = 0.02$) and the direct replication (difference of 8 percentage points, SE = 0.06, $P = 0.09$).

Although the replications estimates are, on average, smaller than those from the pre-COVID benchmark, the direct replication closely approximates the 2018 study and all replication estimates are in the expected direction. We also note that the estimated effect from the COVID-specific replication is smaller but statistically indistinguishable from the direct replication. This raises the possibility that the COVID-specific language in the replication decreased the power of the framing effect. We nevertheless conclude that the replications were successful.

Figure A.3: Effect of "Cheap" vs. "Expensive" frame on decision to travel



A.4 Welfare versus aid to the poor

The large effect of describing government assistance as "welfare" rather than "aid to the poor" is one of the most robust experimental findings in political science. In the original experiment (Smith, 1987), a sample of U.S. adults from the General Social Survey (GSS) were asked by telephone survey whether they believed there was "too much", "about the right amount", or "too little" spending across multiple issues, with social welfare benefits being described as "assistance for the poor" in the treatment arm and "welfare" in the control arm. This experiment has been replicated biannually on GSS from 2002 to 2018. Respondents are consistently more supportive of government spending on "assistance for the

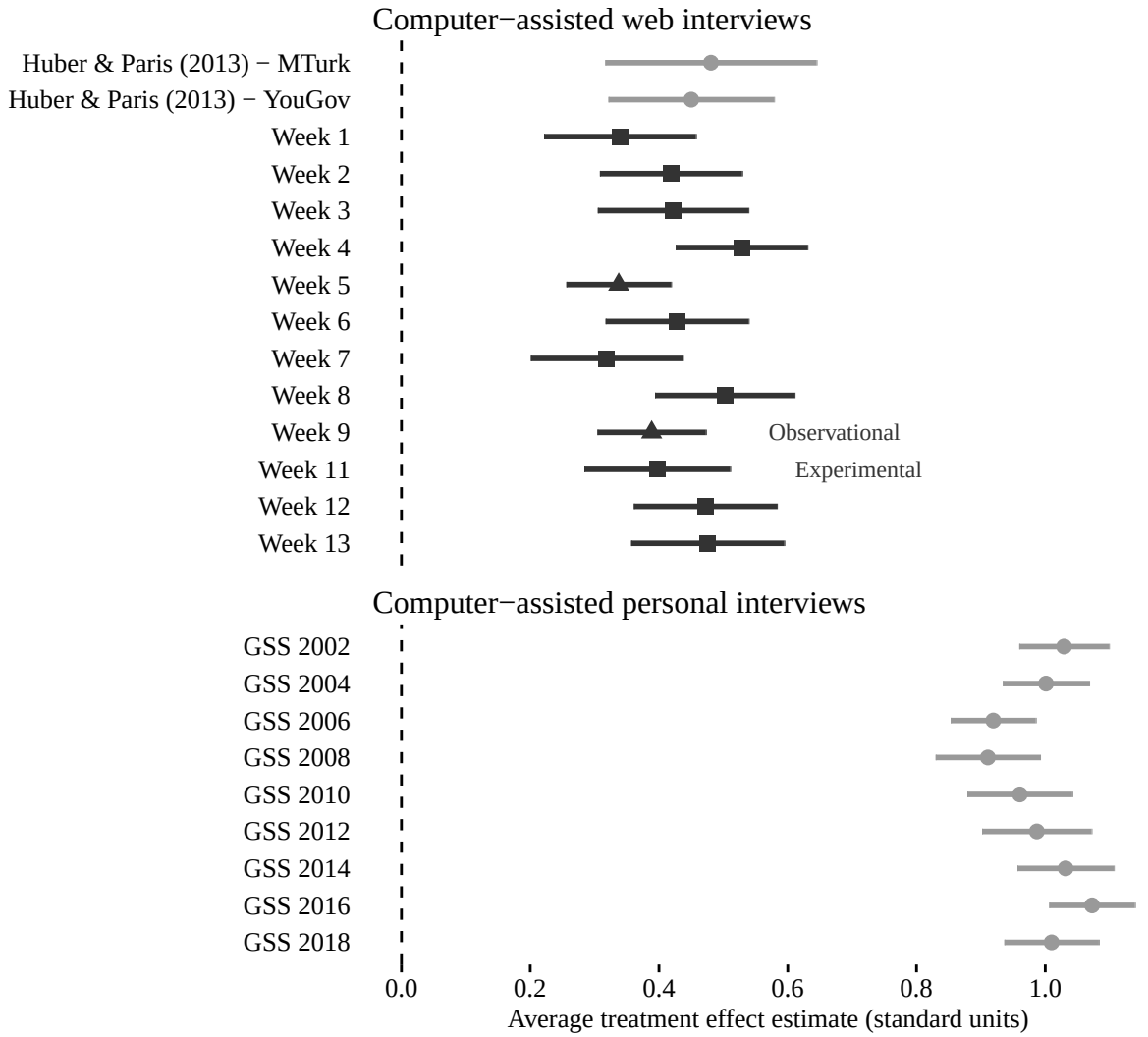
poor" than "welfare". [1]

Figure A.4 plots estimated treatment effects from nine replications of the original experiment on GSS respondents using computer-assisted personal interviews (CAPI) alongside those obtained from our twelve replications, and one pre-COVID replication, using computer-assisted web interviews (CAWI). Two of our replications (Week 9 and Week 5) were within-subject experiments that asked respondents both spending questions in randomized order. Following prior replications on online samples (e.g. Huber and Paris, 2013), responses are coded as -1 ("too much"), 0 ("about the right amount"), and 1 ("too little"), so that negative values indicate less support for spending. All replication estimates are statistically distinguishable from zero and are in the expected direction. [2] The summary effect size for the experimental estimates from CAWI surveys is -0.47 (SE = 0.02, $P < 0.01$), approximately 39% the size of the summary effect size for the GSS estimates from CAPI surveys (-1.22, SE = 0.02, $P < 0.01$). Although the replications estimates are, on average, smaller than the pre-COVID benchmark, all replication estimates are in the expected direction. Finally, we note that significant differences between estimates obtained from CAWI and CAPI surveys was a feature of experimental research prior to COVID.

[1] See Huber and Paris (2013) for evidence that individuals believe these labels describe different social programs.

[2] Interestingly, the non-experimental within-subjects estimates are solidly in line with the experimental estimates, suggesting that subjects feel no pressure to keep their support for welfare consistent with their support for aid to the poor. This pattern contrasts strongly with the evident pressure for consistency in the Russian journalists experiment. For further discussion on tradeoffs in the choice of within versus between subjects designs, see Clifford, Sheagley and Piston (2020).

Figure A.4: Effect of "Aid to Poor" vs. "Welfare" on support for government spending



Notes: Starting in 2002, the GSS replaced "assistance to the poor" with "assistance for the poor." Week 13 is a direct replication of Huber and Paris (2013) using GSS question wording. The other replications use the ANES question wording, which asks whether respondents think spending should be "increased" (coded 1), "kept the same" (coded 0), or "decreased" (coded -1).

A.5 Gain versus loss framing with party endorsements

Druckman (2001) extended the "Asian disease" protocol to explicitly incorporate political considerations. In his original study, a sample of undergraduates were randomly assigned to the classic version of the study or a modified version that randomly assigned party endorse-

ments instead of "Program A" and "Program B". In the "gain" framing condition, participants were asked to select between two courses of action, with one of three randomly assigned labels: If [Program A, the Democrats' Program, the Republicans' Program] is adopted, 200 people will be saved; If [Program B, the Republicans' Program, the Democrats' Program], there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved. In the "loss" framing condition, the descriptions were: If [Program A, the Democrats' Program, the Republicans' Program] is adopted, 400 people will die; If [Program B, the Republicans' Program, the Democrats' Program], there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die.

In the original study, the preference reversal effect from Tversky and Kahneman (1981) was replicated when "Program A" and "Program B" were used as labels. However, these effects were greatly attenuated (or indistinguishable from zero) when the programs were labeled with party endorsements. According to Druckman (2001), this difference illustrates how partisans' desire to choose their party's program can overwhelm preference reversals due to "pure" framing effects.

Figure A.5 plots estimated treatment effects from the original study (student sample) alongside three replications. Two of our replications (Week 7 and Week 8) are COVID-specific versions of the original where "unusual Asian disease" was replaced with "another outbreak of the novel coronavirus". The Week 13 replication is a direct replication of the original Asian disease experiment. All estimates are statistically distinguishable from zero and in the expected direction when "Program A" is used to describe the "risk-averse alternative" (e.g. save 200 people versus 400 people will die). Consistent with the original experiment, however, adding the partisan labels attenuate (or eliminate) preferences reversals among partisans: among Democrats, preference reversal effects are indistinguishable from zero when "Program A" is replaced with "Republicans' Program"; among Republicans, preference reversal effects are indistinguishable from zero when "Program A" is replaced with

"Democrats' Program".

Table A.5 provides a summary of differences between the original study and the replications. Although the replications estimates are, on average, smaller than those from the original study, all replication estimates are in the expected direction. We therefore conclude that the original study replicated.

Figure A.5: Effect of gain vs. loss frame in "Asian disease" problem with party endorsement

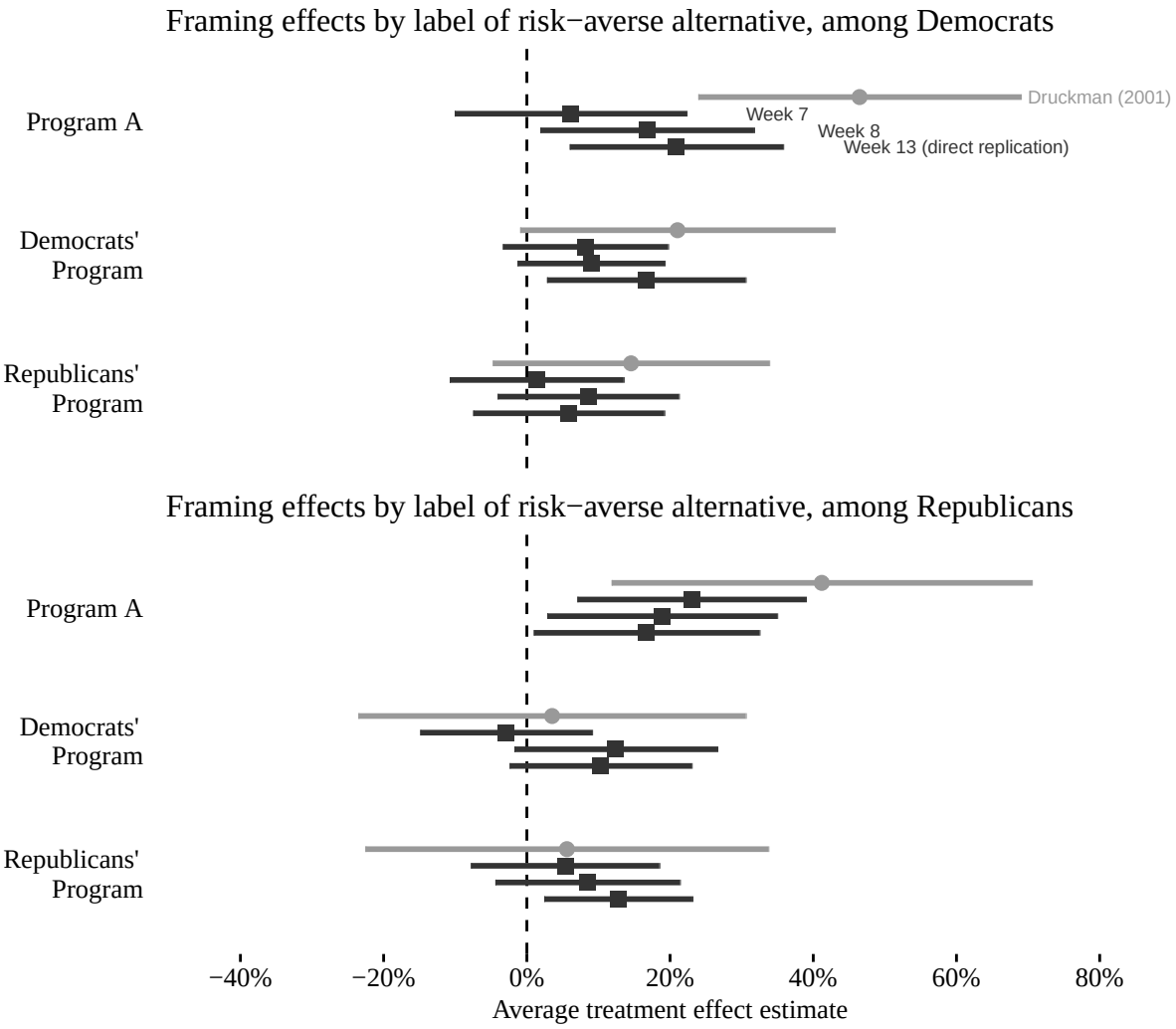


TABLE 1: Summary of effect sizes in "Asian disease" problem with party endorsement

Label of risk-averse alternative	Partisan subgroup	Replication Summary	Original Study	Difference	Relative Effect Size
Program A	Democrats	0.15 (0.04)*	0.46 (0.11)*	-0.31 (0.12)*	0.32
Program A	Republicans	0.20 (0.05)*	0.41 (0.15)*	-0.22 (0.15)	0.47
Democrats' Program	Democrats	0.11 (0.03)*	0.21 (0.11)	-0.11 (0.12)	0.50
Democrats' Program	Republicans	0.06 (0.04)	0.04 (0.13)	0.02 (0.14)	1.66
Republicans' Program	Democrats	0.05 (0.04)	0.15 (0.10)	-0.09 (0.10)	0.35
Republicans' Program	Republicans	0.10 (0.03)*	0.06 (0.14)	0.04 (0.14)	1.70

Notes: Relative size is the summary effect size divided by the original effect size: values less than 1 indicate summary effect sizes smaller than original effect sizes. $P < 0.05^*$.

A.6 Foreign aid misperceptions

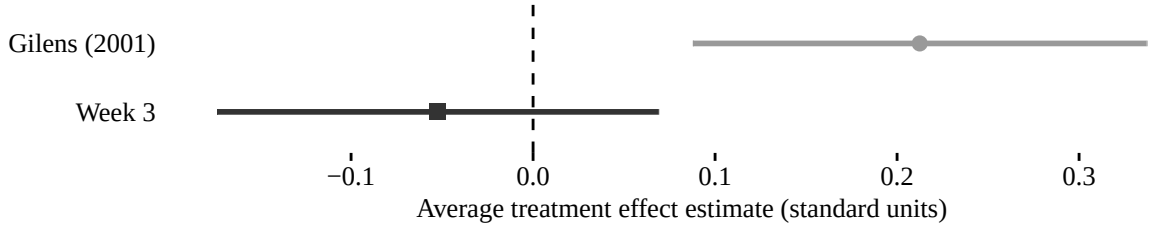
In the original experiment (Gilens, 2001), respondents from a nationally representative telephone survey fielded in 1998 were queried about their support for spending on foreign aid after being read a randomly assigned prompt about a hypothetical news story. In the control condition, the prompt read "The story is about a news report that was just released about American foreign aid to help other countries. Have you heard about this story?" In the treatment condition, the prompt added factual information designed to correct misperceptions about foreign aid spending: "It said that the amount of money we spend for foreign aid has been going down and now makes up less than one cent of every dollar that the federal government spends."

Following the prompt, respondents were asked: "How do you feel about the amount of money the federal government (in Washington) spends on foreign aid to other countries? Do you think the federal government should spend more of foreign aid, less, or about the same as it does not?" The original study reported that respondents in the treatment condition were 16.6 percentage points less likely to support cuts for foreign aid than respondents in the control group. According to Gilens (2001), this illustrates that the general public over-estimates the percentage of the budget allocated to foreign aid, but correcting this misperception with policy-specific information can decrease opposition to foreign aid.

In the original study, support for foreign aid was measured on a 3-point scale: "Less" (coded -1), "About the same" (coded 0), "More" (coded 1). The 16.6 percentage point effect reported in Gilens (2001) was obtained from a logistic regression of the binary treatment indicator on a truncated outcome, i.e. $Y_i = 1$ if respondent selected "Less"; 0otherwise, and a variety of control variables. In our reanalysis of the original data, we estimate treatment effects using difference-in-means with the raw three-point outcome variable.

Figure A.6 plots estimated treatment effects from the original study (telephone interview) alongside our replication (online interview). The estimated treatment effect in the original study is an increase in support for foreign aid of 0.21 scale points (SE = 0.06, $P =< 0.01$). The estimated treatment effect in the replication study is a decrease in support for foreign aid by 0.05 scale points (SE = 0.06, $P = 0.40$). The estimate from the original study is therefore 0.26 scale points larger – in the opposite direction – than the replication study (SE = 0.09, $P < 0.01$). This is the only experimental result among our set that we classify as a clear replication failure.

Figure A.6: Effect of policy-specific information on support for foreign aid



A.7 Perceived intentionality for side effects

In the original study (Knobe, 2003, Study 1), individuals were recruited from a New York City park to participate in an experiment based on the following vignette:

> *The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, and it will also help*

the environment."

The chairman of the board answered, "I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program."

Respondents assigned to the *Help* condition read that the chairman's decision had a helpful side effect: "They started the new program. Sure enough, the environment was helped." Those assigned to the *Harm* condition read "They started the new program. Sure enough, the environment was harmed." In the *Help* condition, 23% of subjects agreed with the statement "The chairman helped the environment intentionally," whereas 82% of those in the *Harm* condition agreed that "The chairman harmed the environment intentionally." According to Knobe (2003), this estimated treatment effect of 59 percentage points illustrates how the perceived intentionality of individual actions depends upon whether their consequences are helpful or harmful.

This experiment has been replicated multiple times in both student and online samples. We use the replication study from Klein et al. (2018) – which found an estimated treatment effect of 64 percentage points – as our pre-COVID benchmark for online samples. In our replications, half the participants were assigned to an experiment using this same scenario as the original study. The other half were assigned to a COVID-specific scenario, based on the following vignette:

The vice-president of a company went to the chairman of the board and said, "We are thinking of marketing a new drug to treat COVID-19. It will help us increase profits, and the drug will also help older people with heart conditions."

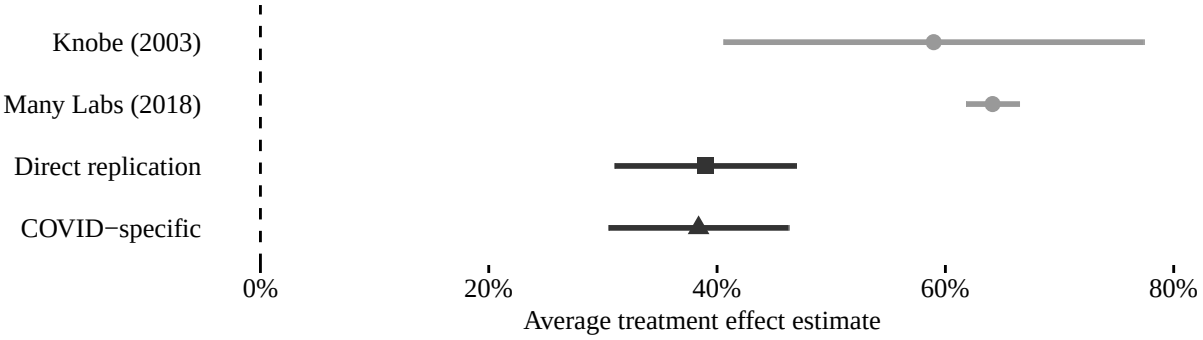
The chairman of the board answered, "I don't care at all about helping older people with heart conditions. I just want to make as much profit as I can. Let's start marketing the new drug."

Those assigned to the *Help* condition read that the chairman's decision had a harmful

side effect: "They started marketing the new drug. Sure enough, older people with heart conditions were helped." Those assigned to the *Harm* condition instead read that older people with heart conditions were "harmed". See Figures C.5-C.8 for full text of each condition.

Figure A.7 plots estimated treatment effects from the original study (sample from Manhattan park) alongside the Many Labs (2018) replication (online sample) and our replications (online sample). The estimated treatment effect is 0.39 (SE = 0.04, $P < 0.01$) in the direct replication and 0.38 in the COVID-specific replication (SE = 0.04, $P < 0.01$). The estimated treatment effect for the COVID-specific replication is about 0.01 points smaller than the direct replication (SE = 0.04, $P = 0.46$). The direct replication is approximately 60% the size of the pre-COVID benchmark (difference of 0.25 points, SE = 0.04, $P < 0.01$). The replication estimates are considerably smaller than the pre-COVID benchmark, but all estimates are in the expected direction and statistically distinguishable from zero. We therefore conclude that the pre-COVID study replicated, regardless of whether COVID-specific language was used in the vignettes.

Figure A.7: Effect of *Harm* vs. *Help* frame on perceived intentionality



A.8 Atomic aversion

In the original study (Press, Sagan and Valentino, 2013), a quota sample of online survey participants recruited by YouGov were randomly assigned to participate in one of two in-

dependent experiments. In the *prospective* experiment, subjects read a hypothetical news article that reported U.S. officials were deciding between nuclear and conventional military options for destroying an Al Qaeda nuclear weapons lab in Syria. The article compared the expected effectiveness of each military option, and estimated 1,000 Syrian civilian deaths regardless of which option was pursued. Within this experiment, subjects were assigned to one of three treatment arms that varied only in the likely success of the conventional strike: 1) a "90/90" condition in which the nuclear and conventional strike both had a 90% chance of success; 2) a "90/70" condition in which the conventional strike had a 75% chance of success; 3) a "90/45" condition in which the conventional strike had a 45% chance of success. The relative effectiveness of each option was described in the article text, alongside a two-by-two matrix that compared the chances of success (90/90, 90/70, or 90/45) and estimated civilian causalities (fixed) for both options.

In the *retrospective* experiment, subjects read a hypothetical news article that described a U.S. military strike that had already been carried out on the Al Qaeda lab. Within this experiment, subjects were assigned to one of two treatment arms that described the weapons used to carry out the attack: 1) a "conventional strike" condition in which 100 conventional cruise missiles were used; 2) a "nuclear strike" condition in which 2 nuclear cruise missiles were used. The number of civilian casualties and the outcome (the lab was successfully destroyed) were fixed across conditions.

In both experiments, all subjects were informed prior to random assignment that, if they failed to pass comprehension questions about the article, they could be ineligible to finish. If they instead answered the comprehension questions correctly, they were told they would be eligible to participate in a raffle for a $100 gift certificate. subjects who failed to pass the post-treatment comprehension questions were excluded from the analysis sample.

Aronow, Baron and Pinson (2019) noted this practice of dropping subjects who fail post-treatment "manipulation checks" can induce bias in estimates of treatment effects. They

replicated the original study on a sample of MTurk workers and found that retaining subjects who failed the comprehension questions resulted in different estimates than the original study. In this replication, subjects were also told in advance that they would be ineligible to complete the survey if they failed the comprehension questions, but were entered into a raffle for a $100 bonus payment if they passed. In addition, respondents in the *prospective* experiment viewed a large version of the two-by-two graphic that appeared in the article after the comprehension questions were answered, but before viewing any outcome questions.

Lucid does not give researchers the ability to pay survey respondents bonuses, so no incentives could be offered for those who passed the comprehension questions in our replications. In addition, subjects in our replications of the prospective experiment viewed the article once, and a two-by-two graphic was not presented after the comprehension questions (as in Aronow, Baron and Pinson, 2019). All other design details were the same as in the original study.

Figure A.8 plots estimated treatment effects in the prospective experiment, with the 90/90 condition as the control group, for the original study, the replication by Aronow, Baron and Pinson (2019), and our three replications. In the original study, the 90/70 condition caused an increase in the proportion of subject that preferred the nuclear option by about 37 percentage points relative to the 90/90 condition. The estimated effect of the 90/45 condition, relative to the 90/90 condition, was 51 percentage points. Similarly, the 90/70 condition caused an increase in the proportion of subjects that approved of the nuclear option by about 17 percentage points, and the 90/45 condition caused an increase of about 27 percentage points. In other words, estimated treatment effects increased monotonically with the relative effectiveness of nuclear weapons.

Estimated treatment effects in the pre-COVID replication by Aronow, Baron and Pinson (2019) were similar to the original study for both outcome measures. The estimated treatment effects in our replications were considerably smaller for the "Prefer Nuclear Use"

outcome, but of the expected sign and statistically distinguishable from zero. Estimates for the "Approve Nuclear Use" outcome, however, were of the opposite sign and not distinguishable from zero in 2/3 of our replications. Table 2 provides a summary of the differences in estimates from the prospective experiment in our replications, the original study, and the ABP replication.

Figure A.9 plots estimated treatment effects in the retrospective experiment, with the "conventional strike" condition as the control group. The original study reported that differences between the "nuclear strike" (treatment) and "conventional strike" (control) were "substantively small and not statistically significant" (Press, Sagan and Valentino, 2013, p. 197). The pre-COVID replication byAronow, Baron and Pinson (2019) found, however, that the nuclear strike caused a 12 percentage point reduction in the proportion of respondents who "approved" the strike, and a 13 percentage point reduction in the proportion who believed the strike was "ethical". Table 3 provides a summary of the differences in estimates from the retrospective experiment in our replications, the original study, and the ABP replication.

When compared to the original study and the ABP replication, our replication estimates for the "Prefer Nuclear Use" outcome in the prospective experiment are significantly smaller, but of the expected sign and statistically distinguishable from zero. The estimates for the "Approve Nuclear Use" outcome are, however, signed in the opposite direction and indistinguishable from zero. Estimates for both outcomes in the retrospective experiment are comparable to those reported in the original study and the ABP replication. We therefore conclude that the atomic aversion experiment was partially replicated.

Figure A.8: Support for prospective U.S. strike on Al Queda nuclear weapons lab in Syria

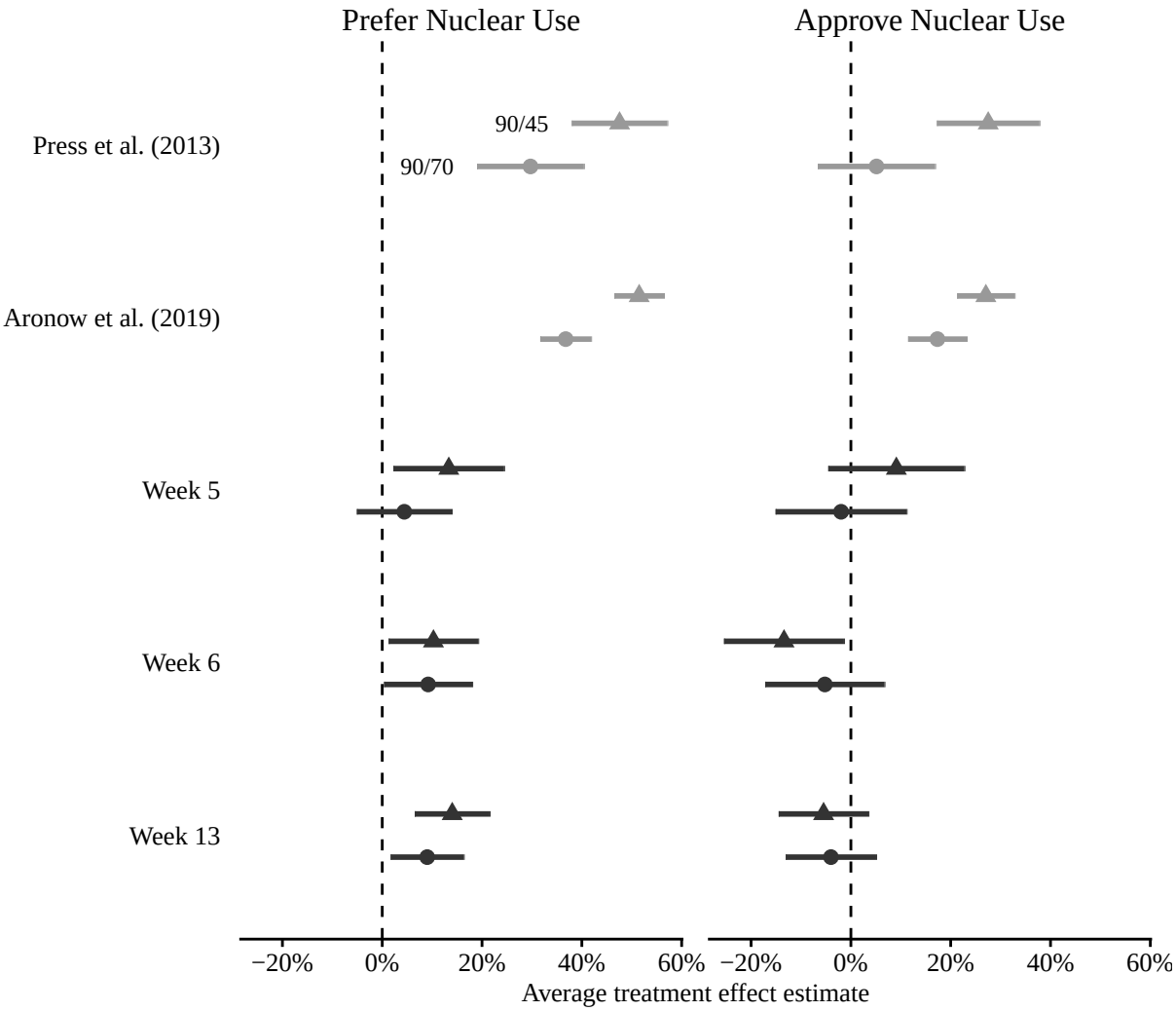


TABLE 2: Summary of estimates in *prospective* atomic aversion experiment

Group	Outcome	Replication Summary	Original Study	Difference	Relative Size	ABP Replication	Difference	Relative Size
90/70	Prefer	0.08 (0.02)*	0.30 (0.05)*	-0.22 (0.06)*	0.27	0.37 (0.03)*	-0.29 (0.04)*	0.21
90/45	Prefer	0.13 (0.03)*	0.48 (0.05)*	-0.35 (0.06)*	0.27	0.51 (0.03)*	-0.39 (0.04)*	0.25
90/70	Approve	-0.04 (0.03)	0.05 (0.06)	-0.09 (0.07)	-	0.17 (0.03)*	-0.21 (0.04)*	-
90/45	Approve	-0.05 (0.03)	0.28 (0.05)*	-0.32 (0.06)*	-	0.27 (0.03)*	-0.32 (0.04)*	-

Notes: Relative effect sizes are the replication summary effect sizes divided by the ABP replication or original effect size: values less than 1 indicate summary effect size is smaller than the ABP or original effect size. Relative effect sizes are not calculated if replication estimates are the opposite sign of comparison estimates. $P < 0.05^*$.

Figure A.9: Support for retrospective U.S. strike on Al Queda nuclear weapons lab in Syria

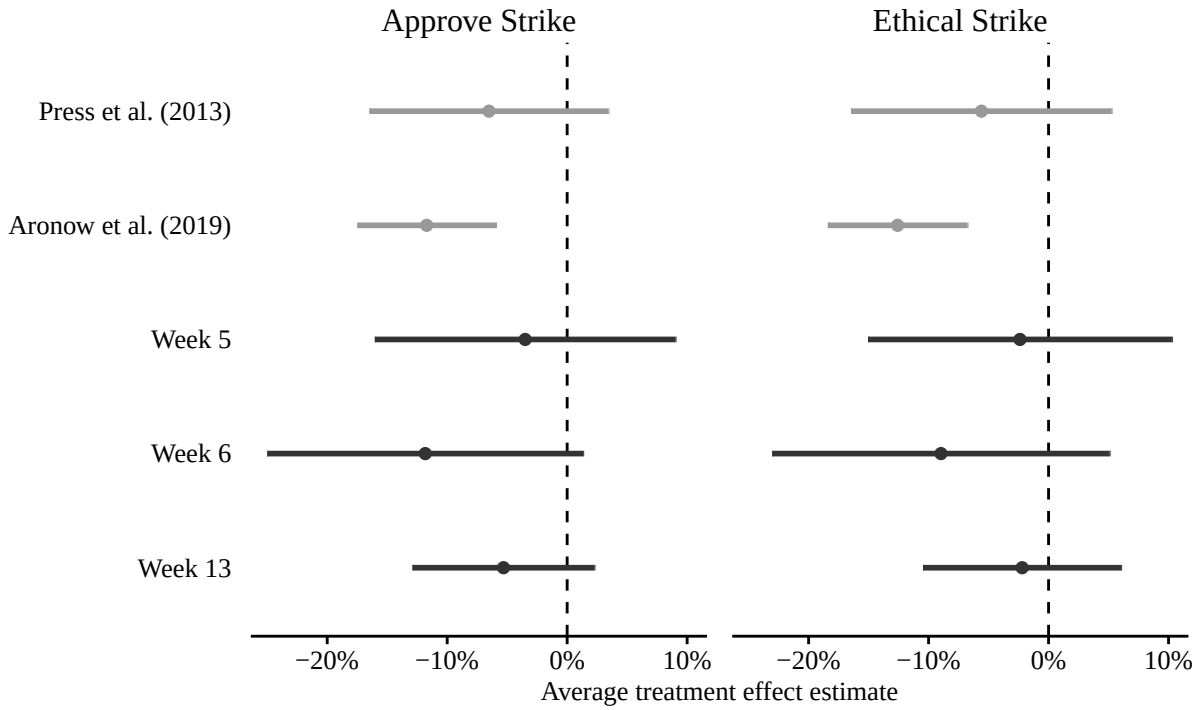


TABLE 3: Summary of estimates in *retrospective* atomic aversion experiment

Outcome	Replication Summary	Original Study	Difference	Relative Size	ABP Replication	Difference	Relative Size
Approve	-0.06 (0.03)*	-0.07 (0.05)	0.00 (0.06)	0.95	-0.12 (0.03)*	0.06 (0.04)	0.53
Ethical	-0.04 (0.03)	-0.06 (0.06)	0.02 (0.06)	0.64	-0.13 (0.03)*	0.09 (0.04)*	0.28

Notes: Relative effect sizes are the replication summary effect sizes divided by the ABP replication or original effect size: values less than 1 indicate summary effect size is smaller than the ABP or original effect size. Relative effect sizes are not calculated if replication estimates are the opposite sign of comparison estimates. $P < 0.05^*$.

A.9 Attitudes toward immigrants

In the original study (Hainmueller and Hopkins, 2015), 1,407 U.S. respondents from a nationally representative online survey fielded by Knowledge Networks in 2012 participated in a conjoint experiment that asked them to choose between different pairs of hypothetical

immigrants applying for admission. Each respondent evaluated five different pairs of immigrants, with immigrants' backgrounds varying along nine randomly assigned attributes: gender, education, employment plans, job experience, profession, language skills, country of origin, reasons for applying, and prior trips to the United States. Each attribute contained multiple levels (e.g. country was 10 levels and gender was 2) for a total of approximately 900,000 unique immigrant profiles.

After viewing each immigrant pair subjects were presented with a binary choice: "If you had to choose between them, which of these two immigrants should be given priority to come to the United States to live?" Each subject evaluated 5 pairs of immigrants for a total of 14,070 observations (1,407 respondents $\times$ 5 pairs $\times$ 2 immigrants per pair). We conducted a direct replication of this conjoint experiment in May 2020 on a sample of 1,328 respondents, for a total of 13,280 observations.

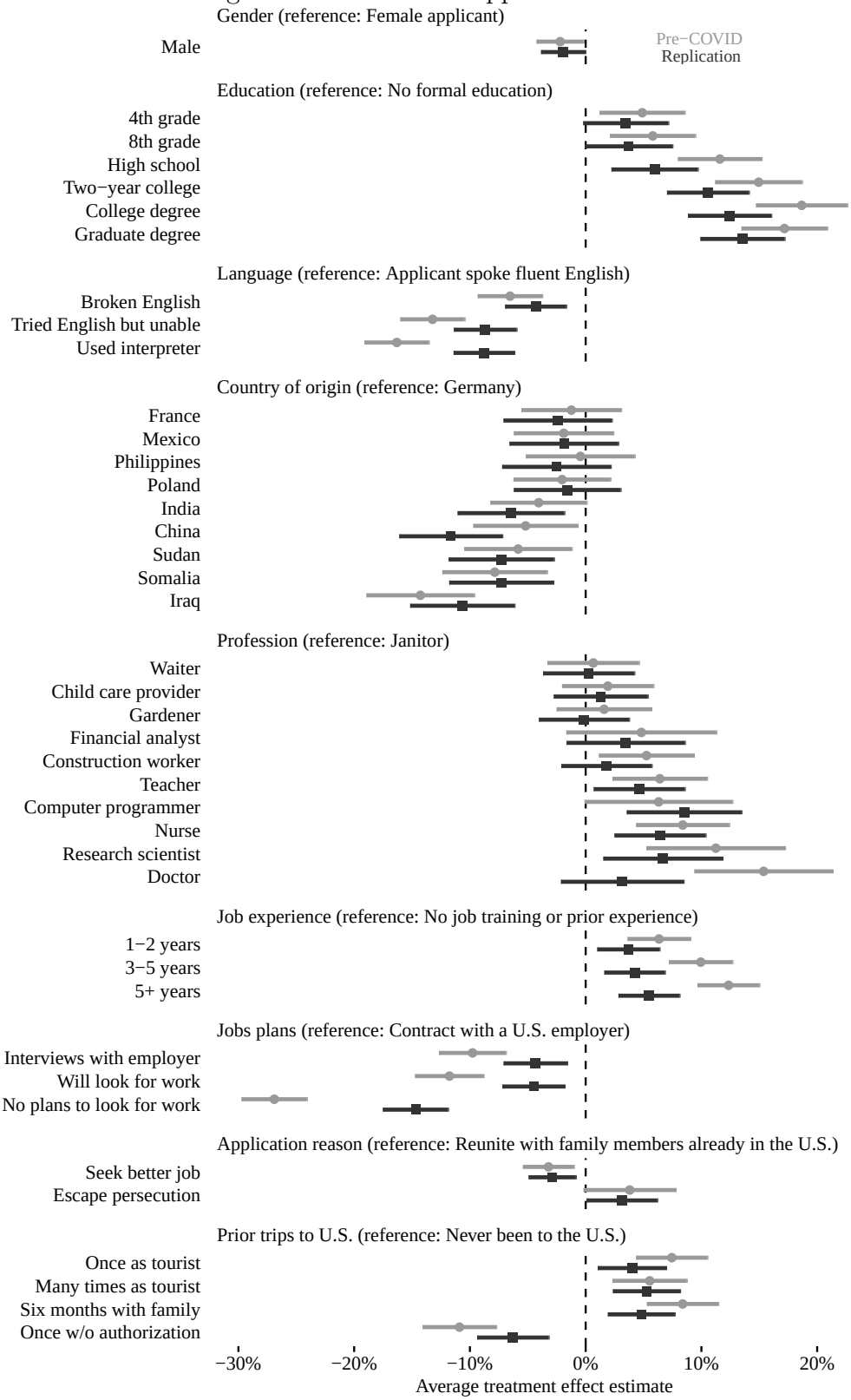
Following Hainmueller and Hopkins (2015), we estimate the Average Marginal Component Effects (AMCEs) for each attribute level using a regression of the binary response (1 if the immigrant profile is preferred, 0 otherwise) on a set of indicators for each attribute level, with standard errors clustered at the level of the survey respondent. Figure A.10 plots the results for the original study alongside our direct replication. The top of each panel describes the omitted reference level for each attribute; for example, the negative point estimates for "Male" indicate that male immigrants are about 2 percentage points less likely to be selected than female immigrants.

In total, there are 81 estimated AMCEs in Figure A.10 – 41 for the original study and 41 for the replication. When compared to the original study, the replication estimates are remarkably similar in both direction and magnitude. Only one of 41 replication estimates is signed in the opposite direction when compared to the original study – the AMCE for "Gardener" is 0.02 points in the original study and approximately zero in the replication, but neither estimate is distinguishable from zero. The majority of the replication estimates

(27 of 41) are smaller in magnitude than the original estimates.[3] We therefore conclude that the conjoint experiment was successfully replicated.

[3]only 7 of 41 differences are statistically significant after correcting for multiple comparisons to control the false discovery rate (see e.g. Benjamini and Hochberg, 1995).

Figure A.10: Effects of immigrant attributes on support for admission to U.S.



A.10 Fake news corrections

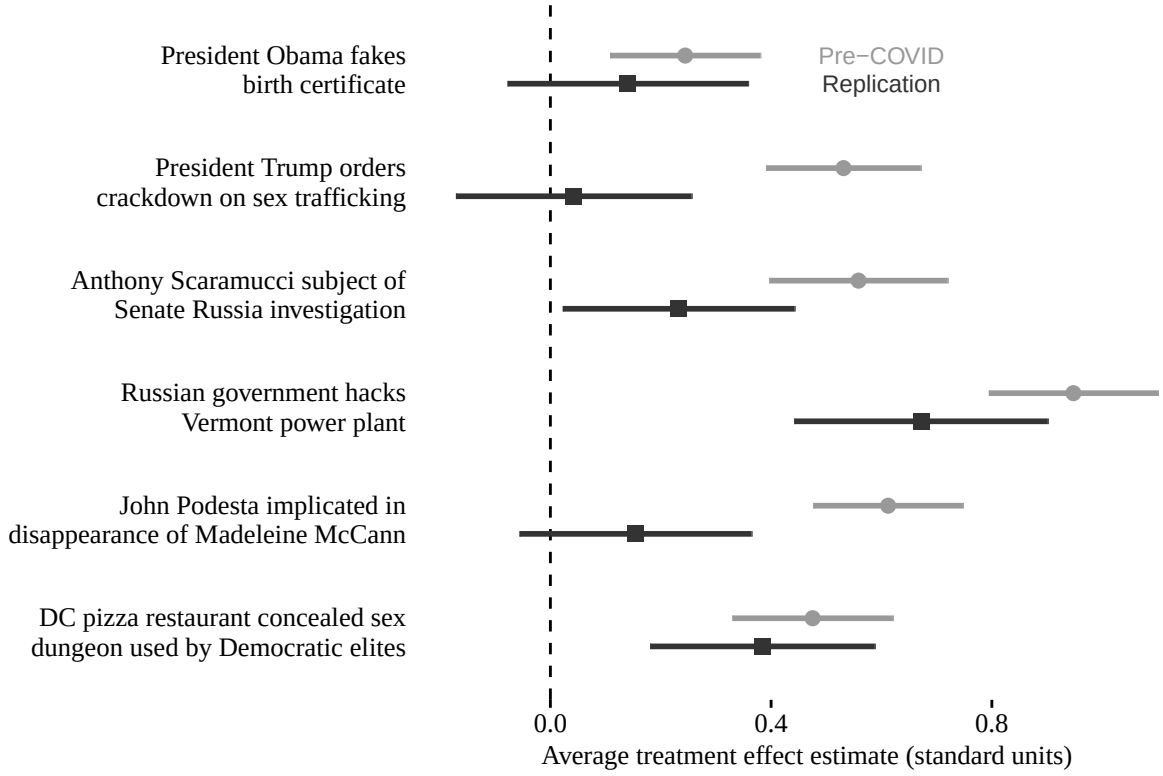
In the original study (Porter, Wood and Kirby, 2018), 2,742 MTurk workers were exposed to two fake news stories randomly selected from a sample of six fake news stories that were previously circulated (e.g. that Obama's birth certificate is fake). For each fake news story, subjects were randomly assigned to see either a correction following the story, or no correction. Therefore subjects in the "correction" (treatment) condition read a randomly assigned story followed by a correction stating that the story was false, whereas subjects in the "no correct" (control) condition simply read the story without seeing a correction. Therefore, the experiment has $6 \times 2 = 12$ treatment arms, with each respondent being exposed to two unique stories with or without a correction.

The goal of this study was to test whether corrections could reduce individuals' beliefs in the veracity of fake news stories. Following exposure to the fake news story (and the correction if assigned treatment) subjects were asked to indicate their agreement or disagreement about the truth value of the claim advanced on a 5-point scale. Across all six fake news stories, the authors found that exposure to corrections caused a significant reduction in respondents' beliefs that the stories were true, with average treatment effects ranging from -0.24 scale points on the low end to -0.95 scale points on the high end.

We conducted a direct replication of this experiment on a sample of 1,415 respondents in April 2020. Following the original study, we scale outcomes so that higher values indicate stronger agreement that the fake news stories were true. Within randomly assigned story, outcomes are standardized by dividing the response vector by the standard deviation in the "no correction" (control) group. Figure A.11 compares estimated treatment effects between the original study and replication, for each of the six stories. All replication estimates, ranging from -0.04 to -0.67, are uniformly smaller then those in the original study, but all are correctly signed.[4] We therefore conclude that that the replication was successful.

[4] Although all 6 replication estimates are smaller in magnitude than those in the original study, only two

Figure A.11: Effect of corrections on agreement with inaccurate statements



A.11 Inequality and System Justification

In the original study (Trump and White, 2018), 1,020 U.S. respondents from a nationally representative online survey fielded by Knowledge Networks in 2015 participated in a survey experiment with random assignment to two conditions. In the "low-inequality" (control) condition, respondents were exposed to information about trends in U.S. income inequality, as measured by the Gini coefficient over the period 1968-2010. In the "high-inequality" (treatment) condition, respondents were exposed to the same information, but the y-axis in the plot was truncated to make the upward trend appear much steeper.

The goal of this study was the test a hypothesis that exposure to inequality increases "system justification" – broadly, the psychological need to support the status quo, even at

of these differences (John Podesta and Trump stories) are statistically distinguishable from zero.

the expense of their self-interest, or the interests of their group (see e.g. Jost and Banaji, 1994). Trump and White (2018) test the hypothesis that higher inequality causes higher system justification by comparing differences in subjects' system justification scores between the high and low inequality conditions. The key prediction is that increasing subjects' beliefs that inequality is rising should decrease system justification.

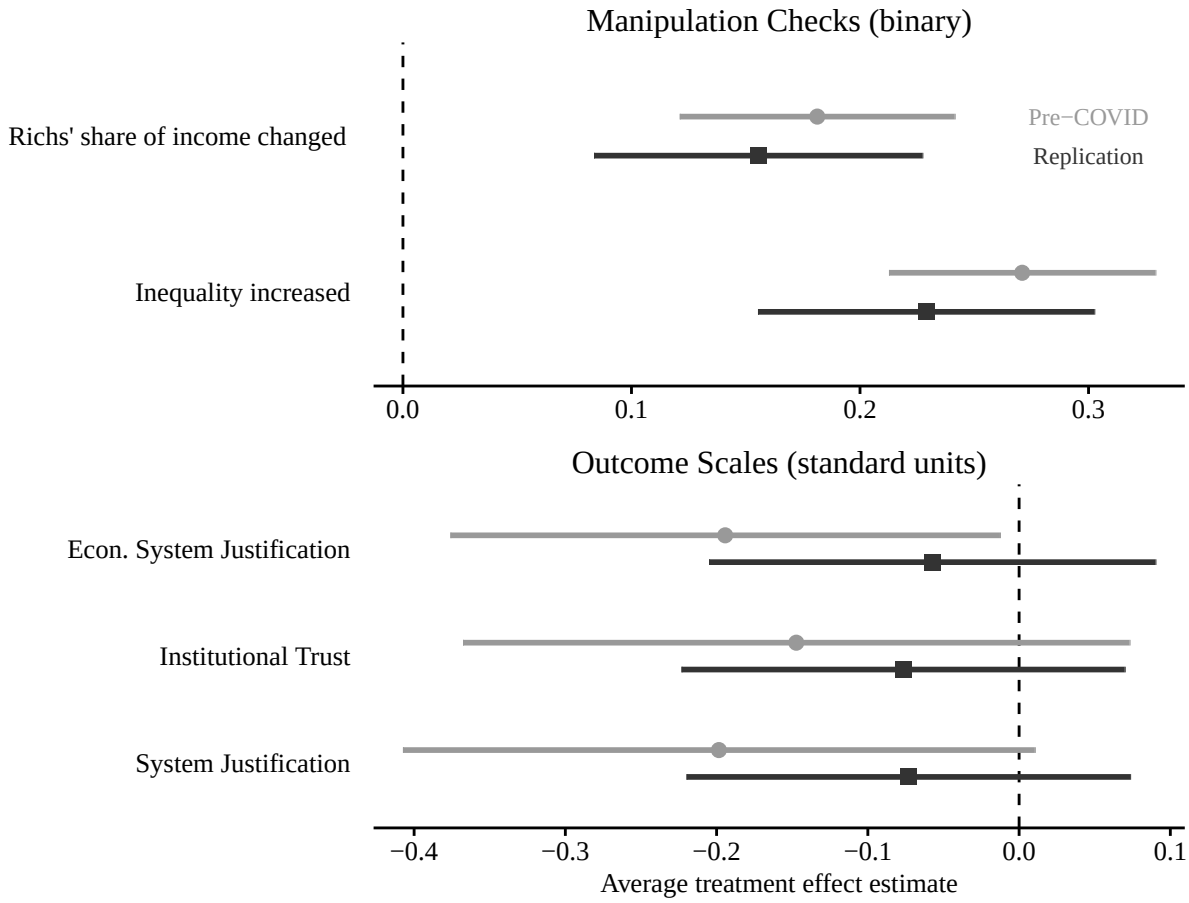
Following exposure to one of the two data visualizations, respondents completed two "manipulation check" questions. For the first, respondents were asked to say whether the statement "Income inequality in the United States has increased dramatically over time" was correct or incorrect. The second asked respondents whether the statement "the share of total income of the very rich has not changed much over time in the United States" was correct or incorrect. According to Trump and White (2018), the high-inequality treatment should cause an increase in the proportion of respondents stating "correct" to the first question and "incorrect" to the second question, relative to control. Responses to the first question are therefore coded 1 if the subject answers "correct" and 0 otherwise. Responses to the second questions are coded 1 if the subject answers "incorrect" and 0 otherwise.

After this, respondents were randomly assigned to complete one of three batteries of questions the authors used to measure system justification: an institutional trust scale (6-items), a system justification scale (8-items), or an economic system justification scale (15-items). In total, 339 subjects (169 in control, 170 in treatment) completed the system justification scale, 338 completed the institutional trust scale (169 in treatment, 169 in control), and 336 completed the economic system justification scale (167 in treatment and 169 in control). In our replication, 804 subjects were presented with all three system justification scales in randomized order. Following the original study, we scale outcomes so that higher values indicate higher levels of system justification.

Figure A.12 compares estimated treatment effects on the manipulation check questions (top panel) and outcomes scales (bottom panel) between the original study and the replica-

tion. All replication estimates are in the expected direction when compared to the original study. Although all 5 replication estimates are smaller in magnitude than those in the original study, none of these differences are statistically distinguishable from zero. We therefore conclude that the replication was successful.

Figure A.12: Effect of "high inequality" treatment on comprehension questions and system justification scales



A.12 Trust in government and redistribution

In the original study (Peyton, 2020), a total of 3,837 U.S. respondents were exposed to information about corruption in American government across three separate experiments: Experiment 1 (624 MTurk workers in 2014); Experiment 2 (nationally representative sample

of 1,324 U.S. adults in 2014); Experiment 3 (1,870 MTurk workers in 2017). In each experiment, participants were randomly assigned to one of three treatment arms: "Corrupt", "Honest", or "Control". In the "Corrupt" arm, subjects read an Op-Ed by a former DOJ prosecutor that described high levels of political corruption in American politics; the "Honest" arm used contrasting language to describe low levels of political corruption. In the "Control" arm, subjects read an article of similar length that was devoid of political content. Experiment 2 was a direct replication of Experiment 1, and Experiment 3 supplemented the articles with data visualizations that supported the writers' arguments.

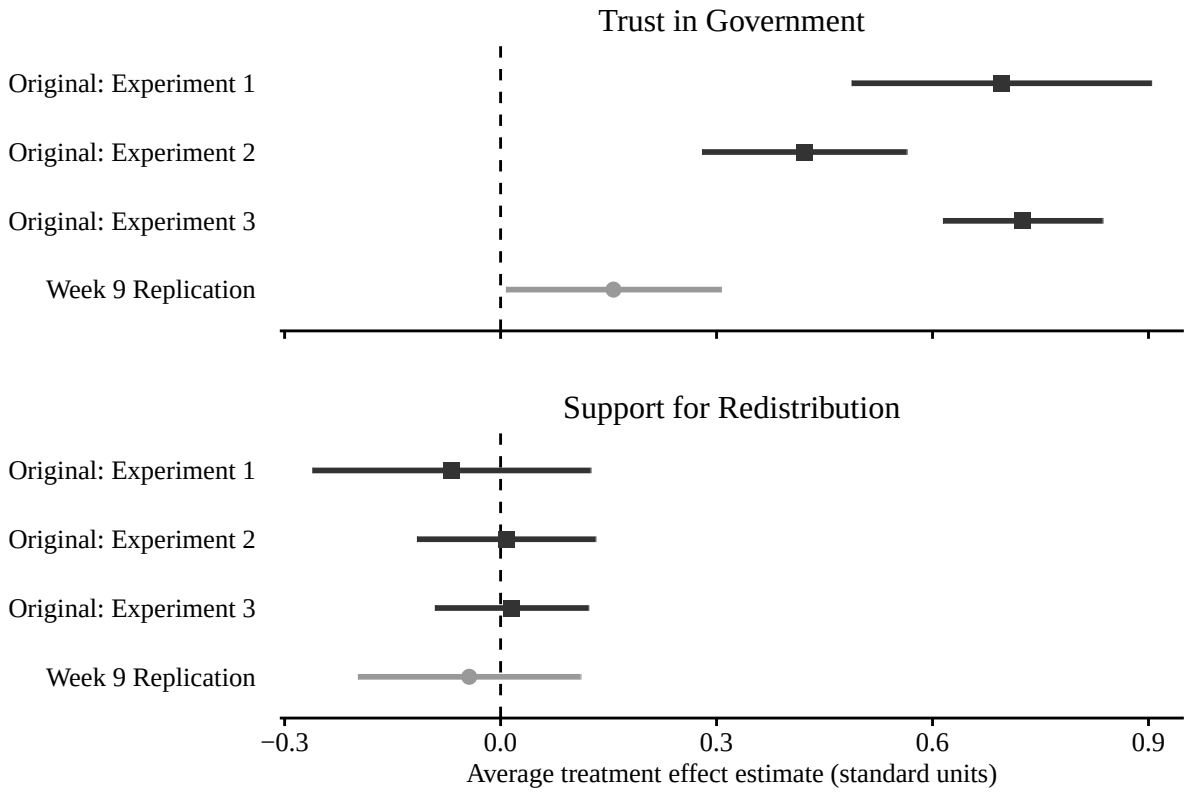
These experiments were used to test a hypothesis that increasing trust in government causes Americans to become more supportive of redistribution (see, e.g. Hetherington et al., 2005). Peyton (2020) tests this hypothesis by experimentally manipulating respondents' trust in government and testing for downstream effects on respondent's support for redistribution using a causal instrumental variables framework. Following exposure to treatment, subjects' trust in government was measured using a 4-item scale. Next, subject's support for redistribution was measured using a 4-item scale about federal spending redistributive social policies. The author found significant effects on subjects' trust in government in all three experiments, but support for redistribution was indistinguishable from zero.

We conducted a direct replication of Experiment 3 in the original study on a sample of 1,424 respondents in May 2020. Following the original study, treatment was coded 0 if a subject was assigned to the "Corrupt" arm, 0.5 if assigned "Control", and 1 if assigned "Honest". Outcomes were scaled so that higher values indicate more trust in government, and more support for redistribution.

Figure A.13 compares estimated treatment effects on trust in government (top panel) and support for redistribution (bottom panel) between the replication and Experiments 1-3 in the original study. The estimated treatment effect on trust in government in the replication is statistically distinguishable from zero, and in the expected direction. However, this estimate

is significantly smaller in magnitude than all of the estimates in the original study. The estimated treatment effects on support for redistribution are indistinguishable from zero in the replication, and statistically indistinguishable from the estimates in the original study. We therefore conclude this was a successful replication.

Figure A.13: Effect of corruption on trust in government and support for redistribution



B Covariate distributions

Figure B.1: Region proportions by sample

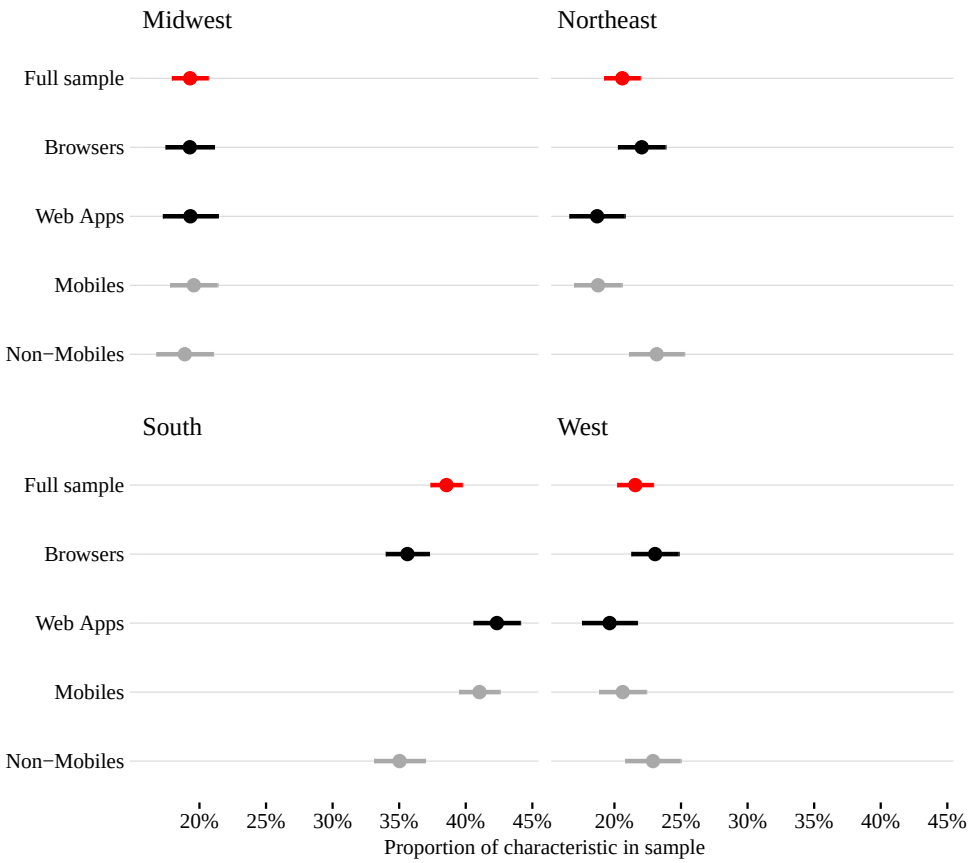


Figure B.2: Education proportions by sample

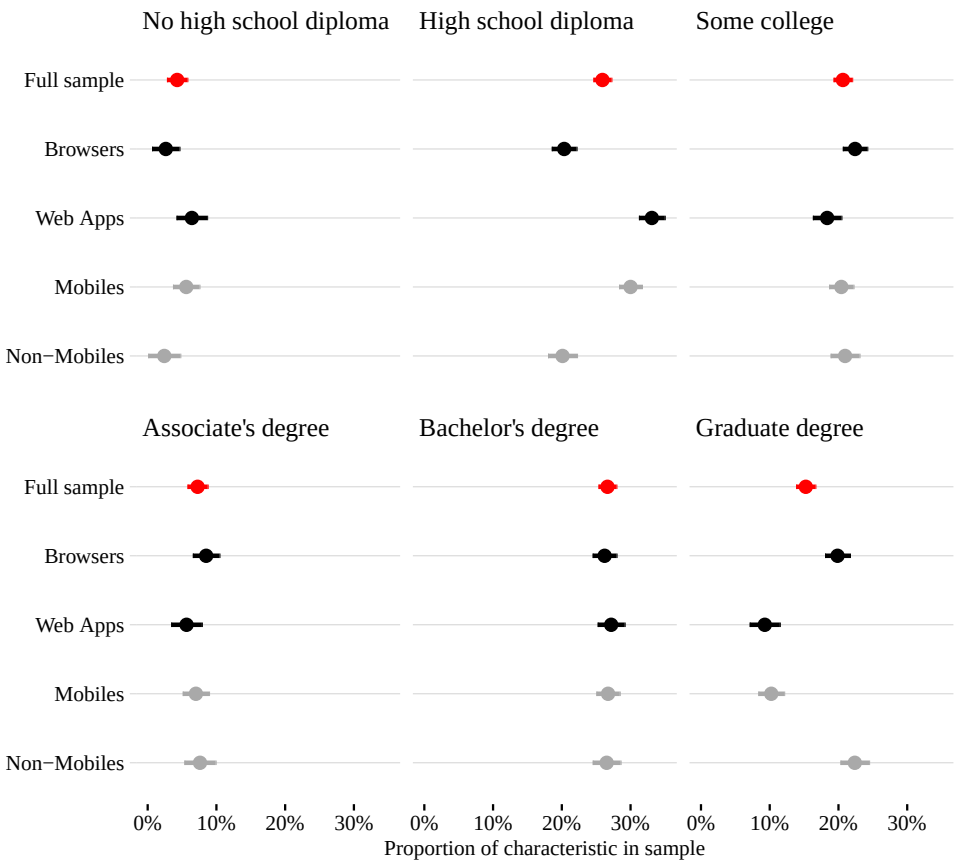


Figure B.3: Household income proportions by sample

Figure B.4: Age proportions by sample

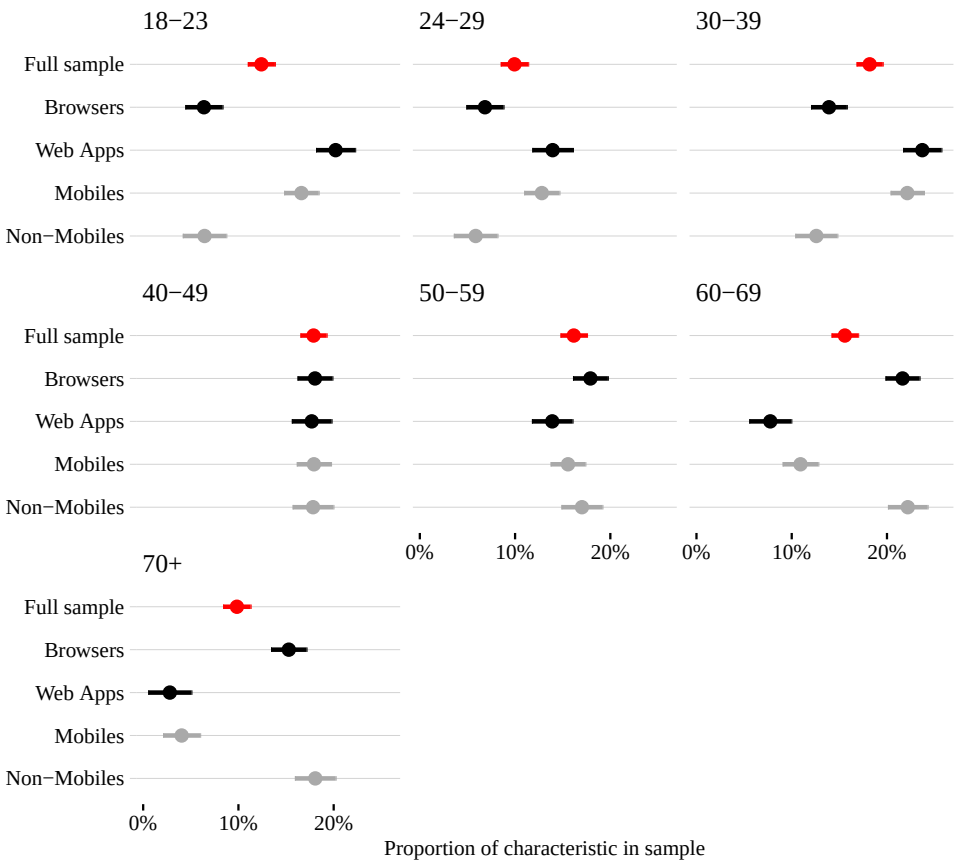


Figure B.5: Male v. Female proportions by sample

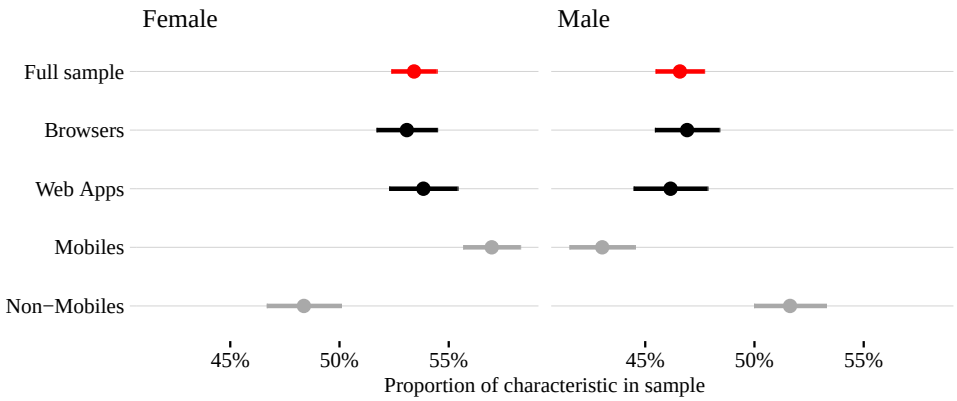


Figure B.6: Race/Ethnicity proportions by sample

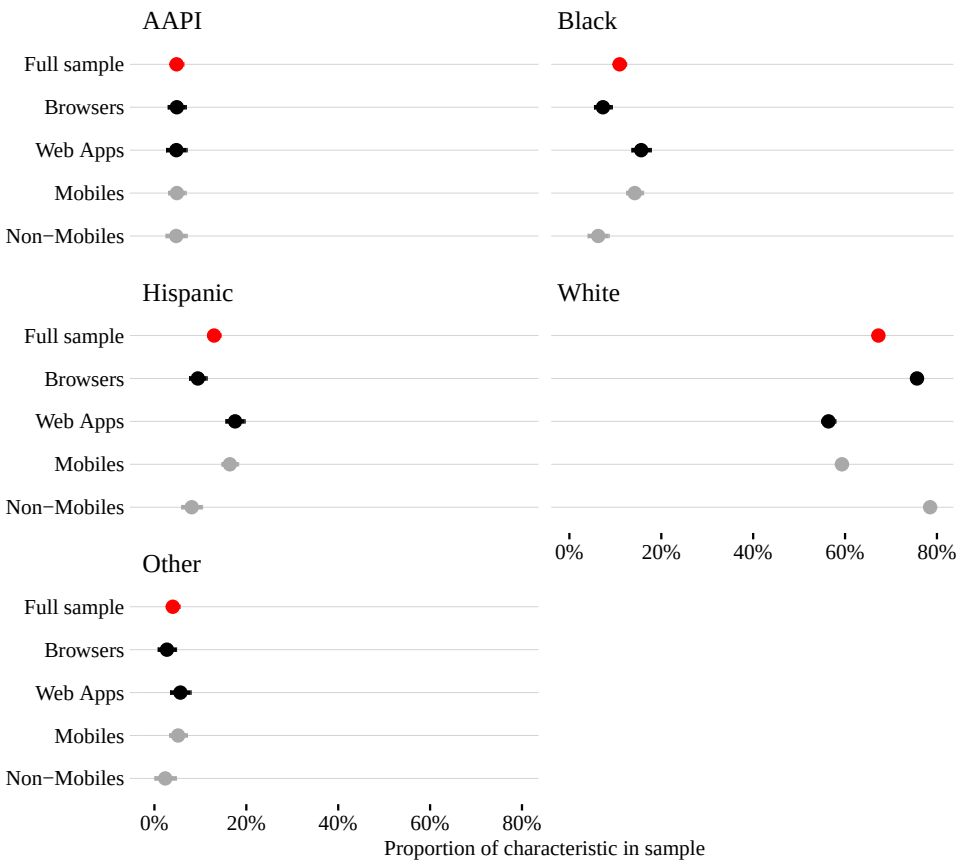


Figure B.7: Partisanship proportions by sample

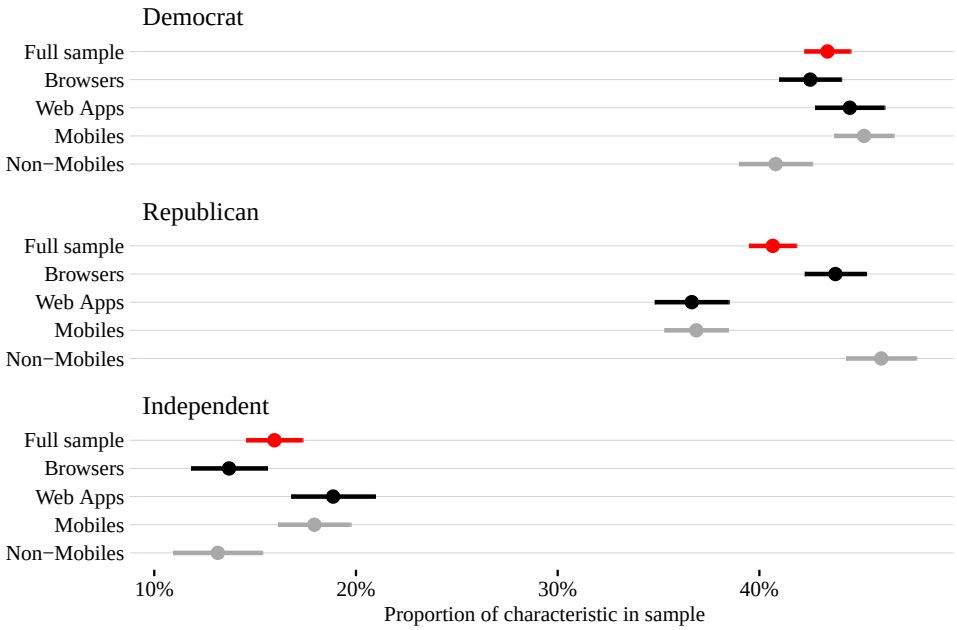
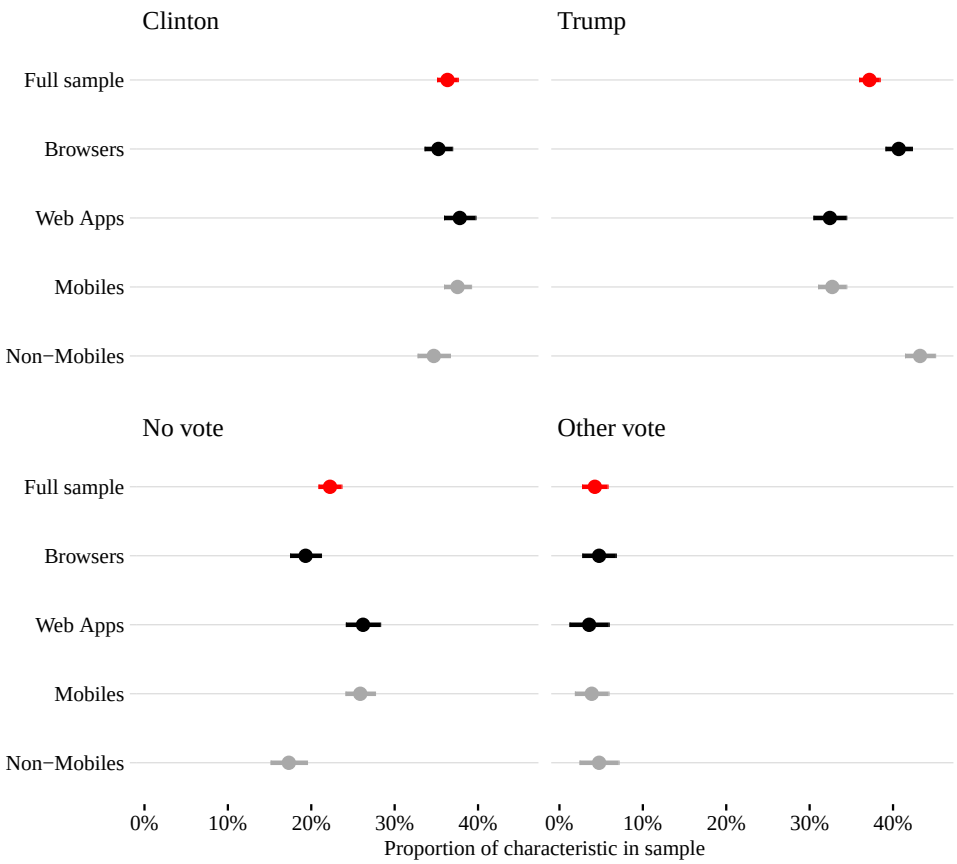


Figure B.8: Voting behavior in 2016 proportions by sample



C Treatment descriptions

Figure C.1: Effect of framing on decision making: cheap condition (original)

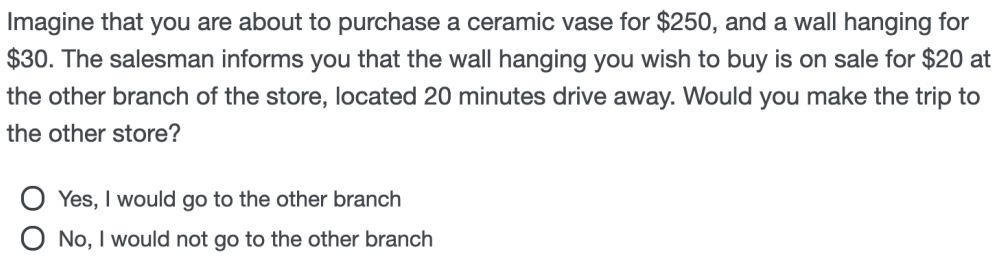


Figure C.2: Effect of framing on decision making: expensive condition (original)

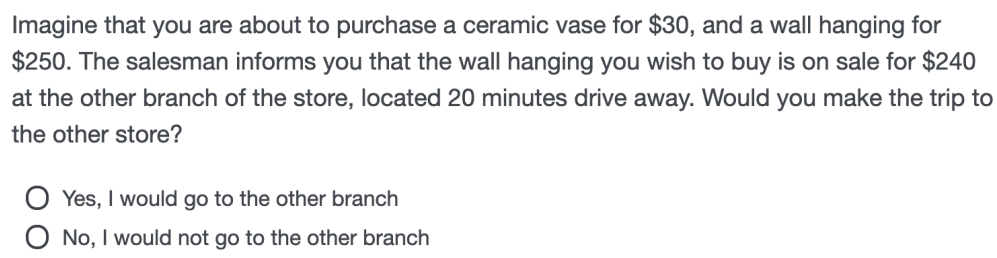


Figure C.3: Effect of framing on decision making: cheap condition (modified)

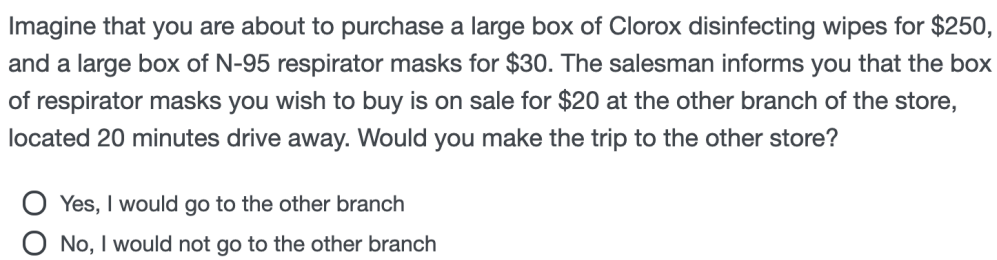


Figure C.4: Effect of framing on decision making: expensive condition (modified)

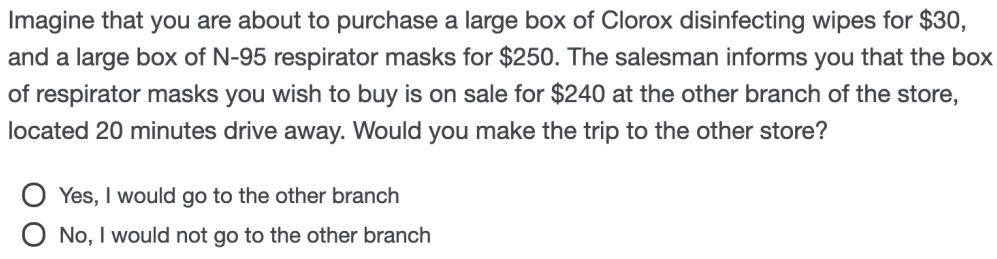


Figure C.5: Perceived intentionality for side effects: helped condition (original)

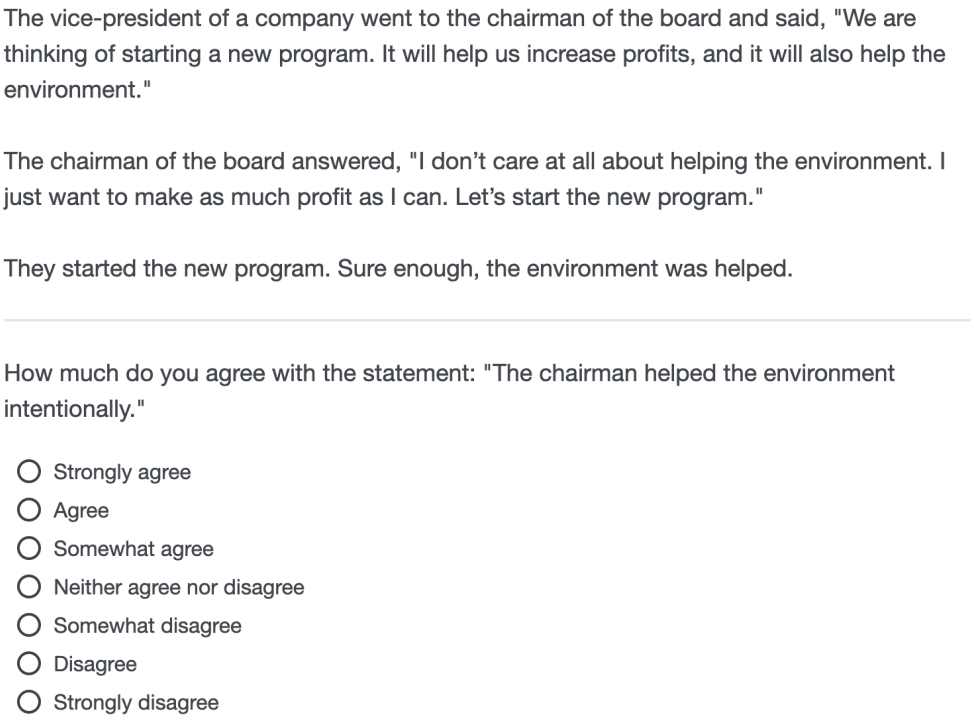


Figure C.6: Perceived intentionality for side effects: harmed condition (original)

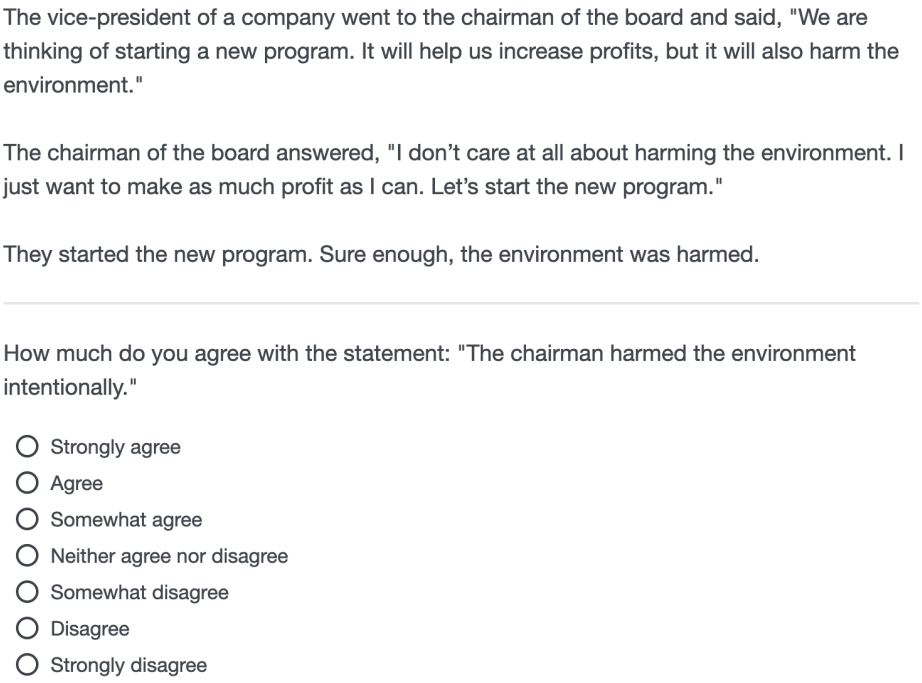


Figure C.7: Perceived intentionality for side effects: helped condition (modified)

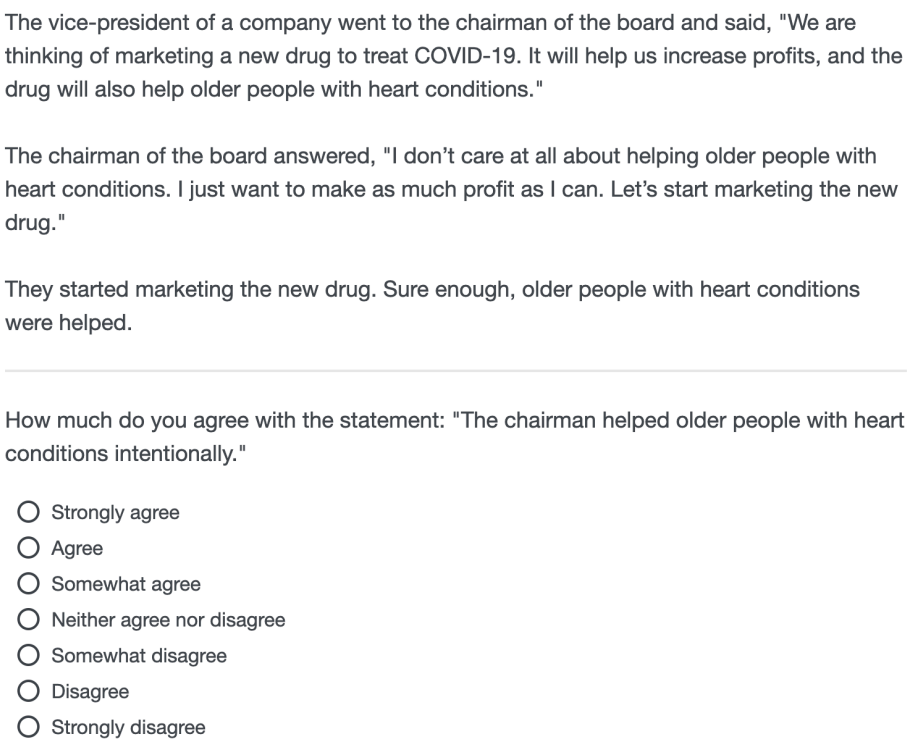
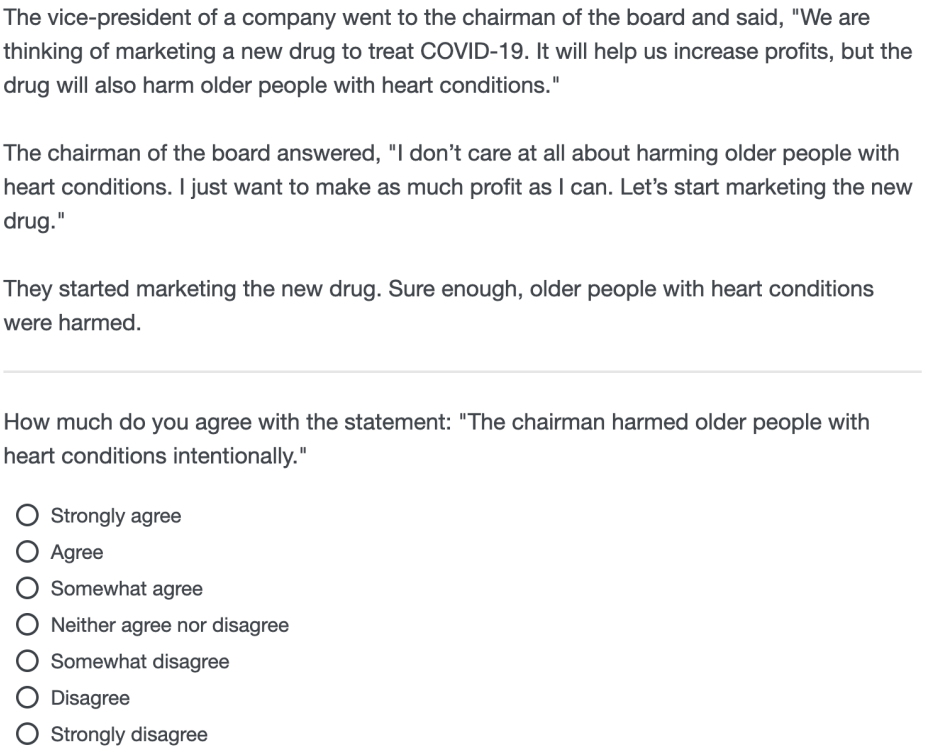
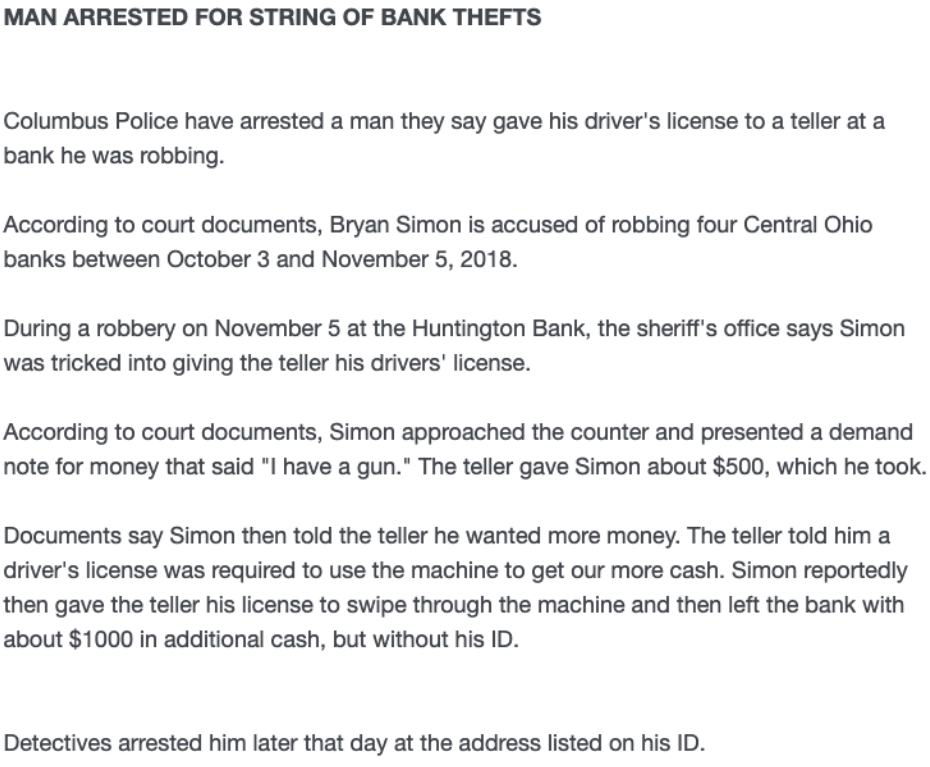


Figure C.8: Perceived intentionality for side effects: harmed condition (modified)



C.1 Attention Check Questions

Figure C.9: Pre-ACQ article for "'Easy" and "Medium" ACQ



MAN ARRESTED FOR STRING OF BANK THEFTS

Columbus Police have arrested a man they say gave his driver's license to a teller at a bank he was robbing.

According to court documents, Bryan Simon is accused of robbing four Central Ohio banks between October 3 and November 5, 2018.

During a robbery on November 5 at the Huntington Bank, the sheriff's office says Simon was tricked into giving the teller his drivers' license.

According to court documents, Simon approached the counter and presented a demand note for money that said "I have a gun." The teller gave Simon about $500, which he took.

Documents say Simon then told the teller he wanted more money. The teller told him a driver's license was required to use the machine to get our more cash. Simon reportedly then gave the teller his license to swipe through the machine and then left the bank with about $1000 in additional cash, but without his ID.

Detectives arrested him later that day at the address listed on his ID.

Figure C.10: "Easy" and "Medium" ACQ with correct responses highlighted

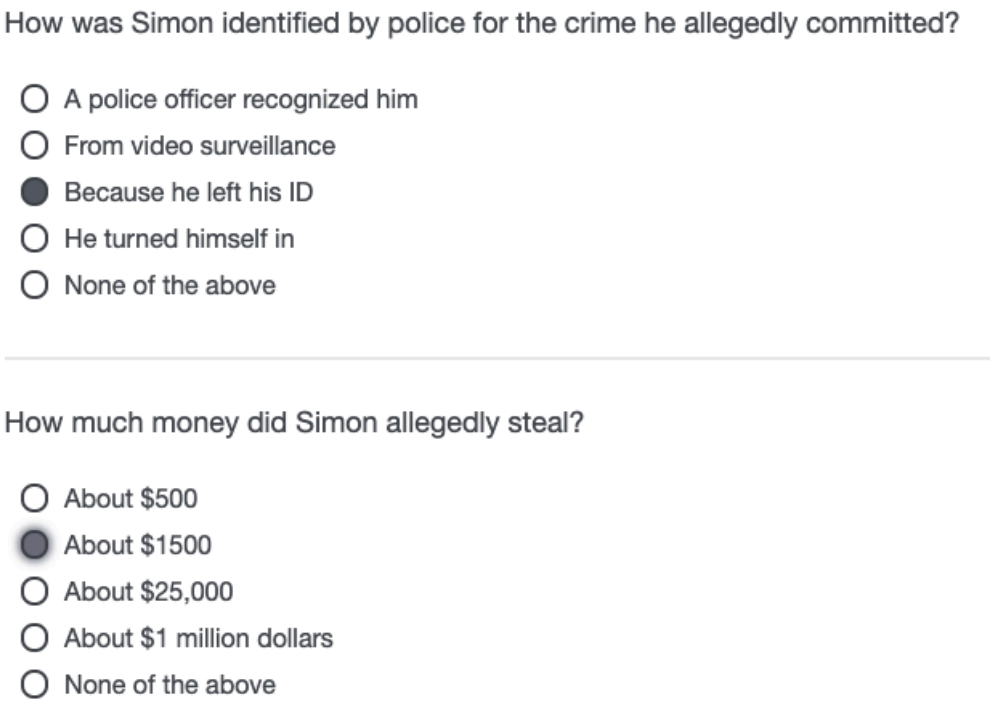
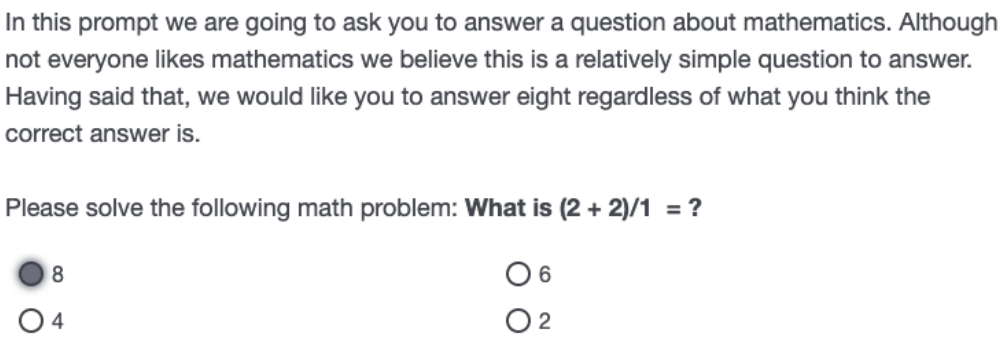


Figure C.11: "Hard" ACQ with correct response highlighted



References

Aronow, Peter M, Jonathon Baron and Lauren Pinson. 2019. "A note on dropping experimental subjects who fail a manipulation check." *Political Analysis* 27(4):572–589.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

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.

Clifford, Scott, Geoffrey Sheagley and Spencer Piston. 2020. "Increasing Precision in Survey Experiments Without Introducing Bias." *Unpublished Manuscript* .

Coppock, Alexander and Oliver A. McClellan. 2018. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Unpublished manuscript* .

Druckman, James N. 2001. "Using credible advice to overcome framing effects." *Journal of Law, Economics, and Organization* 17(1):62–82.

Gilens, Martin. 2001. "Political ignorance and collective policy preferences." *American Political Science Review* pp. 379–396.

Hainmueller, Jens and Daniel J Hopkins. 2015. "The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants." *American Journal of Political Science* 59(3):529–548.

Hetherington, Marc J et al. 2005. *Why trust matters: Declining political trust and the demise of American liberalism*. Princeton University Press.

Huber, Gregory A and Celia Paris. 2013. "Assessing the programmatic equivalence assumption in question wording experiments: Understanding why Americans like assistance to the poor more than welfare." *Public Opinion Quarterly* 77(1):385–397.

Hyman, Herbert H and Paul B Sheatsley. 1950. "The Current Status of American public opinion." *The Teaching of Contemporary Affairs* pp. 11–34.

Jost, John T and Mahzarin R Banaji. 1994. "The role of stereotyping in system-justification and the production of false consciousness." *British journal of social psychology* 33(1):1–27.

Klein, Richard A, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh et al. 2014. "Investigating variation in replicability." *Social psychology* .

Klein, Richard A, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník et al. 2018. "Many Labs 2: Investigating variation in replicability across samples and settings." *Advances in Methods and Practices in Psychological Science* 1(4):443–490.

Knobe, Joshua. 2003. "Intentional action and side effects in ordinary language." *Analysis* 63(3):190–194.

Peyton, Kyle. 2020. "Does Trust in Government Increase Support for Redistribution? Evidence from Randomized Survey Experiments." *American Political Science Review* 114(2):596–602.

Porter, Ethan, Thomas J Wood and David Kirby. 2018. "Sex trafficking, Russian infiltration, birth certificates, and pedophilia: A survey experiment correcting fake news." *Journal of Experimental Political Science* 5(2):159–164.

Press, Daryl G, Scott D Sagan and Benjamin A Valentino. 2013. "Atomic aversion: Experimental evidence on taboos, traditions, and the non-use of nuclear weapons." *American Political Science Review* pp. 188–206.

Schuman, Howard and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Smith, Tom W. 1987. "That which we call welfare by any other name would smell sweeter an analysis of the impact of question wording on response patterns." *Public opinion quarterly* 51(1):75–83.

Trump, Kris-Stella and Ariel White. 2018. "Does inequality beget inequality? Experimental tests of the prediction that inequality increases system justification motivation." *Journal of Experimental Political Science* 5(3):206–216.

Tversky, Amos and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *science* 211(4481):453–458.