The Generalizability of Online Experiments Conducted During The COVID-19 Pandemic

Kyle Peyton, Gregory A. Huber, and Alexander Coppock *

November 28, 2020

Abstract

The disruptions of the COVID-19 pandemic led many social scientists toward online survey experimentation for empirical research. Generalizing from the experiments conducted during a period of persistent crisis may be challenging due to changes in who participates in online survey research and how the participants respond to treatments. We investigate the generalizability of COVID-era survey experiments with 33 replications of 12 pre-pandemic designs fielded across 13 surveys on American survey respondents obtained from Lucid between March and July of 2020. We find strong evidence that these experiments replicate in terms of sign and significance, but at somewhat reduced magnitudes that are possibly explained by increased inattentiveness. These findings mitigate concerns about the generalizability of online research during this period. The pandemic does not appear to have fundamentally changed how subjects respond to treatments, provided they pay attention to treatments and outcome questions. In this light, we offer some suggestions for renewed care in the design, analysis, and interpretation of experiments conducted during the pandemic.

*Kyle Peyton is Postdoctoral Fellow in Law and Social Science, Yale Law School (kyle.peyton@yale.edu); Gregory A. Huber is Forst Family Professor of Political Science, Yale University (gregory.huber@yale.edu); Alexander Coppock is Assistant Professor of Political Science, Yale University (alex.coppock@yale.edu). Thanks to the Center for the Study of American Politics and the Institution for Social and Policy Studies for research support, to Peter Aronow, Josh Kalla, Lilla Orr, and John Ternovski for helpful comments and feedback, to Ethan Porter for sharing replication materials, and to Antonio Arechar, Matt Graham, Patrick Tucker, Chloe Wittenberg, and Baobao Zhang for sharing pre-COVID survey data. Previous versions of the manuscript were presented and benefited from feedback at Université de Montréal and Australian National University. This research was approved by the Yale University Institutional Review Board (Protocol number 1312013102).

During the COVID-19 pandemic, social scientists across the globe have been forced to abandon or postpone research projects that require face-to-face interactions, travel, or even simply leaving the house. Survey experimentation is a form of research that can continue despite these restrictions, because treatments can be delivered and outcomes collected within the same online session. As a result, scholars across many fields including political science responded to the disruption in their work by embracing survey experiments. A natural question is whether these experiments will yield generalizable knowledge due to the extraordinary times. Munger (2020) emphasizes that "temporal validity" is an important feature of external validity. In particular, critics may be concerned that extrapolations from studies conducted during the pandemic to "normal" times will be misleading.

Why might such studies be temporally invalid, or more precisely, why would the effects of treatment during this period be different from the effects of the same treatments deployed earlier or later? The pandemic is an unprecedented event. Many people have fallen ill, died, or lost loved ones, and nearly all people have experienced restrictions on their movement and patterns of normal social interactions. Two important factors relevant for survey experimentation in particular may have changed. First, the same individuals may respond differently to the same treatment during pandemic and non-pandemic periods. For example, survey respondents may be in a heightened state of anxiety over their health, their economic conditions, and national politics. Anxiety and other emotional state variables have been shown to affect information processing (Gadarian and Albertson, 2014), willingness to dissent (Young, 2019), and even to condition the effects of other treatments (Valentino et al., 2009). Second, even if responses to treatment are not different, the types of people participating in these experiment may have changed. For example, extreme changes to how people spend their time (including working from home and job loss) may mean that the set of people willing to take online surveys may have shifted substantially. Another possibility is that increased demand by survey researchers (academic and otherwise) may have put new pressures on markets for survey respondents, leading to respondent "fatigue" or the recruitment of a vastly different set of survey respondents.

The "UTOS" framework for reasoning about how well experimental results obtained on a sample may generalize to a target population considers the extent to which the units, treatments, outcomes, and setting of the experiment correspond to the target population (Cronbach and Shapiro, 1982). This framework is commonly invoked when comparing the results from experiments conducted on representative national samples to online convenience samples. These studies ask – holding treatments, outcomes, and settings fixed – do our

conclusions depend on who the units are? The summary conclusion appears to be "No." Berinsky, Huber and Lenz (2012), Mullinix et al. (2015), Coppock (2019), Coppock, Leeper and Mullinix (2018) and Coppock and McClellan (2018) all find strong correspondences between survey experimental results obtained on national probability samples and online convenience samples like Amazon's Mechanical Turk and Lucid. The question of temporal validity fits easily in the UTOS framework: Holding treatments and outcomes constant, do the pandemic-induced changes in context and subject pool composition alter the conclusions drawn from survey experiments?

Empirically demonstrating that experiments conducted during COVID do or do not generalize to other times is in principle quite straightforward. Once the pandemic has passed and the social, economic, and political aftershocks have dissipated, replications of COVID-era studies can settle the question of whether those results are similar or different. Unfortunately, it may be a while before normal times resume. In this paper, we take up a closely-related question that we can answer much sooner: do experiments conducted prior to the pandemic generalize to COVID times? We conduct 33 replication studies of 12 previously published survey experiments. In line with previous replication attempts in political science, we find good correspondence across study versions. Our replication estimates nearly always agree with original studies in terms of sign and significance but tend to be somewhat smaller.

What explains these smaller effect sizes? As has been noted for subjects recruited using the Mechanical Turk platform (e.g. Arechar and Rand, 2020) and also using the Lucid platform we employ (e.g. Aronow et al., 2020), respondent attentiveness during online surveys appears to have declined by Spring 2020, and possibly sooner. As is well known, *random* measurement error attenuates correlations toward zero (e.g., Spearman, 1904), but the measurement error induced by inattention is not random, so the inter-correlations across covariates and outcomes (i.e., the descriptive relationships) could be overstated or understated. However, we show below that under the assumption that inattentive answers do not change depending on the randomized treatment assignment, inattention will uniformly shrink treatment effect estimates towards zero (see also Berinsky, Margolis and Sances, 2014). We further show that under this same assumption, we can obtain estimates of the average treatment effect among the attentive by dividing the intention-to-treat estimate by the attentiveness rate. This procedure re-inflates estimates deflated by inattention. The substantive implication of this procedure is that replication estimates come even closer to original estimates.

Overall, our findings should give experimentalists and their critics greater confidence that

survey experiments conducted during the pandemic should generalize to other times and contexts. First, setting aside inattentiveness, the sign and significance of treatment effects appear consistent with pre-COVID 19 estimates, implying there are no wholesale changes in responses to these interventions. Second, increases in inattention will attenuate treatment effects towards zero. This problem can be addressed either by screening out inattentive subjects or by adjusting estimates by the attention rate. Overall, we find that treatments generate similar effects during the pandemic as before, though we discuss possible caveats to this finding in the concluding section of the paper.

1 Design

We recruited weekly samples of approximately 1,000 U.S. based participants via Lucid over 13 weeks during the spring and early summer of 2020. Lucid is an aggregator of online survey respondents from many providers. Importantly, Lucid collects demographic information from all respondents before the enter a survey, enabling them to construct sets of subjects who are quota sampled to US census margins (Coppock and McClellan, 2018). The targeting is close but not perfect – and even if it were perfect, samples collected via Lucid should properly be considered convenience samples, since balance on the demographic marginal distributions does not imply balance on the joint distributions. Moreover, because Lucid does not provide probability samples, the real concern is balance on the (joint) distribution of unobserved and unmeasured variables beyond demographics.

Like previous investigations on the suitability of convenience samples for academic research, we focus on survey experiments in particular. Using online convenience samples for descriptive work is generally inadvisable because the samples may be different from target populations in both observable and unobservable ways. However, for theoretical settings in which an average treatment effect among a diverse (but nevertheless unrepresentative) sample would be informative, survey experiments with online convenience samples can be an effective tool. For this reason, we restrict our focus to the ways in which the pandemic might have changed the treatment effect estimates obtained with online convenience samples.

1.1 Selection Criteria

We conducted 33 replications across 12 unique studies, chosen based on the following criteria:

1. *Suitable for online survey environment.* All replications were administered through a web browser using Qualtrics survey software, and the original studies must have been fielded in a similar format.

2. *Length of study.* Time constraints ruled out studies that could not be administered in a survey module of 3-5 minutes.

3. *Design transparency.* The original study design and outcome measures were clearly described in the published manuscript or supporting appendix.

4. *Design complexity and effect size.* Sample size constraints ruled out some two-arm studies with small effect sizes, as well as more complex studies with elaborate factorial designs.

5. *Theoretically and politically importance.* The studies all considered theoretically important questions, with many being published in top journals.

These criteria are similar to those used in other meta-scientific replication projects (e.g., Klein et al., 2014, 2018). We also aimed for coverage across political science sub-fields, and for a mix of classic and contemporary studies. Our set of studies is not a random sample of all survey experiments conducted by social and political scientists, but they do cover a wide range of designs. The full set of studies is listed in Table 1.

Table 1: Summary of thirty-three replications conducted across twelve original studies

Original study	Experimental design	COVID-era replication	Direct replication	Replicated
Russian reporters and American news (Hyman & Sheatsley, 1950)	Two-arm	Week 3	Yes	Yes
Effect of framing on decision making (Tversky & Kaheneman, 1981)	Two-arm	Week 7	Split sample	Yes
Gain versus loss framing (Tversky & Kaheneman, 1981)	Two-arm	Weeks 1, 3, 7, 8, 13	Week 13 only	Yes
Welfare versus aid to the poor (Smith, 1987)	Two-arm	Weeks 1-9, 11-13	Yes	Yes
Gain vs. loss framing + party endorsements (Druckman, 2001)	Six-arm	Weeks 7, 8, 13	Week 13 only	Yes
Foreign aid misperceptions (Gilens, 2001)	Two-arm	Week 3	Yes	No
Perceived intentionality for side effects (Knobe, 2003)	Two-arm	Week 7	Split sample	Yes
Atomic aversion (Press, Sagan, & Valentino, 2013)	Five-arm	Weeks 5, 6, 13	Week 13 only	Partial
Attitudes towards immigrants (Hainmueller & Hopkins, 2015)	Factorial (conjoint)	Week 8	Yes	Yes
Fake news corrections (Porter, Wood, & Kirby, 2018)	Mixed factorial (2x6)	Week 4	Yes	Yes
Inequality and system justification (Trump & White, 2018)	Two-arm	Week 2	Yes	Yes
Trust in government and redistribution (Peyton, 2020)	Three-arm	Week 9	Yes	Yes

2 Results

We present our summary results for all 12 studies in the main text, and describe each study in detail Section A of the online appendix. For all pre-COVID studies, we were able to obtain the original effect size(s) and, for most studies, we were able also able to obtain at least one replication of the original effect size(s). In total, we obtained 102 replication estimates and 97 pre-COVID estimates, across 12 unique studies. To provide an overall summary, we calculate (pooled) summary estimates for each set of replication studies and their pre-COVID benchmarks. For each treatment-outcome pair, we calculate the summary effect size using the precision-weighted average. For studies with one outcome and a simple experimental design, we compute a single difference; and for studies with multiple outcomes and/or treatments, we compute a difference for each unique treatment-outcome pair. Except for the conjoint experiment, all within-study estimates (binary and ordinal) are standardized using Glass's Δ, which scales outcomes by the standard deviation in the control group (Glass, 1976).

Across the 12 studies, we obtained 138 summary effect size estimates (82 for the conjoint and 56 for the remaining studies). Figure 1 compares the 28 estimated summary effects from the pre-COVID studies (horizontal axis) with their 28 replications (vertical axis). All replication summary estimates were smaller in magnitude than their pre-COVID estimates, with 24 of 28 signed in the same direction. Of the 24 correctly signed estimates, 10 were significantly smaller in replication. Of the 4 incorrectly signed estimates, 3 were significantly different – the foreign aid misperceptions study and 2 of 6 estimates from the atomic aversion study. Figure 2 plots the analogous information for the 41 conjoint estimates and their 41 replications, all of which are signed in the same direction. Of these, 35 of 41 were smaller in replication (11 of 35 significant differences) and 6 of 41 were larger in replication (1 of 6 significant differences).

Pooling across the 65 of 69 correctly signed pairs presented in Figures 1-2, the replication estimates were, on average, 74% as large as the pre-COVID estimates. We present the individual replication estimates for each of the twelve studies, alongside their pre-COVID estimates, in Appendix A. Within each study, estimates on non-binary outcome measures are standardized using Glass's Δ. For binary outcome measures, we report the unstandardized estimates as these have a straightforward interpretation as the percentage point difference between the reference category and the level(s) of treatment.

Figure 1: Comparison of 28 summary effect sizes across 11 studies (conjoint excluded)

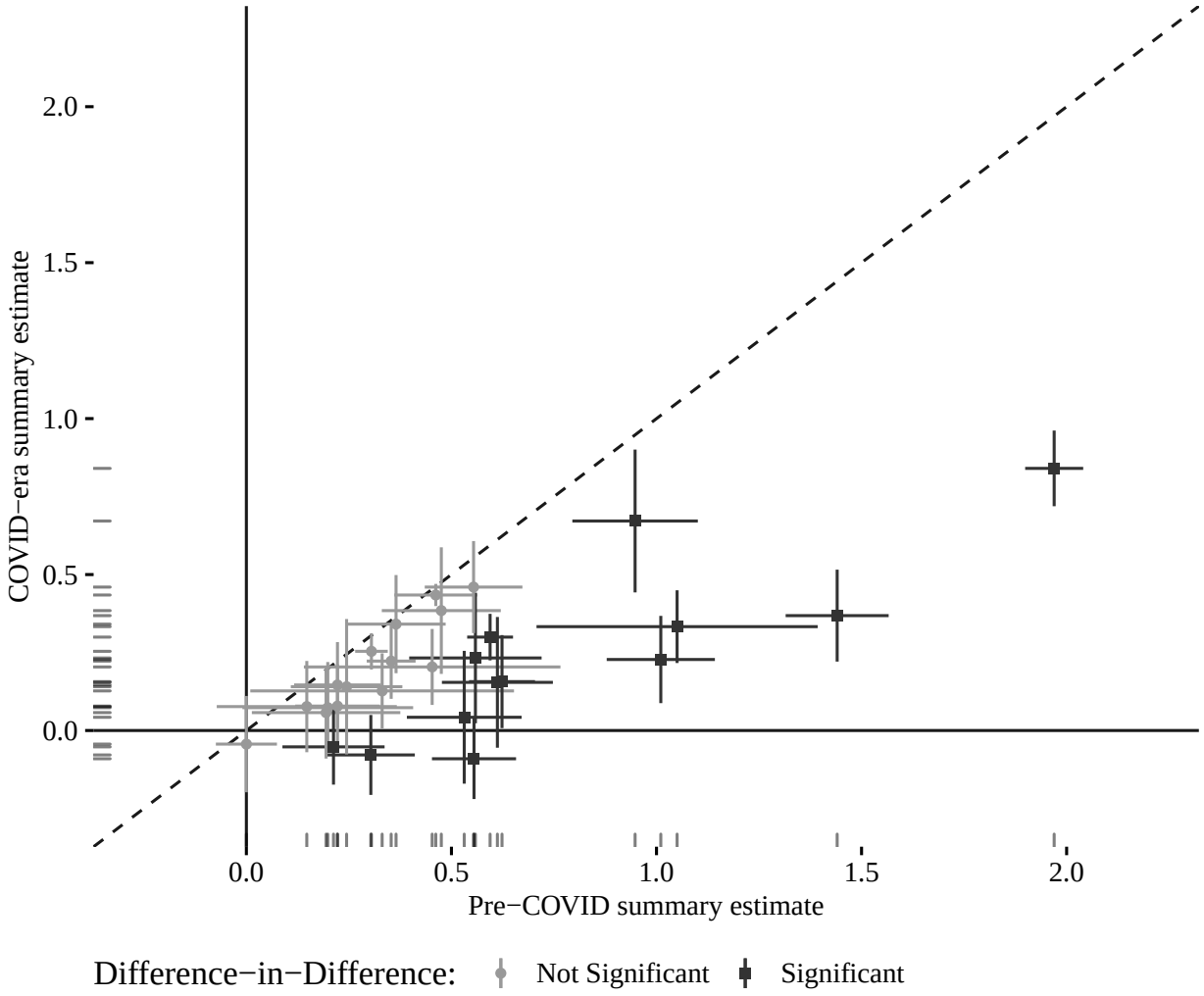
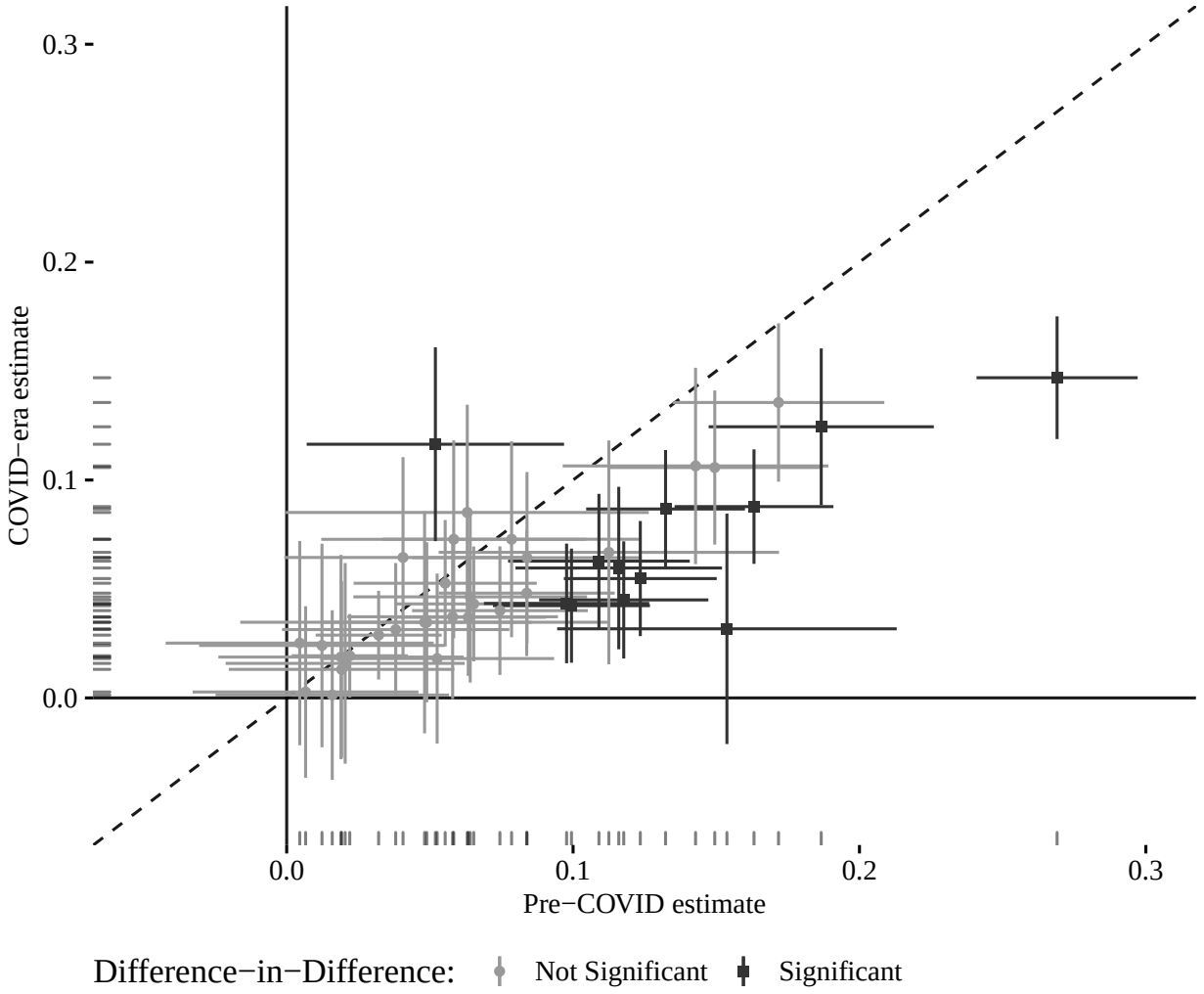


Figure 2: Comparison of 41 summary effect sizes in conjoint experiment



3 Inattentiveness in online samples

For the most part, survey experiments conducted during COVID replicated the pre-pandemic results in terms of sign and significance, but not magnitude. Our hunch is that the culprit is an increase in the fraction of inattentive respondents in the subject pool. Inattention leads to measurement error in self-administered surveys (Berinsky, Margolis and Sances, 2014, 2016; Berinsky et al., 2019): some subjects read and answer questions quickly while others take their time. We can sympathize with survey takers who rush – perhaps the questions are uninteresting or they have other things on their mind. Recent work has documented a decline attentiveness on popular online platforms since at least January 2020, including both

Lucid (Aronow et al., 2020) and MTurk (Arechar and Rand, 2020).

The main problem is that inattentive subjects provide survey responses, but these survey responses are likely unaffected by the treatment they are assigned to receive because they do not meaningfully engage with experimental stimuli. Moreover, we do not know the extent of the inattention problem. In addition to ignoring treatment information, inattentive subjects may also ignore the details of the questions used to measure outcomes. As a result, we do not know if their responses reflect their untreated state or reflect nothing more than haphazard clicking. In either case, we cannot learn the effect of treatment among inattentive subjects. Fortunately, under the assumption that those who are inattentive provide responses that are unaffected by treatment assignment, we can estimate the treatment effect among those who are attentive, as explained below.

3.1 Causal inference with inattentive subjects

In this section, we formalize the attentiveness problem in the potential outcomes framework. Let $Z_i \in \{0, 1\}$ denote a subject's treatment assignment (0 for control, 1 for treatment), $A_i \in \{0, 1\}$ denote their attentiveness (0 if inattentive, 1 if attentive), and $Y_i \in \mathbb{R}$ denote their observed outcome.[1] We make the following assumptions:

A1 (Inattentive Potential Outcomes)	$Y_i = \begin{cases} Y_i(0) & \text{if } Z_i = 1, A_i = 1; \\ Y_i(1) & \text{if } Z_i = 0, A_i = 1; \\ Y_i(\mathcal{I}) & \text{if } Z_i = 1, A_i = 0; \\ Y_i(\mathcal{I}) & \text{if } Z_i = 0, A_i = 0. \end{cases}$
A2 (Independence)	$(Y_i(0), Y_i(1), Y_i(\mathcal{I}), A_i) \perp\!\!\!\perp Z_i$
A3 (Non-zero Attentiveness)	$\Pr(A_i = 1) > 0$

A1 states that attentive subjects in the treatment condition reveal their treated potential outcome $Y_i(1)$ and attentive subjects in the control condition reveal their untreated potential outcome $Y_i(0)$. Inattentive subjects, however, reveal their "inattentive" potential outcome $Y_i(\mathcal{I})$ *regardless* of treatment assignment.

A2 states that the joint distribution of potential outcomes and attentiveness are independent of treatment assignment. This assumption is satisfied by random assignment of Z_i. A3 states there are at least some attentive subjects in the sample who can be identified from

[1] We present the case of a binary treatment for ease of exposition, but this framework can be generalized to situations with multiple levels of treatment and/or attentiveness.

the observed data if a pre-treatment measure of attentiveness is available for all subjects.

Under assumptions A1-A3, we can define the following estimands:

$$
\begin{aligned}
\text{CATE} \mid \text{Attentive} &= \mathbb{E}[Y_i(1) - Y_i(0) | A_i = 1] && (1)\\
&= \mathbb{E}[Y_i(1) | Z_i = 1, A_i = 1] - \mathbb{E}[Y_i(0) | Z_i = 0, A_i = 1]\\
&= \mathbb{E}[Y_i | Z_i = 1, A_i = 1] - \mathbb{E}[Y_i | Z_i = 0, A_i = 1]\\
\text{CATE} \mid \text{Inattentive} &= \mathbb{E}[Y_i(\mathcal{I}) - Y_i(\mathcal{I}) | A_i = 0] && (2)\\
&= \mathbb{E}[Y_i(\mathcal{I}) | Z_i = 1, A_i = 0] - \mathbb{E}[Y_i(\mathcal{I}) | Z_i = 0, A_i = 0]\\
&= \mathbb{E}[Y_i | Z_i = 1, A_i = 0] - \mathbb{E}[Y_i | Z_i = 0, A_i = 0]\\
&= 0\\
\text{Intention-to-Treat (ITT)} &= \mathbb{E}[Y_i^*(1) - Y_i^*(0)] && (3)\\
&= \big\{\mathbb{E}[Y_i(1) | Z_i = 1, A_i = 1] \Pr(A_i = 1 | Z_i = 1)\\
&\quad + \mathbb{E}[Y_i(\mathcal{I}) | Z_i = 1, A_i = 0] \Pr(A_i = 0 | Z_i = 1) \big\}\\
&\quad - \big\{\mathbb{E}[Y_i(0) | Z_i = 0, A_i = 1] \Pr(A_i = 1 | Z_i = 0)\\
&\quad + \mathbb{E}[Y_i(\mathcal{I}) | Z_i = 0, A_i = 0] \Pr(A_i = 0 | Z_i = 0) \big\}\\
&= \text{CATE} \mid \text{Attentive} * \Pr(A_i = 1) + \text{CATE} \mid \text{Inattentive} * \Pr(A_i = 0)\\
&= \text{CATE} \mid \text{Attentive} * \Pr(A_i = 1)
\end{aligned}
$$

Under assumptions A1-A3, the Intention-to-Treat (ITT) effect is a weighted average of the conditional average treatment effect (CATE) among the attentive and the CATE among the inattentive (3), where the weights are the sample proportions of attentive and inattentive subjects. Since under assumption A1, the CATE among the inattentive is equal to zero, the ITT must be closer to zero than the CATE among the attentive, because $Pr(A_i = 1) \leq 1$. In other words, the measurement error from inattention shrinks treatment effect estimates toward zero.

This formalization also helps clarify one finer point – even if the inattentive could be inducted to pay attention, the CATE among this group need not equal the CATE among those who pay attention without further inducement. Being the kind of person who does not pay attention could be correlated with responses to treatment.

3.2 Estimating the CATE among the attentive

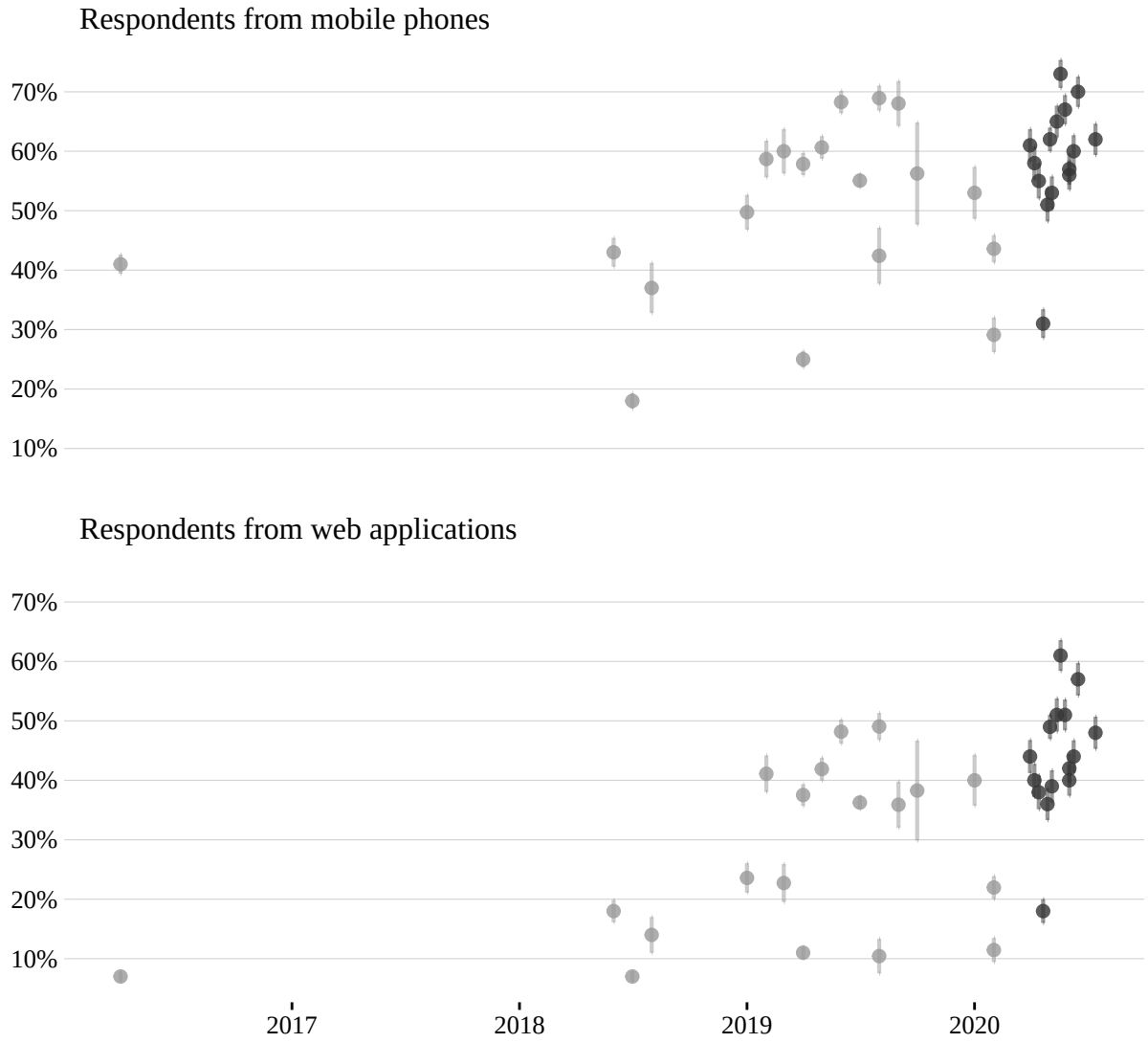
Under the (we think reasonable) assumption that most subjects were paying attention pre-COVID, the ATE estimates from the original studies can be considered CATEs among attentive subjects. Our goal, therefore, is to estimate COVID-era CATEs among the attentive. We explore three approaches. The first and most straightforward way to estimate CATEs among the attentive is simply to include pre-treatment attention check questions (ACQs) that separate the attentive from the inattentive (see Oppenheimer, Meyvis and Davidenko, 2009; Paolacci et al., 2010). Kane, Velez and Barabas (2020) suggest including "mock vignettes" which can be viewed as a task-specific ACQ. Berinsky, Margolis and Sances (2014); Berinsky et al. (2019) urge researchers to use multiple ACQs and classify subjects based on different *levels* of attentiveness. Only two of our surveys include ACQs so we also consider other approaches.

In the absence of an ACQ, we can follow an instrumental variables logic to recover an estimate of the CATE among the attentive. Expression 3 implies that we can estimate the CATE among the attentive by dividing the ITT by an estimate of the attention rate. Again, since we did not measure attention in most of our surveys, we rely on contemporaneous estimates of attentiveness on Lucid during this period. Aronow et al. (2020) report attentiveness rates on the order of 70 to 80%. On average, treatment effect estimates in the fake news studies were 45% the size of the original. Dividing our replication estimates by 0.8 re-inflates our estimates of 57% of the original effect sizes. Our conjoint ACME estimates were on average 87% as big as the originals – re-inflating our estimates returns us to 109% of the original magnitudes. We do not want to push these back-of-the-envelope calculations too far – again, we do not have reliable attentiveness data for our studies in particular, just for Lucid studies conducted around the same time as ours. Nevertheless, the re-inflation procedure does help us to reason about the plausible size of effects among those subjects who do pay attention. We conclude that among the attentive, the pandemic has not meaningfully changed how subjects respond to treatments.

A third approach is to rely on subject-level metadata. Our suspicion is that the large increase in inattention stems from the increase in survey respondents who arrive in the survey environment from a mobile game, so we need a way to identify these subjects. The "User-Agent" string captured from the end user's browser by online survey software like Qualtrics provides detailed information about how respondents arrive at the survey. We offer two approaches for classifying respondents as inattentive on the basis of this string: 1) if they come from a web-application rather than a web-browser; 2) if they come from

a mobile phone rather than a tablet or desktop. Applying this approach across the 13 surveys used for the replication studies, we estimate that the proportion of subjects coming from web-applications (rather than internet browsers) ranged from 0.18 to 0.61, and the proportion of subjects coming from mobiles (rather than desktops or tablets) ranged from 0.31 to 0.73. Based on an additional sample of 36,317 individuals obtained from other Lucid surveys fielded prior to March 2020, we observe that the proportion of subjects coming from both web-based applications and mobiles was on the rise prior to the COVID-19 outbreak. On average, 41% of respondents came from web applications (56% from mobiles) in the 2020 surveys, an increase from 33% (56% from mobiles) in 2019 and just 12% (35% from mobiles) prior to 2019. These results, reported in Figure 3, are consistent with the decline in data quality from January 2020 onward documented by Aronow et al. (2020).

Figure 3: Proportion of Lucid samples arriving from mobile devices and web applications



Notes: Estimates come from a combination of surveys conducted by the authors and UserAgent data shared by Antonio Arechar, Matt Graham, Patrick Tucker, Chloe Wittenberg, and Baobao Zhang.

Overall, 97% of our subjects coming from web-applications were on mobile devices, and 72% of subjects on mobile devices arrived from web-applications. Pooling across the 13 surveys used for the replication studies, we find that subjects from web-applications on average spent roughly 7 minutes less time completing surveys than subjects from web browsers, who spent an average of 21.5 minutes. Subjects from mobile devices spent roughly 6 minutes less time completing surveys than subjects from non-mobile devices. Additionally, in two of

the 13 surveys we included ACQs of varying difficulty and found those on web-applications (or mobiles) were significantly less likely to pass the ACQs, compared to those coming from browsers (or non-mobiles). These results are reported in Table 2.

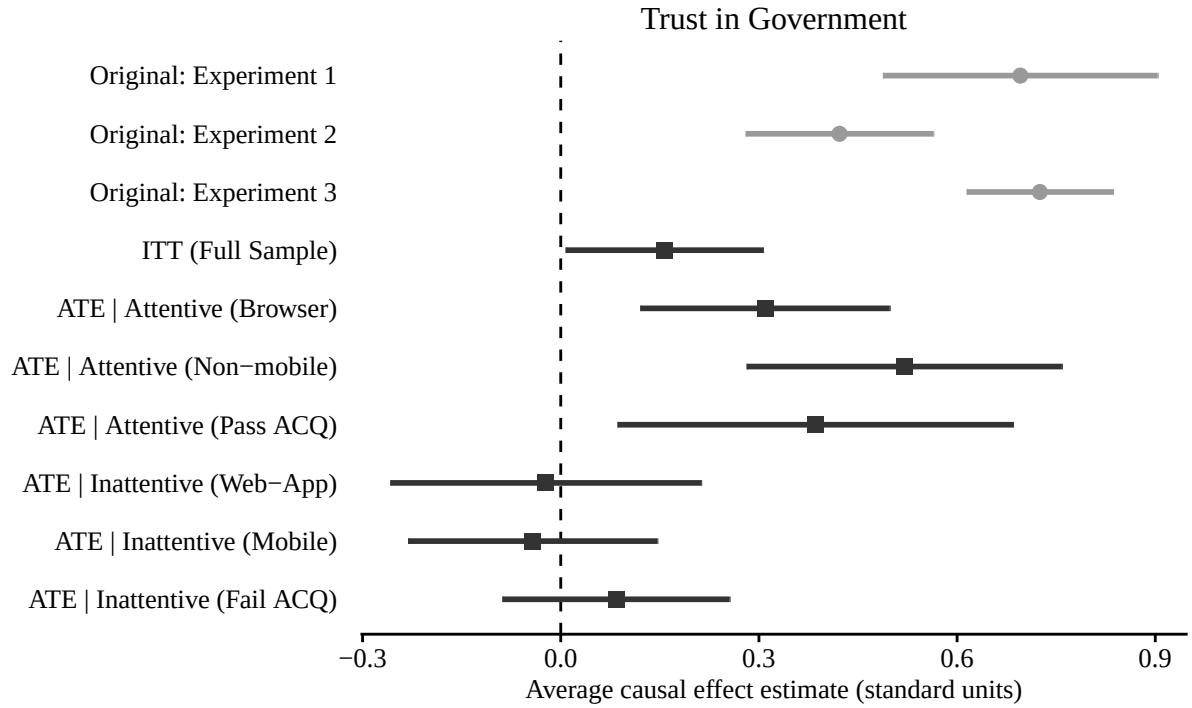
Table 2: ACQ pass rates by attentiveness and level of difficulty

Difficulty	Browser	Web-App	Difference	Mobile	Non-Mobile	Difference
Easy	0.66 (0.02)	0.54 (0.02)	-0.13 (0.03)*	0.58 (0.02)	0.63 (0.02)	-0.05 (0.03)
Medium	0.53 (0.02)	0.36 (0.02)	-0.17 (0.03)*	0.42 (0.02)	0.50 (0.02)	-0.08 (0.03)*
Hard	0.25 (0.02)	0.19 (0.02)	-0.06 (0.02)*	0.19 (0.01)	0.27 (0.02)	-0.08 (0.03)*

Notes: The "Easy" and "Medium" questions were novel ACQs that subjects completed after reading a short news article (see Figure A.X-A.X). The "Hard" ACQ comes from Peyton (2020) and was included with the direct replication of the original study in Week 9. This ACQ, is analogous to an "Instructional Manipulation Check" (see Oppenheimer, Meyvis and Davidenko, 2009) and was passed by 87% of respondents in the original study (see Figure A.X). Standard errors in parentheses. $P < 0.05^*$.

We now illustrate the utility of both the ACQ and metadata approaches. While most of our surveys did not include pre-treatment ACQs, we did include one immediately prior to the direct replication of Peyton (2020). Figure 4 shows that, for the replication of Peyton (2020), estimates of the CATE among those classified as attentive on trust in government using metadata classification are similar in magnitude to those using the attention check classification. Estimates among those coming from a browser are 2.4 times larger than the ITT, and estimates among those coming from non-mobile devices are 3.3 times larger than the ITT. The estimated CATE among those passing the ACQ are 1.8 times larger than the ITT. Estimates of the CATE among those classified as inattentive are indistinguishable from zero regardless of which approach is used. Finally, as Berinsky, Margolis and Sances (2014) show, those who fail ACQs can differ markedly from those who pass on observed characteristics. We find that respondents' metadata is also correlated with covariates; for example, Whites and Republicans were less likely to take surveys on web applications and mobile phones (see Appendix B for differences in background covariates).

Figure 4: Reanalysis: trust in government and redistribution



4 Discussion

In this paper, we have presented the results of 33 replications of 12 studies conducted on Lucid during the early COVID era. We conducted 13 surveys of approximately 1,000 subjects each during a period of massive disruption to life and work across the globe. The main purpose of this investigation was to understand whether survey experimental results obtained during this period of upheaval will generalize to other times, that is, we sought to understand what Munger (2020) calls the "temporal validity" of COVID-era survey experiments.

We considered two main ways in which treatment effects might have been different during this period. First, it is possible that the pandemic and its rippling effects on all aspects of life changed people's political preferences, their states of mind, and their patterns of information processing. If so, then the same experiment conducted on the same group of people before and during the pandemic might generate different conclusions. We do not find support for this idea. Second, it's possible that the pandemic changed *who* takes surveys, or at least that the population of survey takers changed between the time of the original studies and our replications. We think that the increases in inattention documented by others (Arechar

and Rand, 2020; Aronow et al., 2020) provide some indication that the composition of online convenience samples has changed or is changing. In particular, we think that there has been an increase in inattentive survey subjects.

We could in principle want to learn about causal effects among these inattentive subjects. However, we *cannot* estimate treatment effects among this group because they fail to take treatment and because they may also fail to report their outcomes.[2] The immediate implication of all this is that survey researchers should include pre-treatment attention checks (Berinsky, Margolis and Sances, 2014; Permut, Fisher and Oppenheimer, 2019). Post-treatment attention checks can induce bias (Montgomery, Nyhan and Torres, 2018; Aronow, Baron and Pinson, 2019), but pre-treatment attention checks allow researchers to either terminate the interview without paying for the bad data[3] or to condition those subjects out in the analysis. Otherwise, we can reason about the the effects of treatment among the attentive by dividing by the attention rate in contemporaneous studies or by investigating subject metadata, but it is preferable to condition on attention in the design stage rather than adjust during the analysis stage.

In conclusion, the overarching implication of our analyses is that the pandemic did not change subjects' internal states in ways that interacted with our experimental treatments. People still take riskier options when in a loss frame – even when the real-world analogue of the framing experiment could hardly be more salient. People still believe misinformation but can be corrected by fact checks. People still have preferences over immigrants that can be measured via a conjoint experiment. People still prefer funding "aid to the poor" to "welfare." Even in extraordinary times, people still exhibit ordinary political attitudes and behaviors.

[2]It would of course also be desirable to design interventions that manipulated attention, so that we could understand whether treatment effects among those who are at baseline inattentive are similar to those who are otherwise attentive. That said, prior research has shown that inducing attentiveness among inattentive subjects is challenging (Berinsky, Margolis and Sances, 2016). In a separate Lucid study, we randomly assigned whether those who failed an initial attention check were told that they failed and given a second chance to pass. Only 17 of the 410 subjects in the treatment group passed when specifically reminded to re-read the prompt carefully because they had missed something.

[3]For example, in a separate Lucid study fielded on October 29th we filtered out respondents who failed an ACQ at the beginning of the survey (the "Easy" ACQ from Table 2). Immediately prior to this ACQ, we also asked respondents to self-report whether they were recruited to the survey from an online game, and 51% reported they were. We observed substantial differences in pass rates: among those coming from an online game, only 38% passed the ACQ, versus 82% of those not coming from a game.

References

Arechar, Antonio A and David Rand. 2020. "Turking in the time of COVID." *PsyArXiv* .

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.

Aronow, Peter Michael, Joshua Kalla, Lilla Orr and John Ternovski. 2020. "Evidence of Rising Rates of Inattentiveness on Lucid in 2020." *SocArXiv* .

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.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2016. "Can we turn shirkers into workers?" *Journal of Experimental Social Psychology* 66:20–28.

Berinsky, Adam J, Michele F Margolis, Michael W Sances and Christopher Warshaw. 2019. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?" *Political Science Research and Methods* pp. 1–8.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3):613–628.

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* .

Coppock, Alexander, Thomas J. Leeper and Kevin J. Mullinix. 2018. "Generalizability of heterogeneous treatment effect estimates across samples." *Proceedings of the National Academy of Sciences* 115(49):12441–12446.

Cronbach, Lee J. and Karen. Shapiro. 1982. *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Gadarian, Shana Kushner and Bethany Albertson. 2014. "Anxiety, Immigration, and the Search for Information." *Political Psychology* 35(2):133–164.

Glass, Gene V. 1976. "Primary, secondary, and meta-analysis of research." *Educational researcher* 5(10):3–8.

Kane, John V., Yamil R. Velez and Jason Barabas. 2020. "Analyze the Attentive and Bypass Bias: Mock Vignette Checks in Survey Experiments." *Unpublished Manuscript* .

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.

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2:109–138.

Munger, Kevin. 2020. "Knowledge Decays: Temporal Validity and Social Science in a Changing World." *Unpublished Manuscript* .

Oppenheimer, Daniel M, Tom Meyvis and Nicolas Davidenko. 2009. "Instructional manipulation checks: Detecting satisficing to increase statistical power." *Journal of experimental social psychology* 45(4):867–872.

Paolacci, Gabriele, Jesse Chandler, Panagiotis G Ipeirotis et al. 2010. "Running experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5):411–419.

Permut, Stephanie, Matthew Fisher and Daniel M Oppenheimer. 2019. "Taskmaster: A tool for determining when subjects are on task." *Advances in Methods and Practices in Psychological Science* 2(2):188–196.

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

Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15(1):72–101.

Valentino, Nicholas A., Antoine J. Banks, Vincent L. Hutchings and Anne K. Davis. 2009. "Selective Exposure in the Internet Age: The Interaction between Anxiety and Information Utility." *Political Psychology* 30(4):591–613.

Young, Lauren E. 2019. "The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe." *American Political Science Review* 113(1):140–155.

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

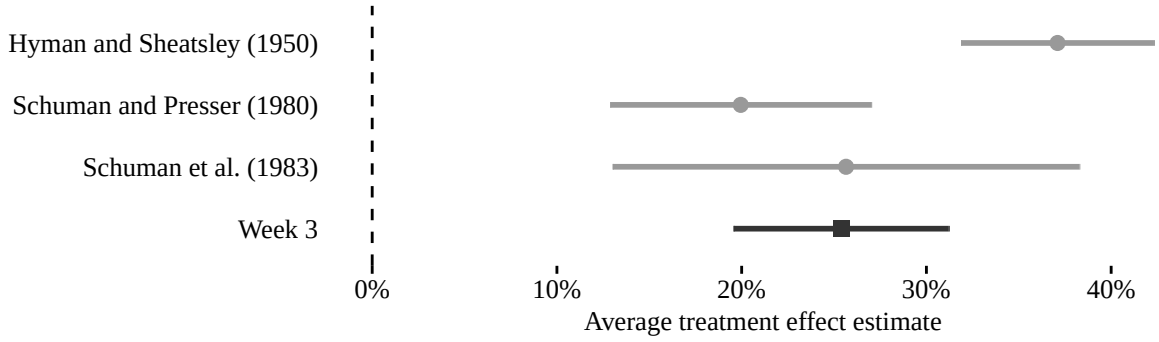
A Individual Studies

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.

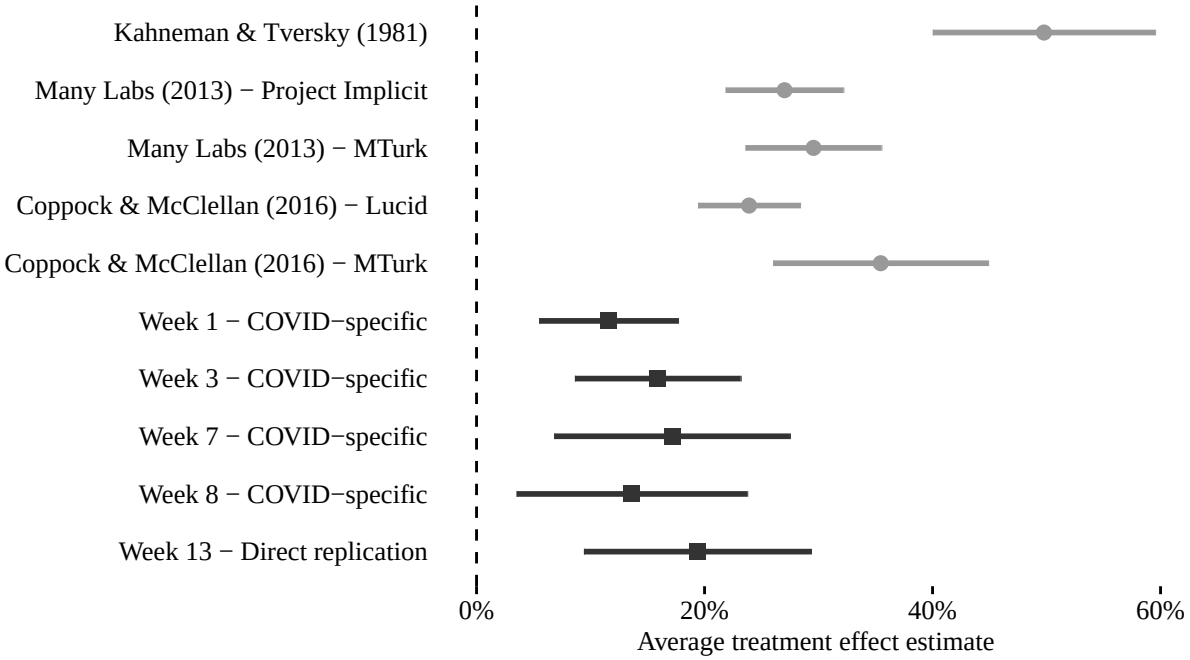
In both framing conditions, the expected number of deaths is 400 for both Program A and Program B. In the original study, 72% selected Program A in the gain frame, whereas 22% selected Program A in the loss frame, for an estimated treatment effect of 50 percentage points. According to Tversky and Kahneman (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?"

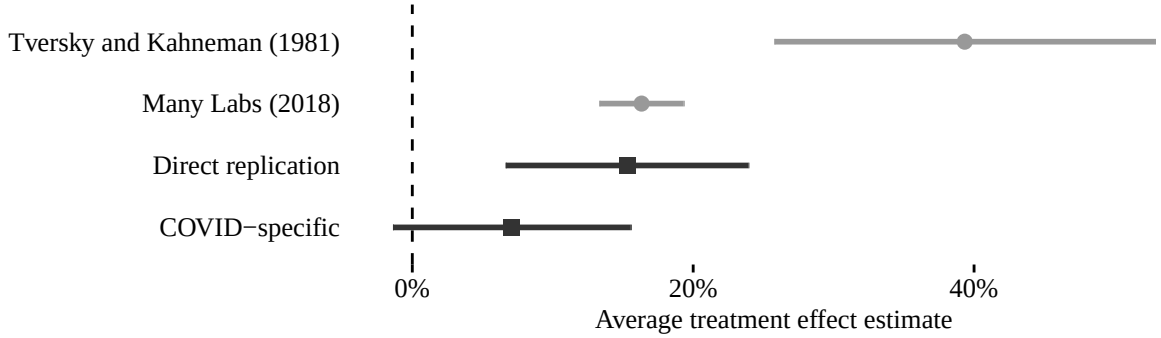
In the replication by Klein et al. (2018), 49% of participants said they would travel to save $10 on the "cheap" wall-hanging whereas 32% said they would travel to save $10 on the "expensive" wall-hanging. In our Week 7 replication study, half the participants were assigned to an experiment using this 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

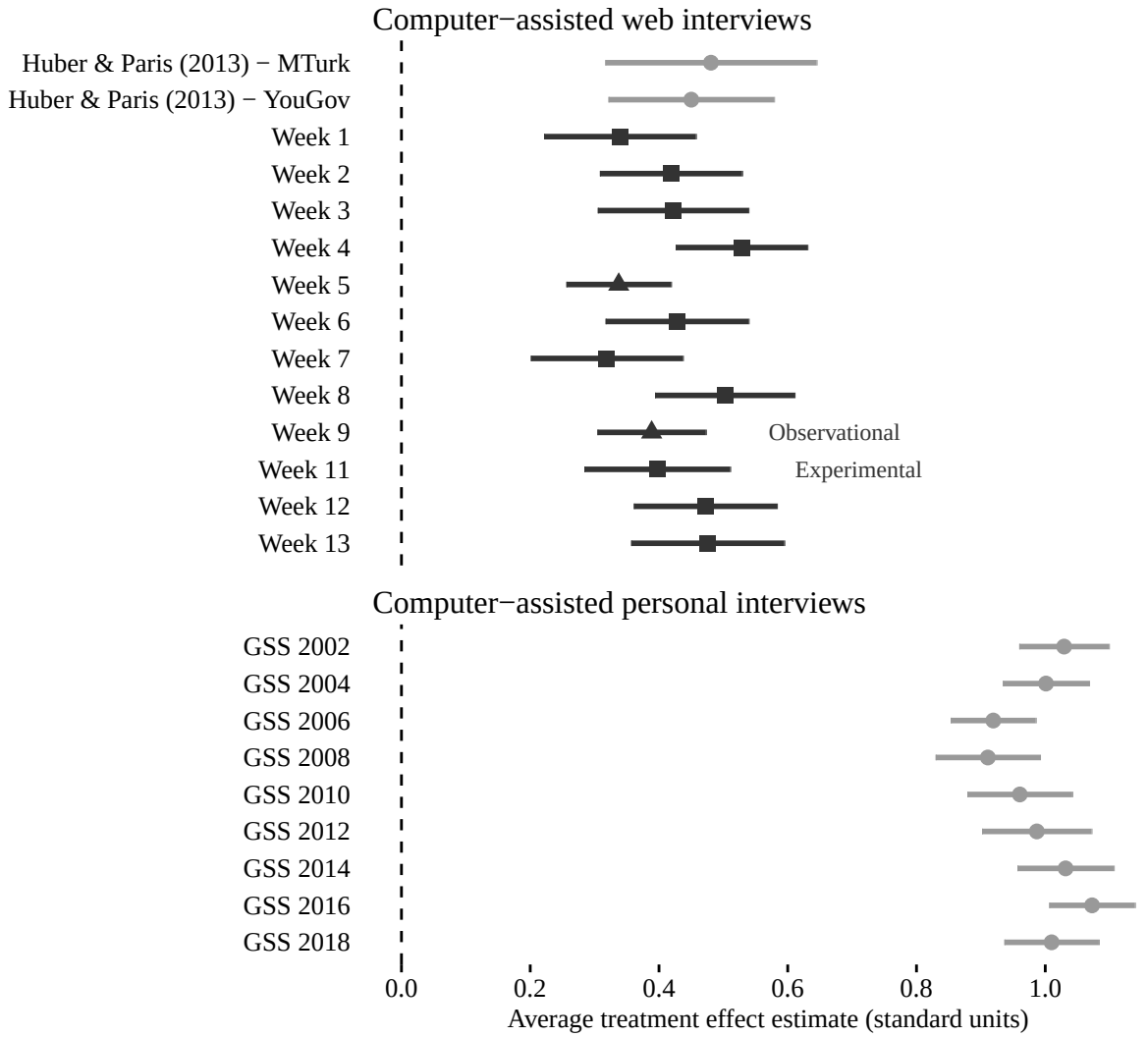
poor" than "welfare". [1]

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

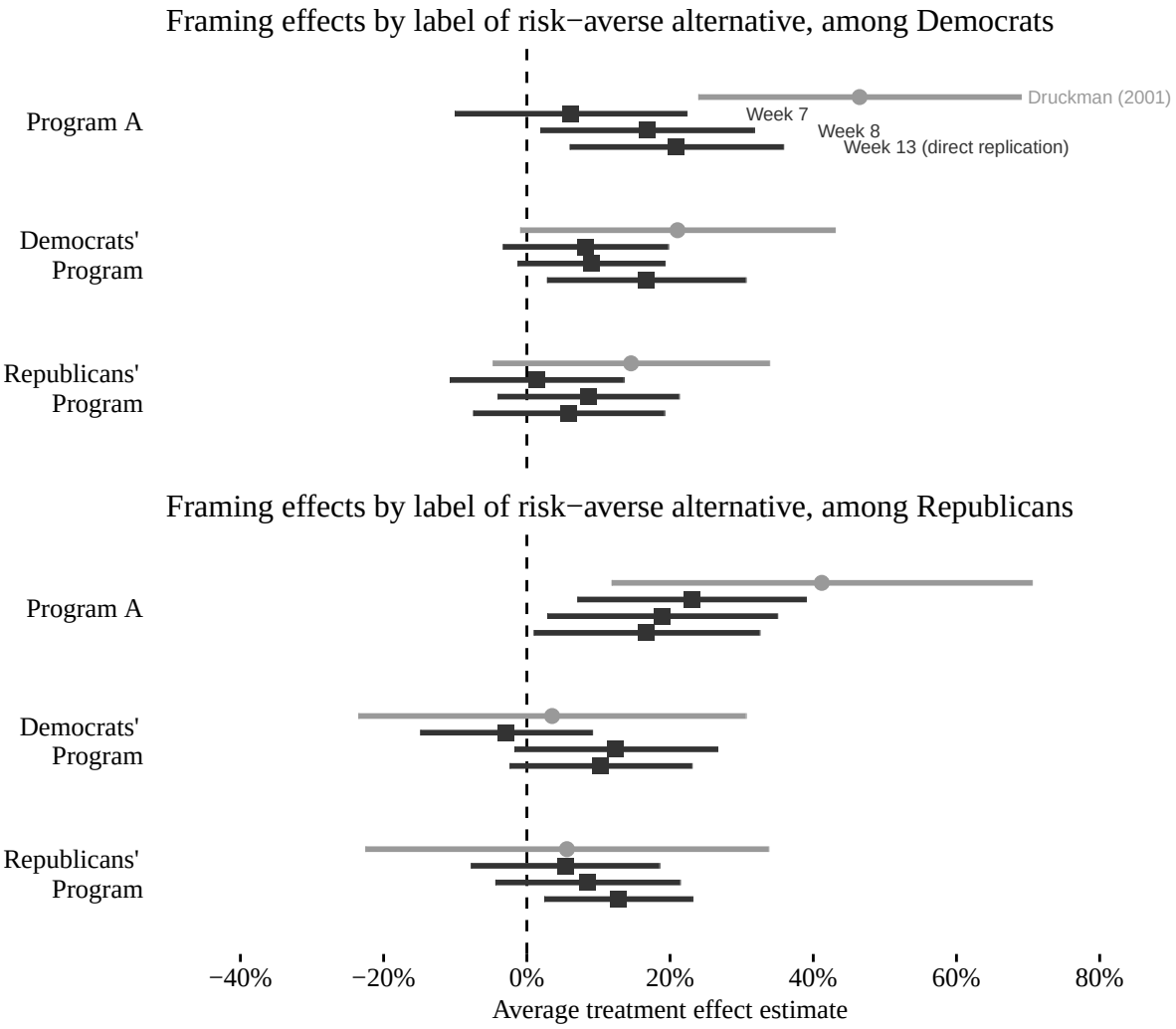


TABLE 1: Summary of effect sizes in "Asian disease" problem with party endorsement

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

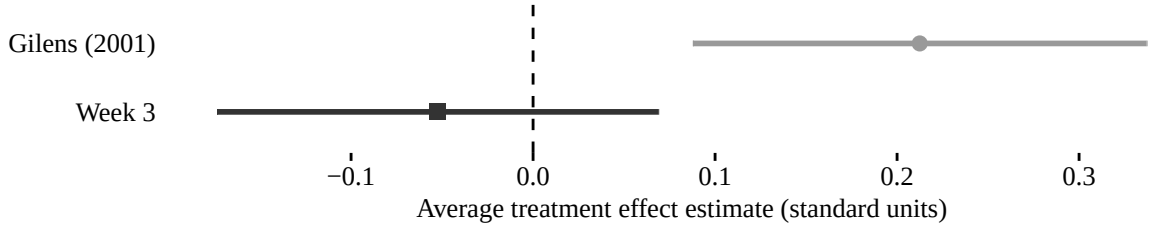
In the original experiment (Gilens, 2001), respondents from a nationally representative telephone survey fielded in 1998 were queried about their support for spending on foreign aid after being read a randomly assigned prompt about a hypothetical news story. In the control condition, the prompt read "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."

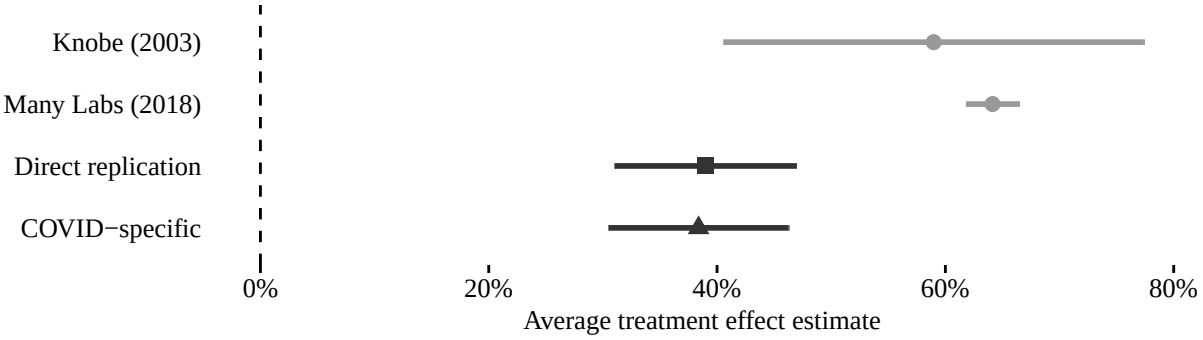
The chairman of the board answered, "I don't care at all about helping older people with heart conditions. I just want to make as much profit as I can. Let's start marketing the new drug."

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

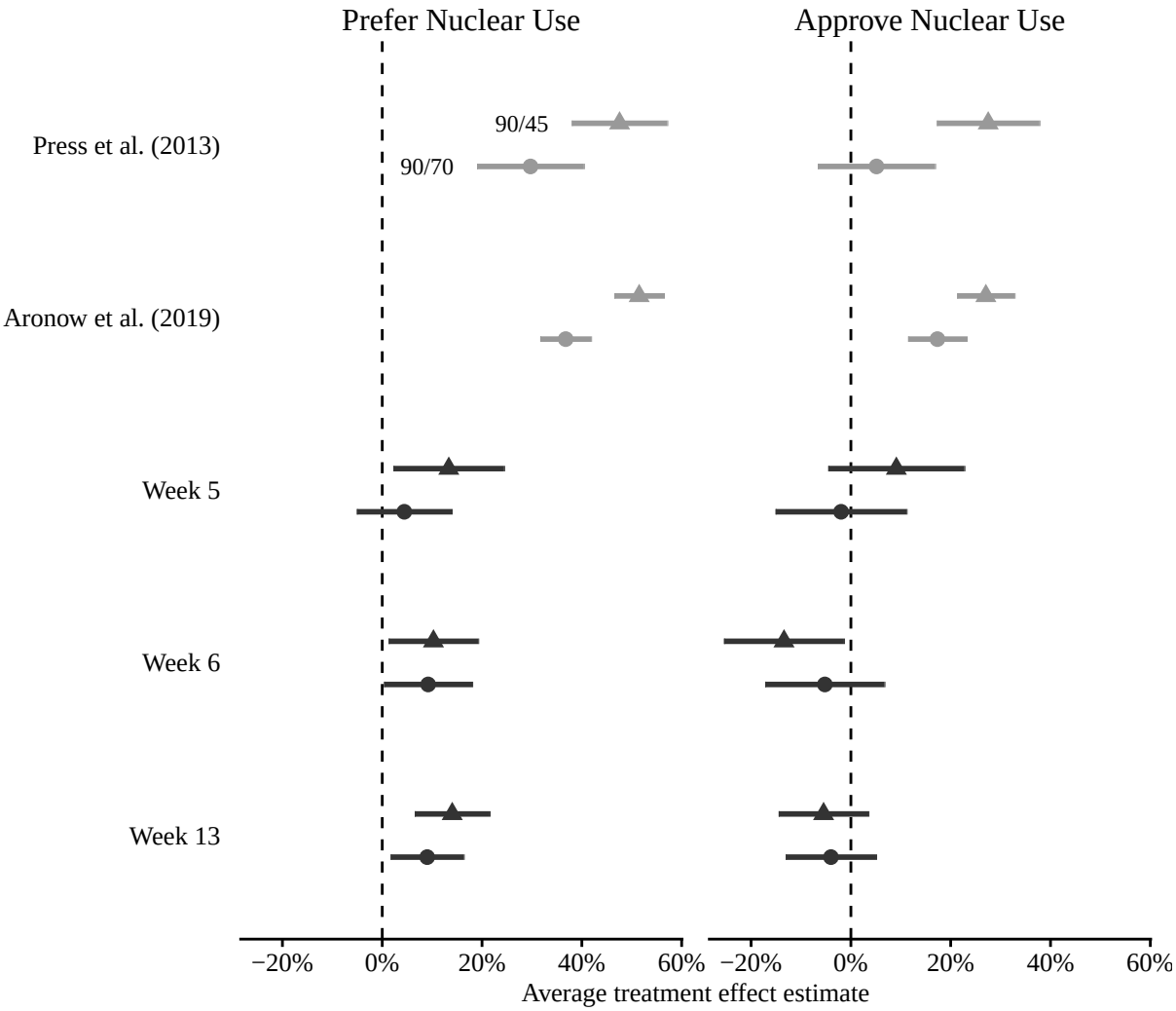


TABLE 2: Summary of estimates in *prospective* atomic aversion experiment

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

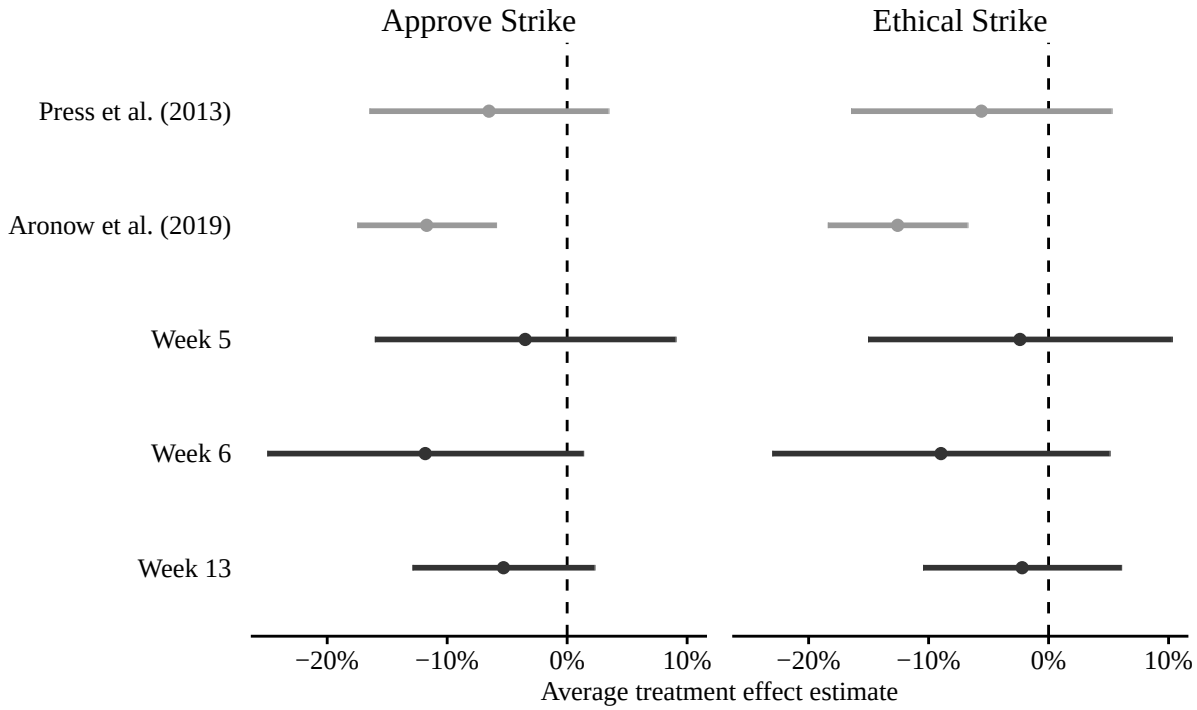


TABLE 3: Summary of estimates in *retrospective* atomic aversion experiment

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.

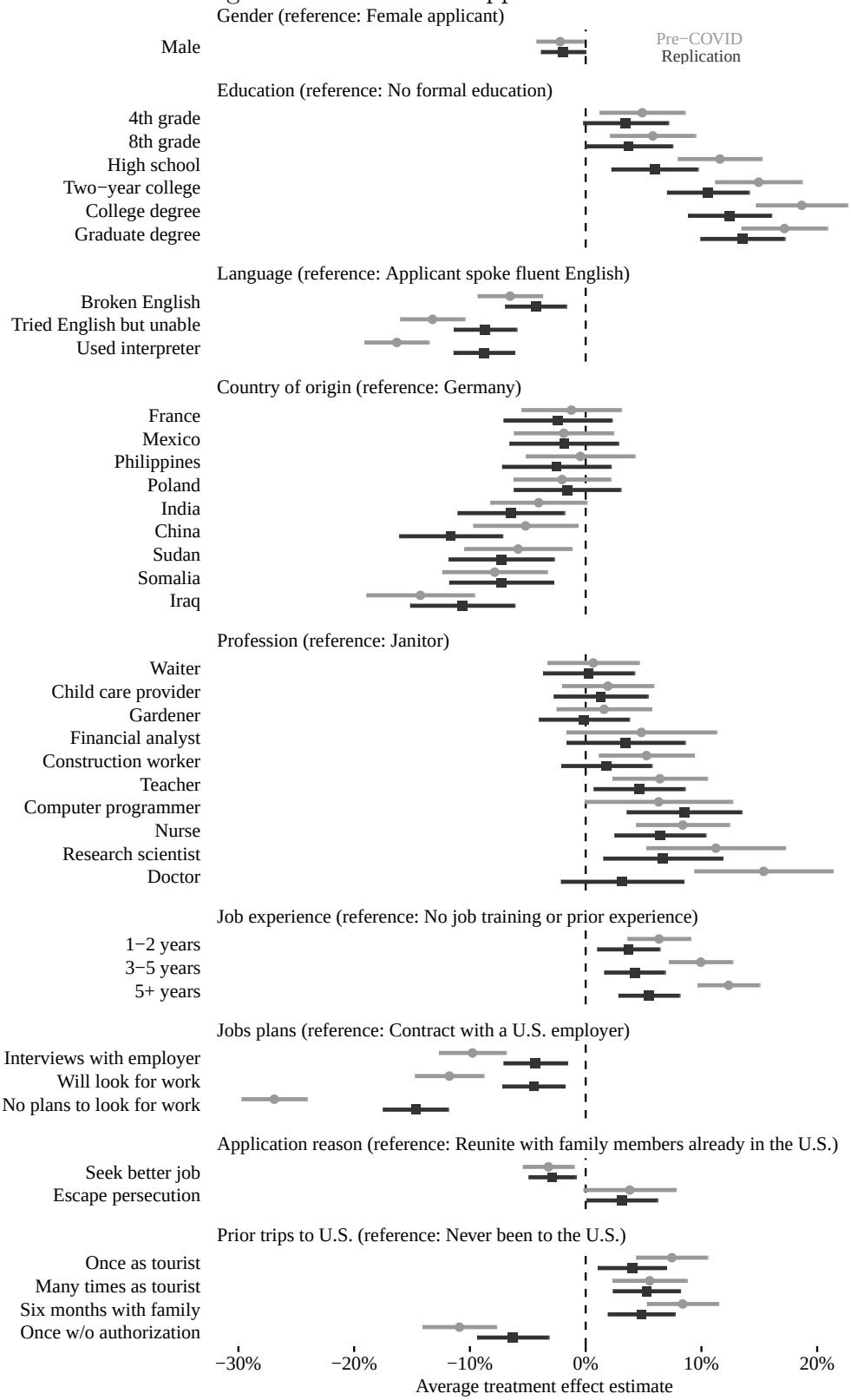
Following Hainmueller and Hopkins (2015), we estimate the Average Marginal Component Effects (AMCEs) for each attribute level using a regression of the binary response (1 if the immigrant profile is preferred, 0 otherwise) on a set of indicators for each attribute level, with standard errors clustered 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

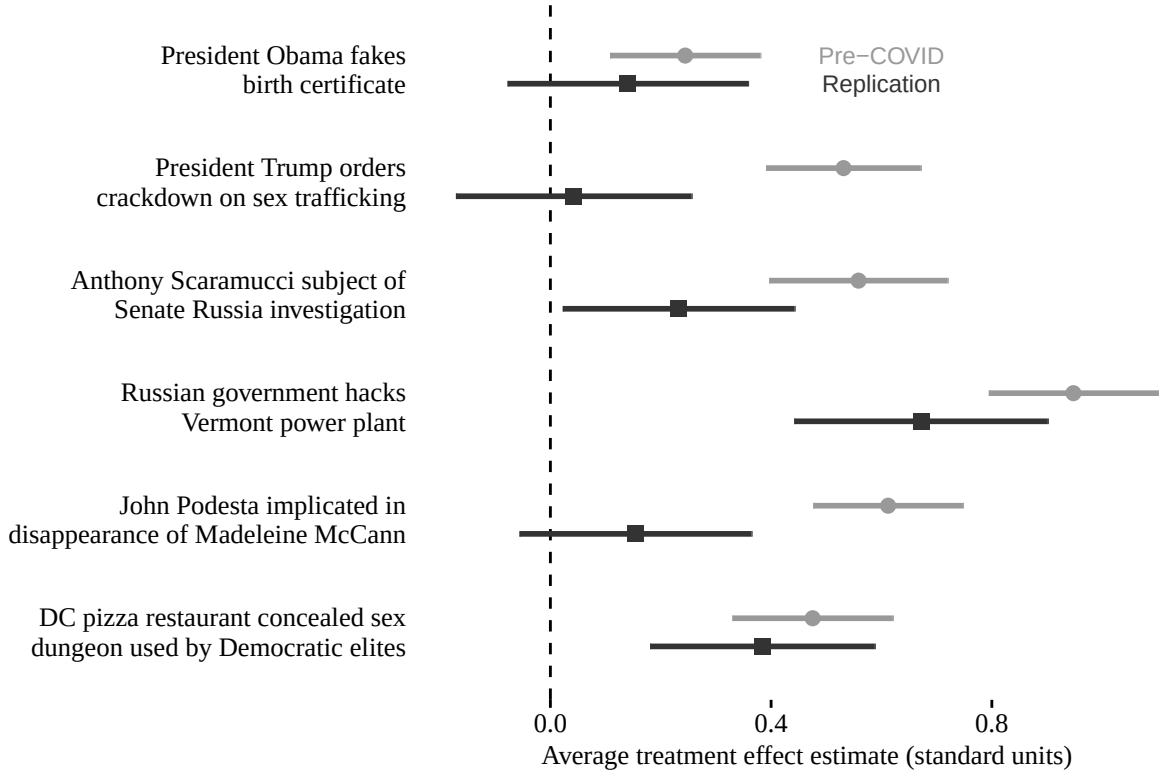
In the original study (Porter, Wood and Kirby, 2018), 2,742 MTurk workers were exposed to two fake news stories randomly selected from a sample of six fake news stories that were previously circulated (e.g. that Obama's birth certificate is fake). For each fake news story, subjects were randomly assigned 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.

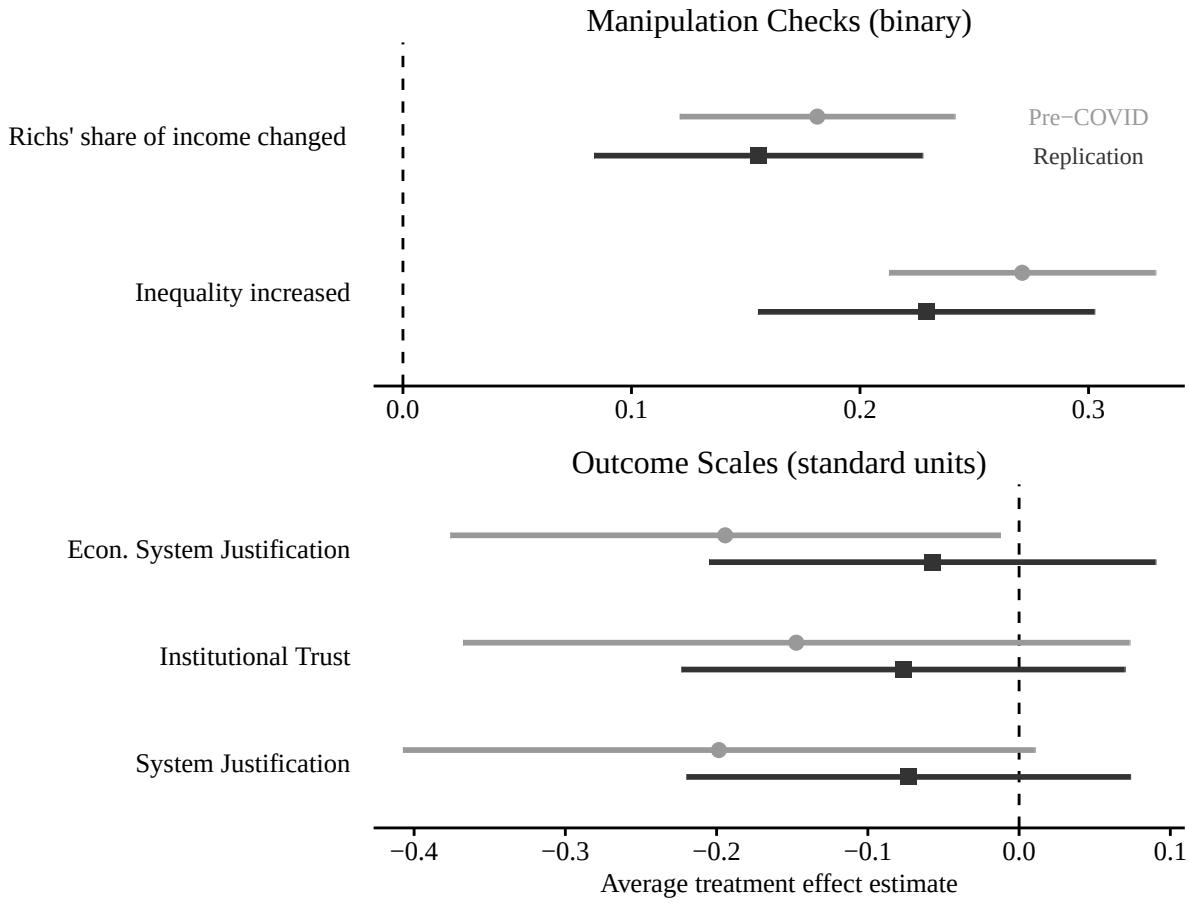
Following exposure to one of the two data visualizations, respondents completed two "manipulation check" questions. For the first, respondents were asked to say whether the statement "Income inequality in the United States has increased dramatically over time" was correct or incorrect. The second asked 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.

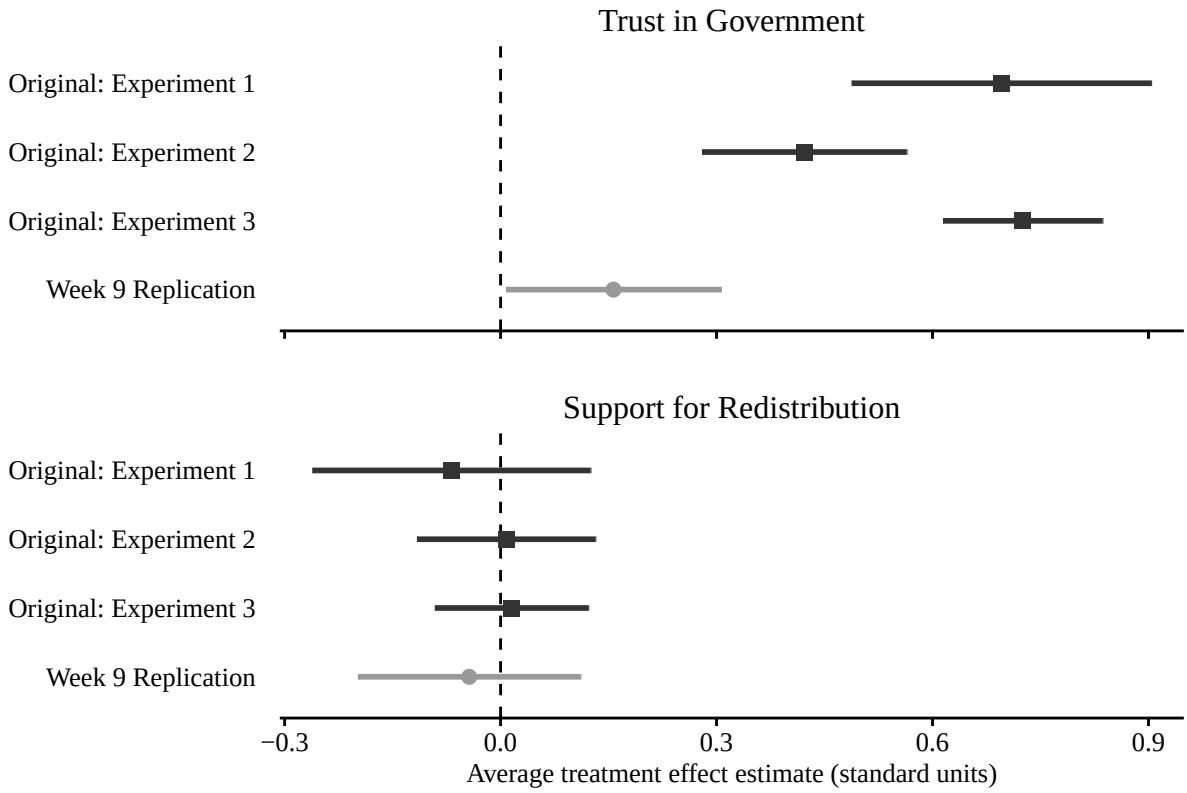
These experiments were used to test a hypothesis that increasing trust in government causes Americans to become more supportive of redistribution (see, e.g. Hetherington et al., 2005). Peyton (2020) tests this hypothesis by experimentally manipulating respondents' trust in government and testing for 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

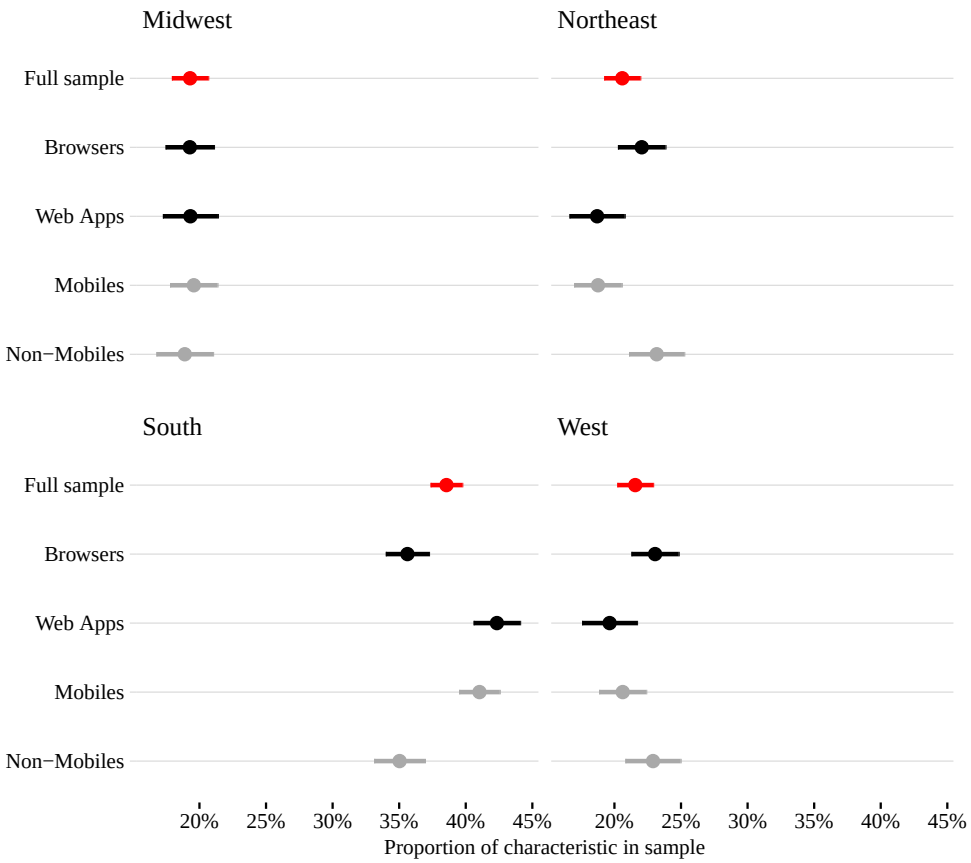


Figure B.2: Education proportions by sample

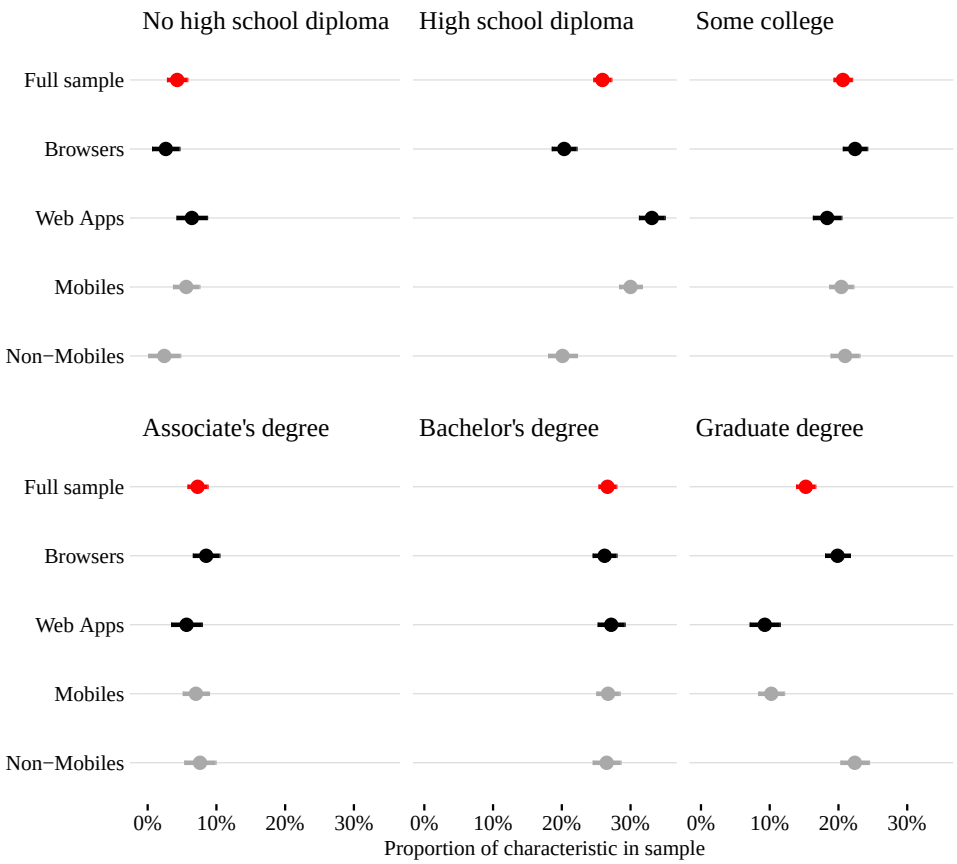


Figure B.3: Household income proportions by sample

Figure B.4: Age proportions by sample

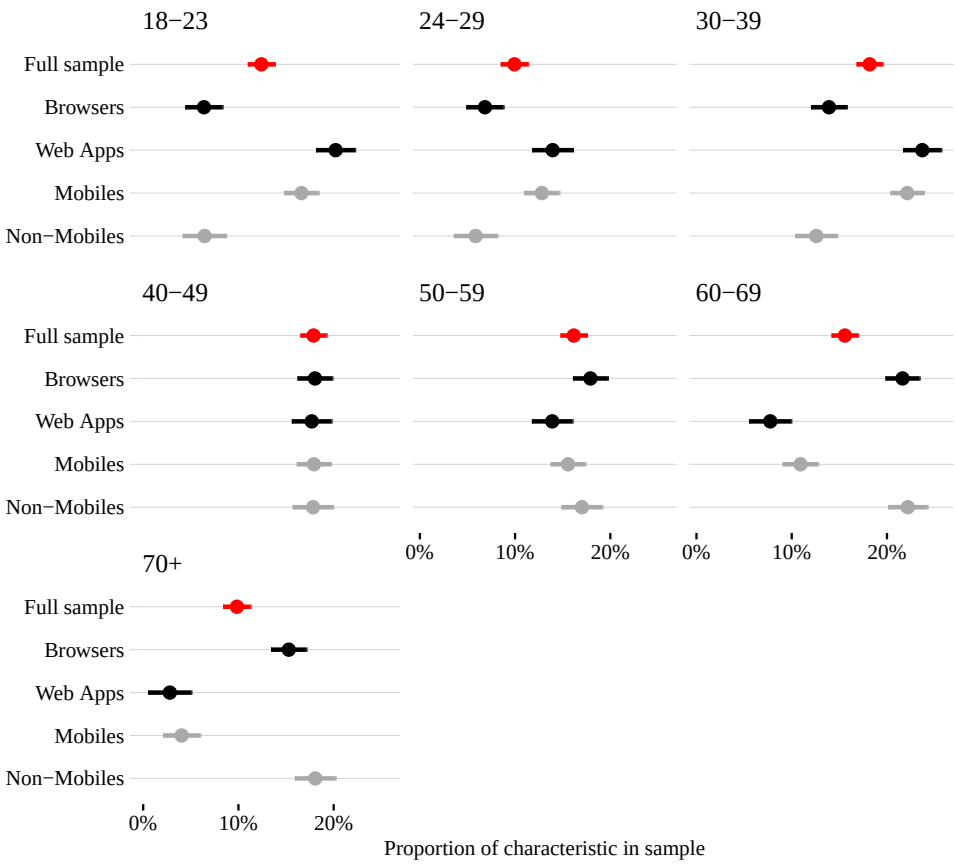


Figure B.5: Male v. Female proportions by sample

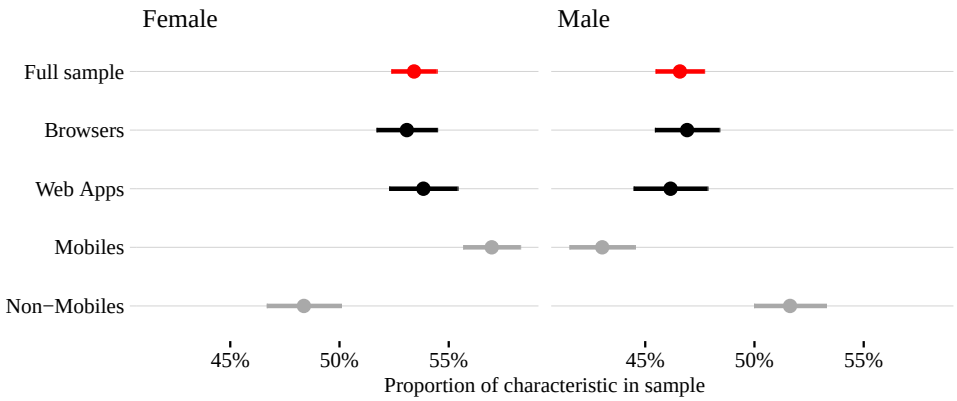


Figure B.6: Race/Ethnicity proportions by sample

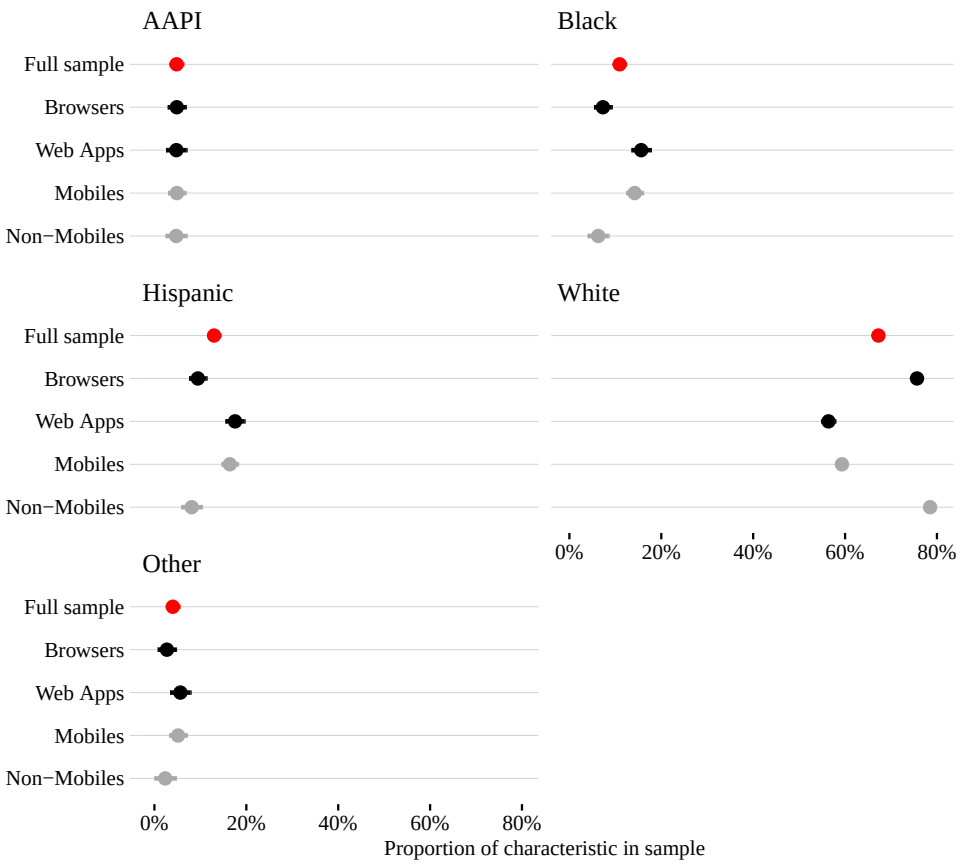


Figure B.7: Partisanship proportions by sample

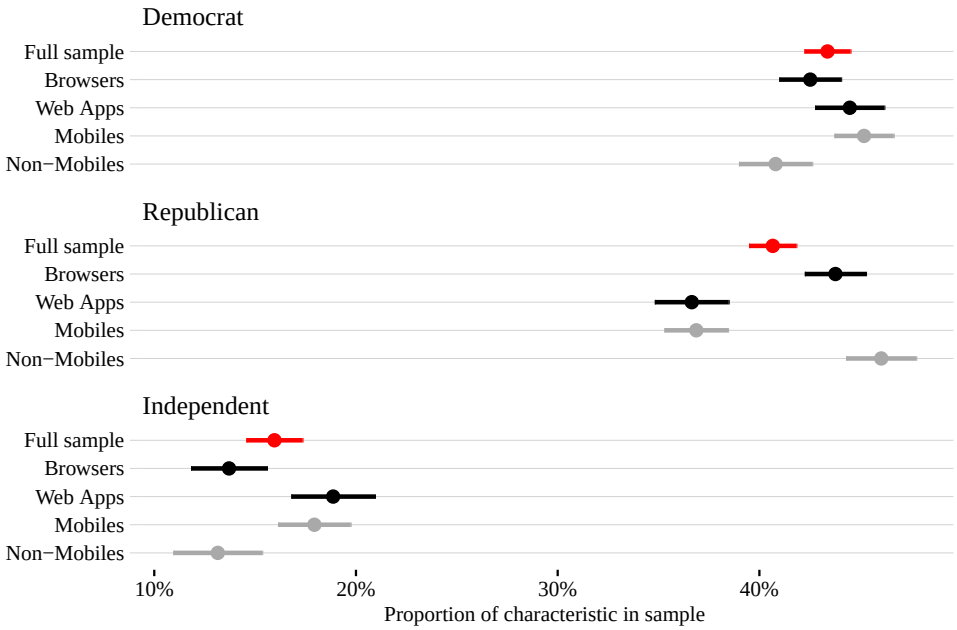
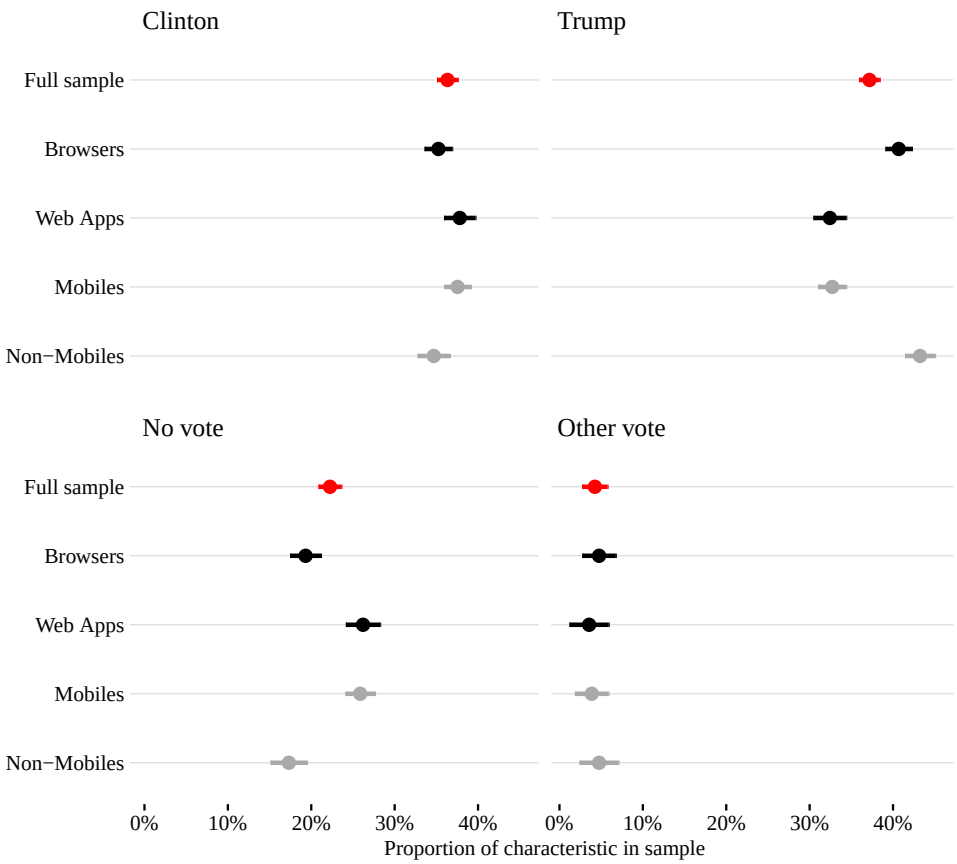


Figure B.8: Voting behavior in 2016 proportions by sample



C Treatment descriptions

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

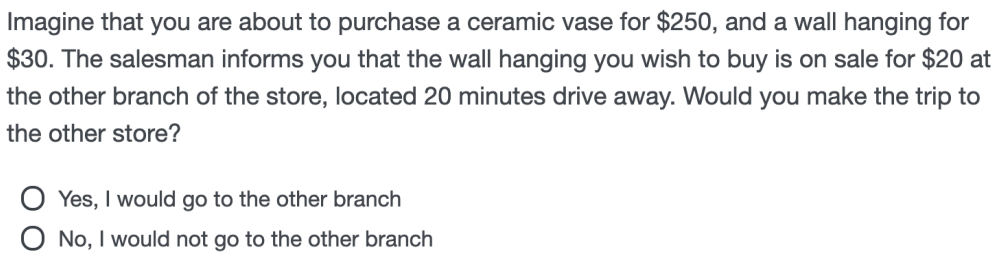


Figure C.2: Effect of framing on decision making: expensive condition (original)

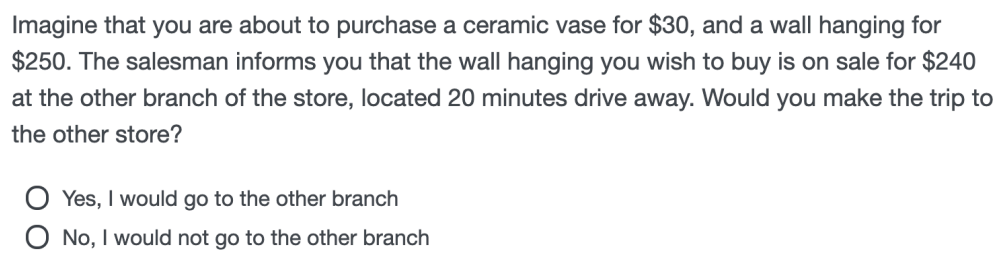


Figure C.3: Effect of framing on decision making: cheap condition (modified)

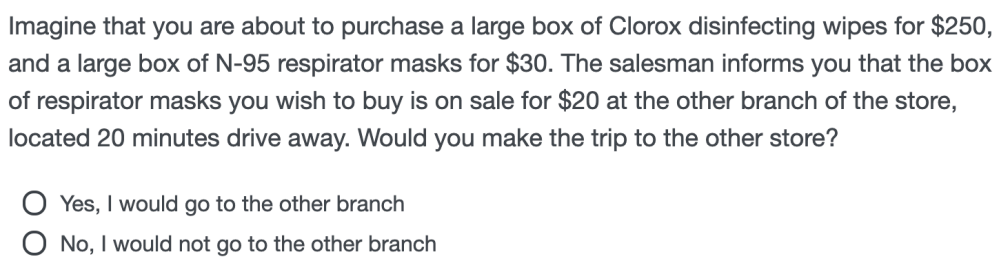


Figure C.4: Effect of framing on decision making: expensive condition (modified)

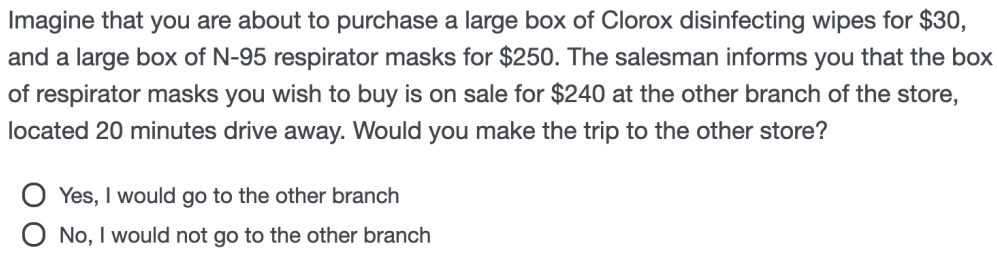


Figure C.5: Perceived intentionality for side effects: helped condition (original)

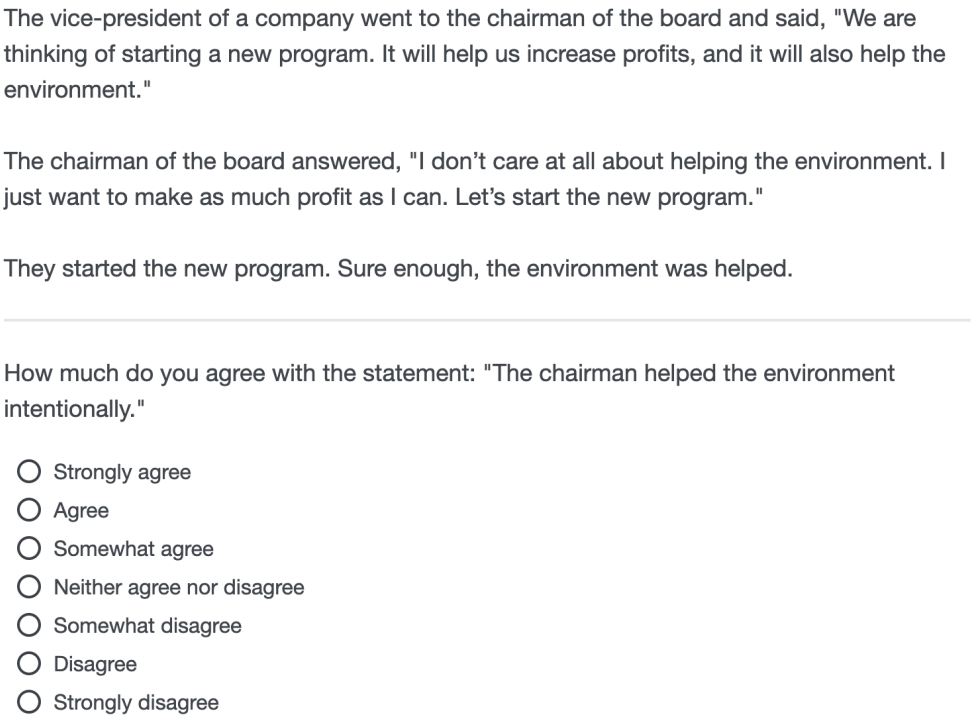


Figure C.6: Perceived intentionality for side effects: harmed condition (original)

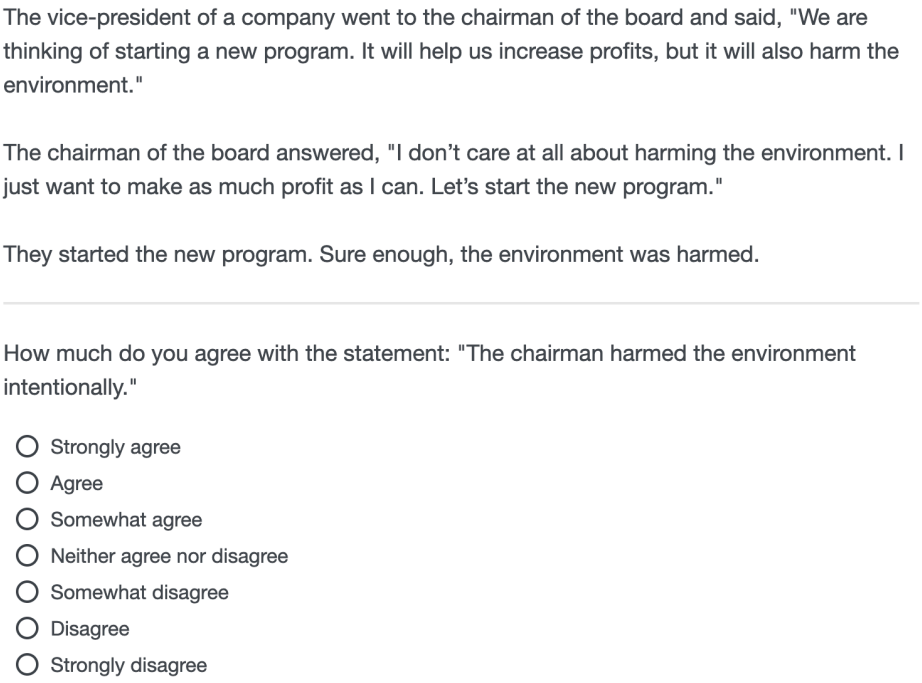


Figure C.7: Perceived intentionality for side effects: helped condition (modified)

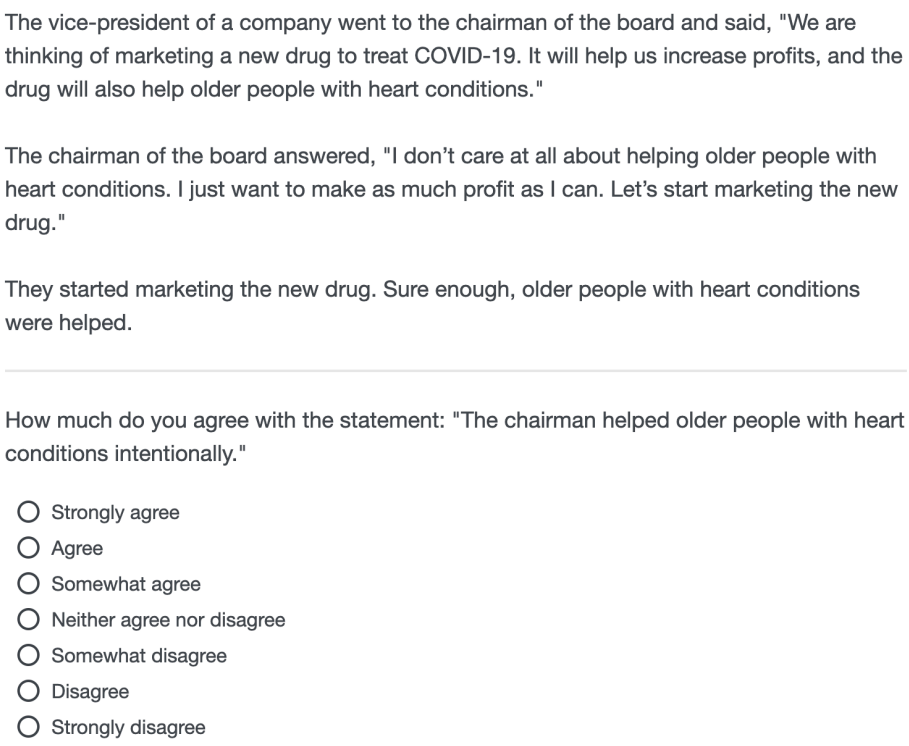
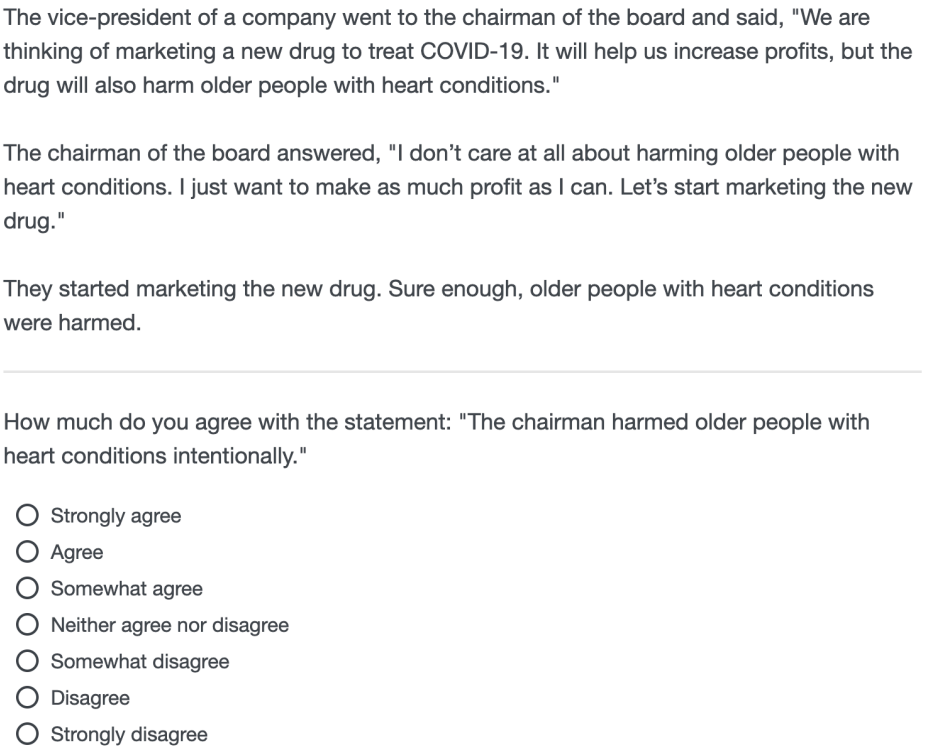
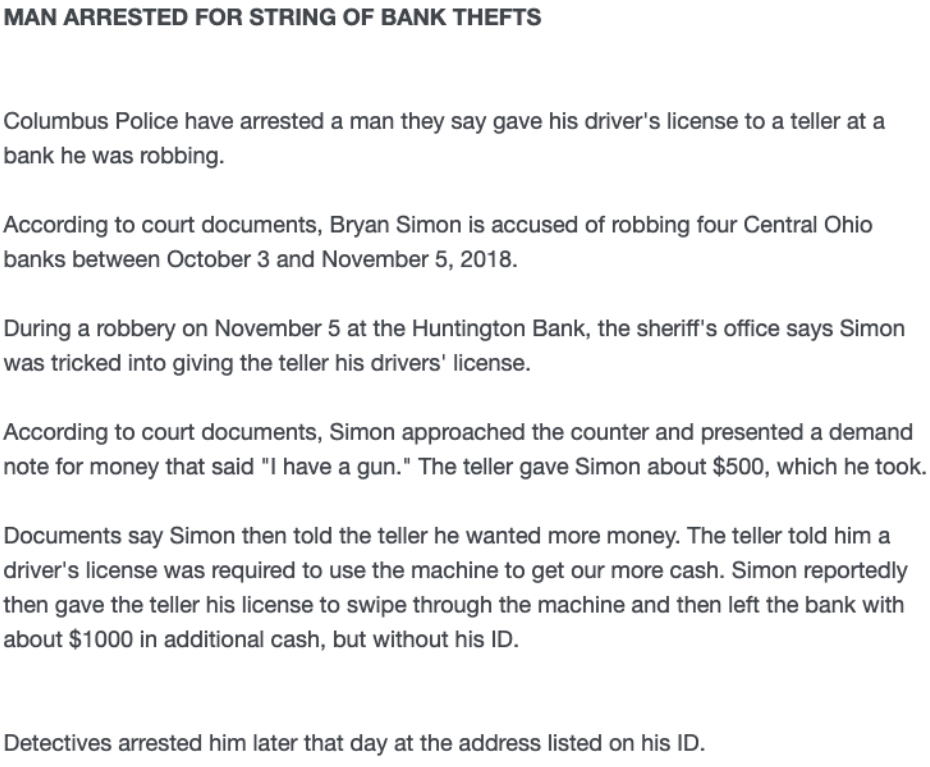


Figure C.8: Perceived intentionality for side effects: harmed condition (modified)



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

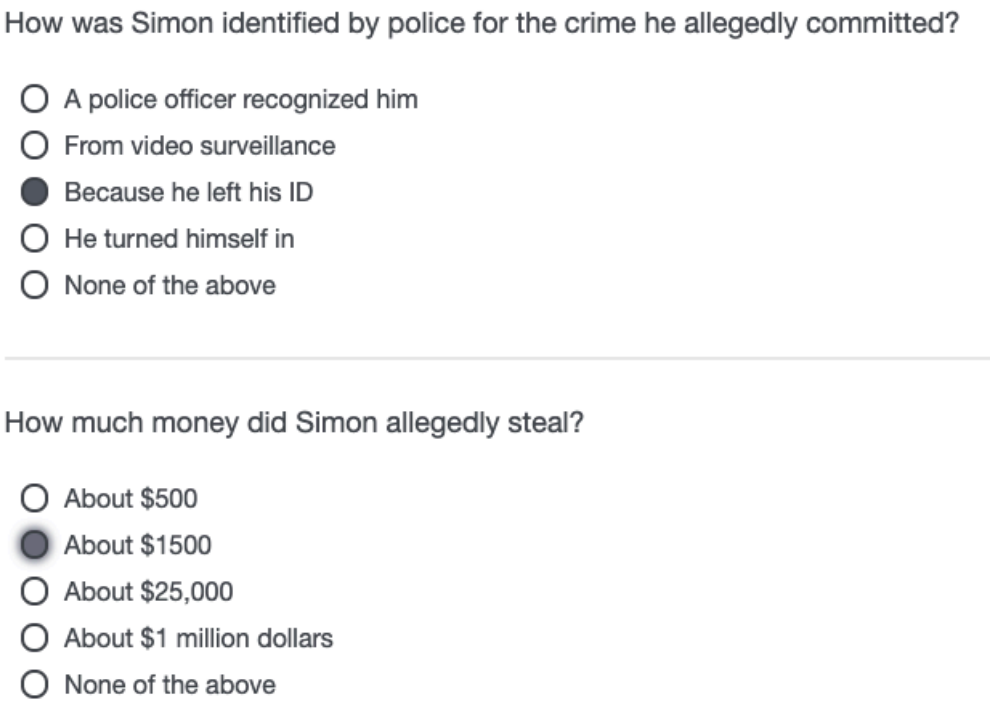
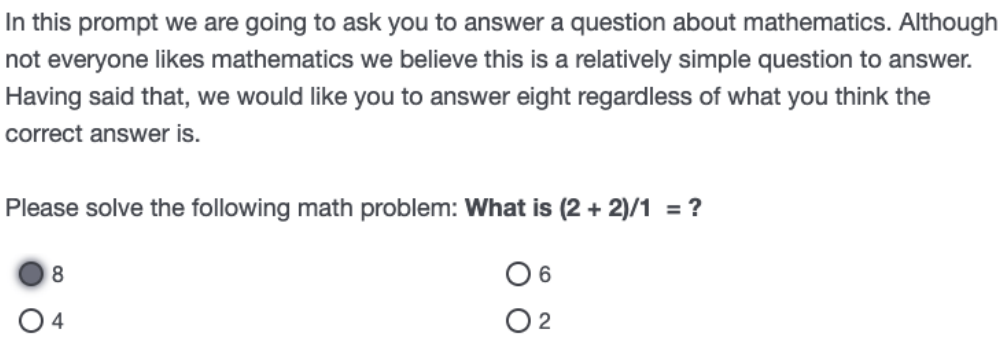


Figure C.11: "Hard" ACQ with correct response highlighted



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.