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

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

Forthcoming, *Journal of Experimental Political Science*
May 11, 2021

Abstract

The COVID-19 pandemic imposed new constraints on empirical research, and online data collection by social scientists increased. Generalizing from experiments conducted during this period of persistent crisis may be challenging due to changes in how participants respond to treatments or the composition of online samples. We investigate the generalizability of COVID-era survey experiments with 33 replications of 12 pre-pandemic designs, fielded across 13 quota samples of Americans between March and July of 2020. We find strong evidence that pre-pandemic experiments replicate in terms of sign and significance, but at somewhat reduced magnitudes. Indirect evidence suggests an increased share of inattentive subjects on online platforms during this period, which may have contributed to smaller estimated treatment effects. Overall, we conclude that the pandemic does not pose a fundamental threat to the generalizability of online experiments to other time periods.

*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, John Ternovski, and Baobao Zhang for helpful comments and feedback, to Ethan Porter for sharing replication materials, to Antonio Arechar, Matt Graham, David Rand, Patrick Tucker, Chloe Wittenberg, and Baobao Zhang for sharing pre-COVID survey data. We thank Allison St. Martin and Patrick Comer of Lucid for enlightening conversations and providing aggregate data on purchases made by academic clients. Previous versions of the manuscript were presented and benefited from feedback at Université de Montréal, Australian National University, UMass Amherst, and York 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 interaction, travel, or even simply leaving the house. Not surprisingly, the use of online surveys – a relatively low cost form of empirical research that does not require in-person data collection – expanded dramatically during the COVID-19 pandemic. For example, purchases by academic clients on Lucid – a widely used marketplace for survey respondents – nearly tripled between 2019 and 2020. Amid this boom in online research activity, concerns have been raised about the external validity of experiments conducted during this period (e.g., IJzerman et al., 2020; Rosenfeld et al., forthcoming). Here we investigate whether extrapolations from studies conducted during the pandemic to other time periods will be misleading.

Empirically demonstrating that experiments conducted during the COVID era do or do not generalize to other times is straightforward in principle. 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 generalize to the post-crisis period. Unfortunately, it may be a while before normal times fully 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 provide an answer using 33 replications fielded during the onset of the COVID-19 pandemic of 12 previously-published survey experiments. Consistent with the findings from recent replication attempts on different samples, we find strong evidence of correspondence. Our COVID-era replication estimates nearly always agree with pre-COVID estimates in terms of sign and significance, but are somewhat smaller in magnitude at an average of 73% of the pre-COVID effect size. We argue that this pattern may be explained by lower levels of attentiveness among online survey respondents during the initial onset of the pandemic. Overall, the replication estimates presented here should mitigate concerns that results from survey experiments conducted during the pandemic will not generalize to other times and contexts.

1 Background

A common framework for understanding how results generalize from one sample to a target population is the "UTOS" framework, which stands for Units, Treatments, Outcomes, and Settings (Cronbach and Shapiro, 1982). The framework predicts that results will exhibit greater correspondence (or generalizability or transportability) between a sample and the

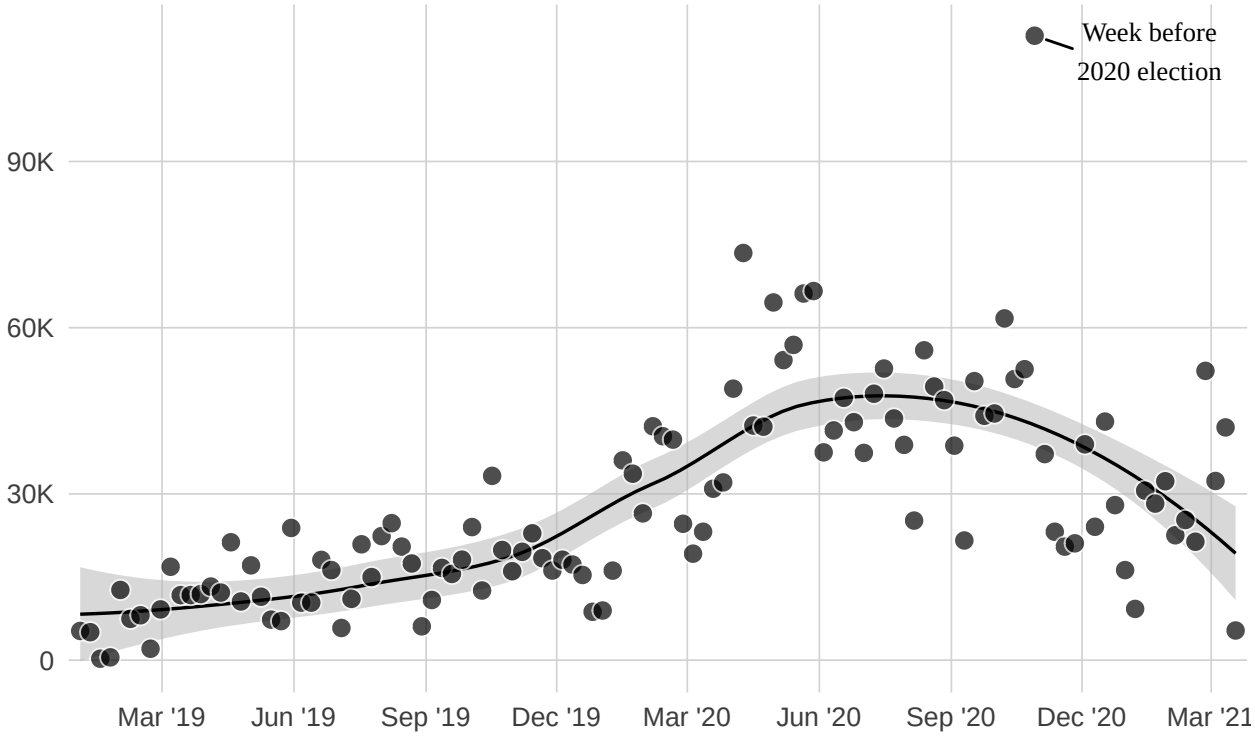
target population with increasing similarity of their UTOS components. A decade's worth of studies in political science have invoked the UTOS framework to argue that survey experiments conducted with online samples generalize to the U.S. national population (Berinsky, Huber and Lenz, 2012; Mullinix et al., 2015; Coppock, 2019; Coppock, Leeper and Mullinix, 2018; Coppock and McClellan, 2019). Recent work by Findley, Kikuta and Denly (2020) elaborates the UTOS framework to differentiate the time dimension from other features of settings, pointing out that the mechanisms that link treatments to outcomes may change over time. Munger (2020) describes growing concerns about generalizability across time as a question of "temporal validity."

The COVID-19 pandemic led to changes in at least two important factors relevant to online survey experiments. First, nearly everyone experienced massive disruptions to their daily lives. These changes may have affected how online survey participants respond to treatment. For example, Americans may have been in a heightened state of anxiety due to a variety of factors, including an unmitigated public health crisis, rising economic insecurity, and chronic political instability. Anxiety and other emotional state variables have been shown to affect information processing (e.g., Gadarian and Albertson, 2014), willingness to dissent (e.g., Young, 2019), and to condition effects of other treatments (e.g., Valentino et al., 2009).

Another important factor is that the pandemic altered, or was coincident with, changes in the composition of markets for online survey respondents. Arechar and Rand (2020), for example, find that an influx of new workers to the Amazon Mechanical Turk (MTurk) platform in March 2020 led to samples that were less attentive, but more demographically diverse and representative of the U.S. population. On the Lucid platform, the demand for survey respondents in 2020 increased by roughly 40% among commercial clients, and by 200% among academic clients (see Figure 1). In the immediate wake of the pandemic, Lucid suppliers – who furnish survey respondents for both academic and commercial clients – struggled to recruit enough new participants to meet this increased demand.[1] Consistent with the possibility that suppliers attempted to meet this increased demand with relatively lower-quality respondents, Aronow et al. (2020) find that respondent attentiveness on Lucid was declining over the period March-July 2020. This decline in attentiveness may also have reflected changes in the types of people who were previously attentive becoming less attentive because of COVID-19 related disruptions.

[1] Conversation with Lucid representatives on 18 March 2021. According to Lucid, supply gradually increased over the summer of 2020, just as demand for survey responses intensified in the lead up to the 2020 election.

Figure 1: Total weekly survey responses completed and sold to academic buyers by Lucid between January 2019 and March 2021



Notes: Points denote weekly survey responses completed and sold to academic customers by Lucid (in thousands). Dark lines and 95% confidence bands from loess smoother. The WHO declared COVID-19 a pandemic on 11 March 2020. Proprietary data provided by Lucid cover the period 1 January 2019 - 18 March 2021. In 2019 Lucid sold 729,284 completed survey responses to academic buyers, compared with 2,185,387 in 2020.

1.1 Design

Given the proliferation of online research conducted during the COVID-19 pandemic, we can articulate concerns about external validity as an empirical question: would the results from an experiment conducted during the pandemic have been different if the experiment were instead conducted in a different period?

Between March and July 2020, we recruited weekly samples of approximately 1,000 U.S. based participants via Lucid. Each survey was between 10 and 15 minutes long (median duration: 12.8 minutes) and was structured in discrete 3-5 minute modules. Lucid collects demographic information from all respondents before they enter any particular survey, enabling the construction of quota samples that approximate U.S. census margins (Coppock

and McClellan, 2019). 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 differ from target populations in both observable and unobservable ways.[2] 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 are an effective and widely-used tool.

1.1.1 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. *Theoretical and political 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 a mix of classic and contemporary studies, and for coverage across political science sub-fields. 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. We conducted at least one direct replication of each study. Modified versions of studies 2, 3, 5, and 7 were also replicated using COVID-specific content. A description of each modified version and their corresponding replication estimates are provided in Online Appendix Section A.

[2]Balance on the demographic marginal distributions does not imply balance on the joint distributions. Moreover, because these are not probability samples, the degree of imbalance on the distribution of unobserved attributes is unknown.

We categorize each replication as a "success" if the COVID-era estimate(s) are correctly signed and statistically distinguishable from zero. For studies with "null" results (studies 11-12), replication was declared successful if estimate(s) were indistinguishable from zero and their pre-COVID benchmarks. A replication "failure" occurs when estimate(s) are incorrectly signed, regardless of whether they are distinguishable from zero. For studies with multiple treatment arms/outcomes, we concluded replication was successful if the preponderance of evidence supports success. In one case, the preponderance of evidence was ambiguous and we concluded the replication attempt was a partial success. Additional details are available in Online Appendix Section A.

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

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

2 Results

We were able to obtain the original effect size(s) for each of the 12 pre-COVID studies listed in Table 1, and at least one pre-COVID replication estimate of the original effect size(s) for 7 of these studies. In total, we obtained 89 pre-COVID estimates and 101 replication estimates.[3] A detailed description of each study, their pre-COVID estimates, and the individual replication estimates that we obtained are provided in Section A of the Online Appendix. Here we provide an overall summary for each study by comparing the pooled estimates from our COVID-era replications with the pooled estimates from the pre-COVID benchmarks (original estimates and any pre-COVID replications).

For each study, we calculate summary effect sizes for each treatment-outcome pair using a precision-weighted average, with weights based on the reciprocal of the variance. We can then compare the pre-COVID and replication estimates by taking the differences between their summary effect size estimates. 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 studies and 56 for the remaining studies. For the non-conjoint studies, Figure 2 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 3 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 (6 statistically significant differences) and 6 of 41 were larger in replication (1 of 6 significant differences).

Pooling across all 65 of 69 correctly signed pairs presented in Figures 2-3, the replication

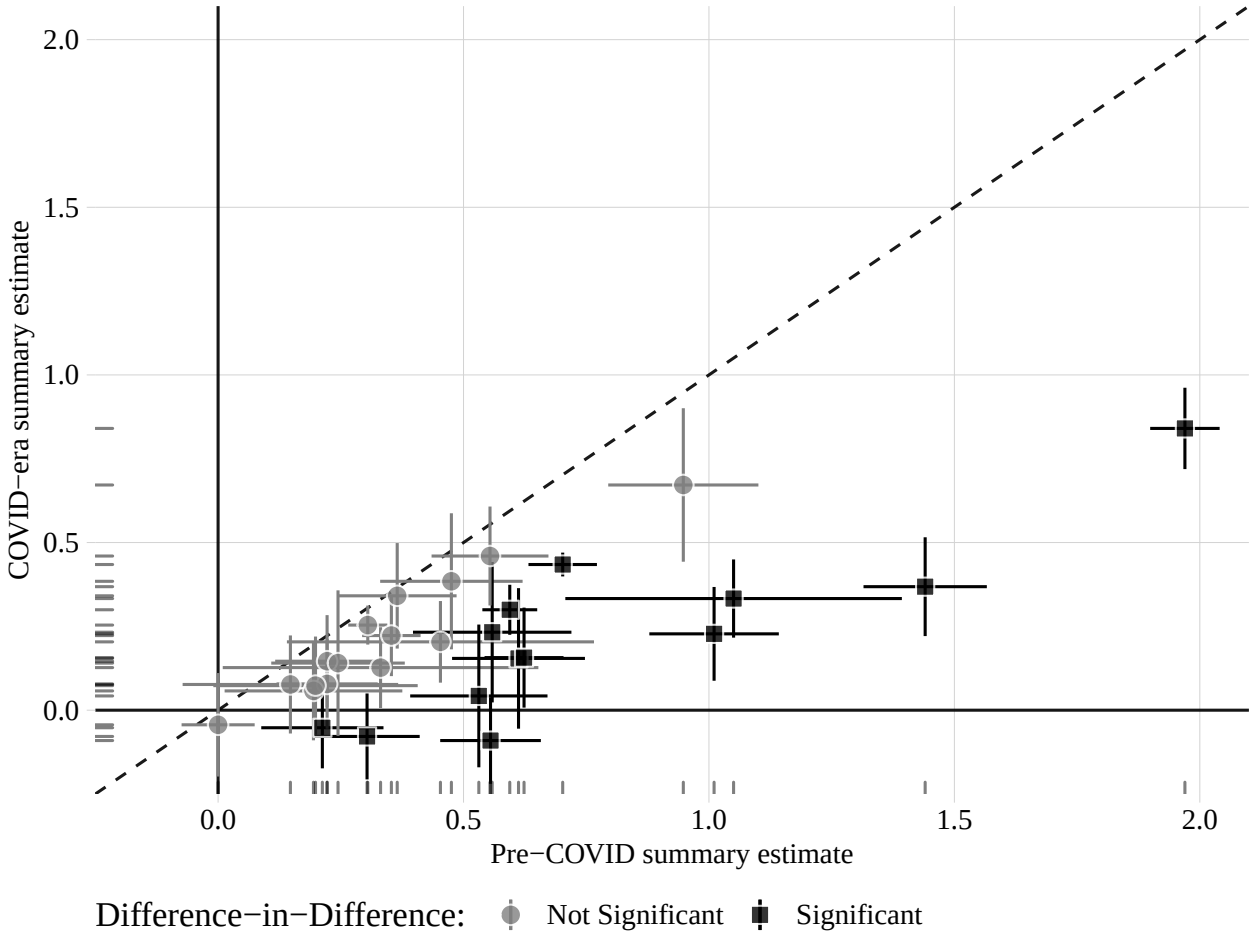
[3]The General Social Survey (GSS) has administered the welfare versus aid to the poor experiment 20 times between 1986 and 2018. If all GSS estimates are included, we have 108 pre-COVID estimates in total. Given the variation in question wordings and survey modes across time, we use the 1986 GSS experiment as our "original" estimate, and take the first two replications conducted on online samples by Huber and Paris (2013) as our pre-COVID replication estimates.

estimates were, on average, 73% as large as the pre-COVID estimates. The 41 correctly signed estimates from the conjoint replication were, on average, 87% as large as the original. The 24 of 28 correctly signed estimates from the other replications were, on average, 49% as large as the pre-COVID summary effect sizes.

When compared with other replication efforts, the correspondence for the conjoint experiment is high whereas the correspondence for the others is modest. For example, in one of the earliest large-scale replications, Open Science Collaboration (2015) replicated 100 experiments from three top psychology journals and found that about 40% replicated and effect sizes were, on average, about half the magnitude of the original effects. In economics, Camerer et al. (2016) replicated 18 survey and lab-based experiments and found effect sizes in the 11 studies that successfully replicated were approximately 66% of the original. Coppock (2019) replicated 40 treatment effect estimates from 12 survey experiments, finding that the median MTurk estimate was 66% of the original effect size.

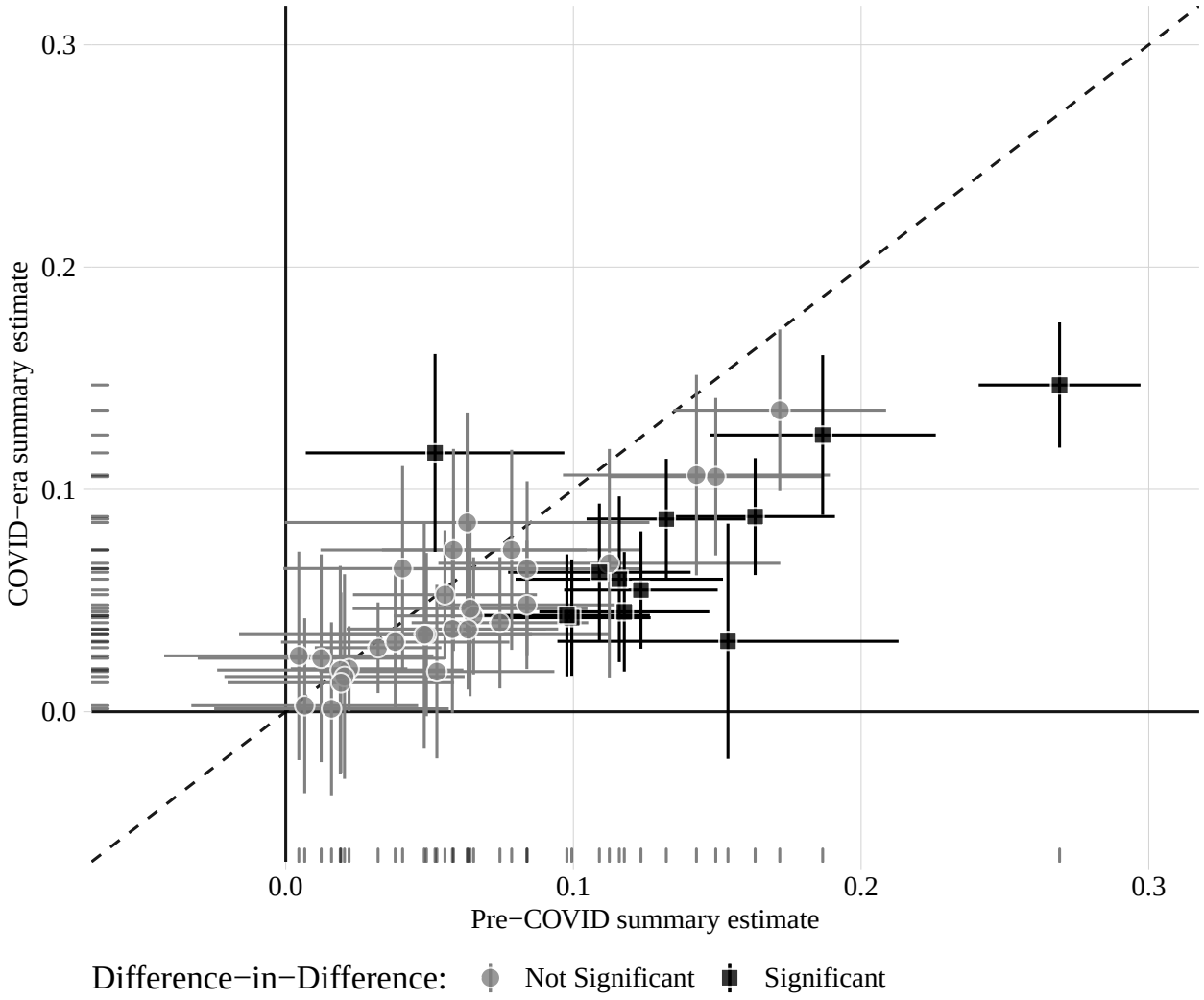
Common explanations for somewhat diminished effect sizes in replication studies include imperfect adherence to the original experimental protocol on the part of replicators and publication pressures that systematically select for larger effect sizes in original study journal submissions. These mechanisms may be operative here as well. At the same time, we do not think they are the main reasons for the smaller effect sizes we find because the experimental protocols are relatively straightforward and the bulk of these studies had been successfully replicated by others before us with larger effect sizes than we find. Prior replications of studies 7 and 8, for example, have found effect sizes at least as large as the original (see Klein et al., 2018; Aronow, Baron and Pinson, 2019).

Figure 2: Comparison of 28 summary effect sizes across 11 studies (conjoint excluded)



Notes: Estimated summary effect sizes and 95% confidence intervals from pre-COVID experiments (horizontal axis) and COVID-era replications (vertical axis). 13 of 28 replication estimates were significantly different from their pre-COVID benchmark at $p < 0.05$.

Figure 3: Comparison of 41 summary effect sizes in conjoint experiments



Notes: Estimated summary effect sizes and 95% confidence intervals from pre-COVID experiments (horizontal axis) and COVID-era replications (vertical axis). 7 of 41 replication estimates were significantly different from their pre-COVID benchmark at $p < 0.05$.

3 Can inattention explain attenuated replication estimations?

In this section, we consider declining attentiveness among survey respondents during the COVID-era as a potential explanation for attenuated treatment effect estimates. This decline in attention – as measured by increased failure rates on Attention Check Questions (ACQs) in Aronow et al. (2020) and Arechar and Rand (2020) – might reflect a genuine decline in attention among online survey respondents, underlying changes in the types of people who participate in online research, or both. Although we cannot distinguish among these possibilities, prior work has demonstrated that inattention leads to measurement error in self-administered surveys. Berinsky, Margolis and Sances (2014), for example, replicated the gain versus loss framing experiment from Tversky and Kahneman (1981) and found an estimated treatment effect of approximately zero among respondents that failed a pre-treatment attention screener.

Measurement error induced by inattention is nonrandom, so the correlations across covariates and survey outcomes can be overstated or understated in descriptive survey work. In the experimental setting, however, inattention generates a form of treatment noncompliance. To fix ideas, suppose that an experimental sample can be partitioned into two types of individuals: the "attentive" and the "inattentive". Imagine the average effect among the attentive is positive and the average effect among the inattentive is zero, because these subjects do not engage with treatment and are therefore unaffected by it. The Average Treatment Effect (ATE) in the full sample is therefore a weighted average of two quantities: the average effect among the attentive, and the average effect among the inattentive. The ATE estimated for the full sample will therefore be closer to zero than the average effect among the attentive. Estimates for the overall ATE attenuate towards zero as the share of inattentive subjects in the sample grows.

For ease of exposition, we have dichotomized subjects as attentive or inattentive, but respondent attention in online surveys is both continuous and dynamic (Berinsky, Margolis and Sances, 2016; Berinsky et al., 2019). It is continuous because subjects may be more or less attentive and it is dynamic because attention may change during a survey. Accounting for these subtleties is important, but the core argument is that estimated treatment effects will attenuate towards zero as inattention increases. If subjects can be classified as one type or the other, then we expect smaller effects where the share of the inattentive is higher. If subjects pay partial attention, we expect that estimated treatment effects among the

partially attentive will be smaller what they would have been had these subjects paid closer attention. Given that the proportion of inattentive subjects in online samples appears to have increased during the period in which our replications were conducted (see Aronow et al., 2020; Arechar and Rand, 2020), this may partly explain why our replication estimates are smaller when compared to pre-COVID estimates.

We can examine this possibility by estimating effects among the attentive and inattentive separately. The most straightforward approach is to include pre-treatment ACQs (see Oppenheimer, Meyvis and Davidenko, 2009; Paolacci et al., 2010).[4] However, only two of our surveys included ACQs, so we consider alternative approaches as well.

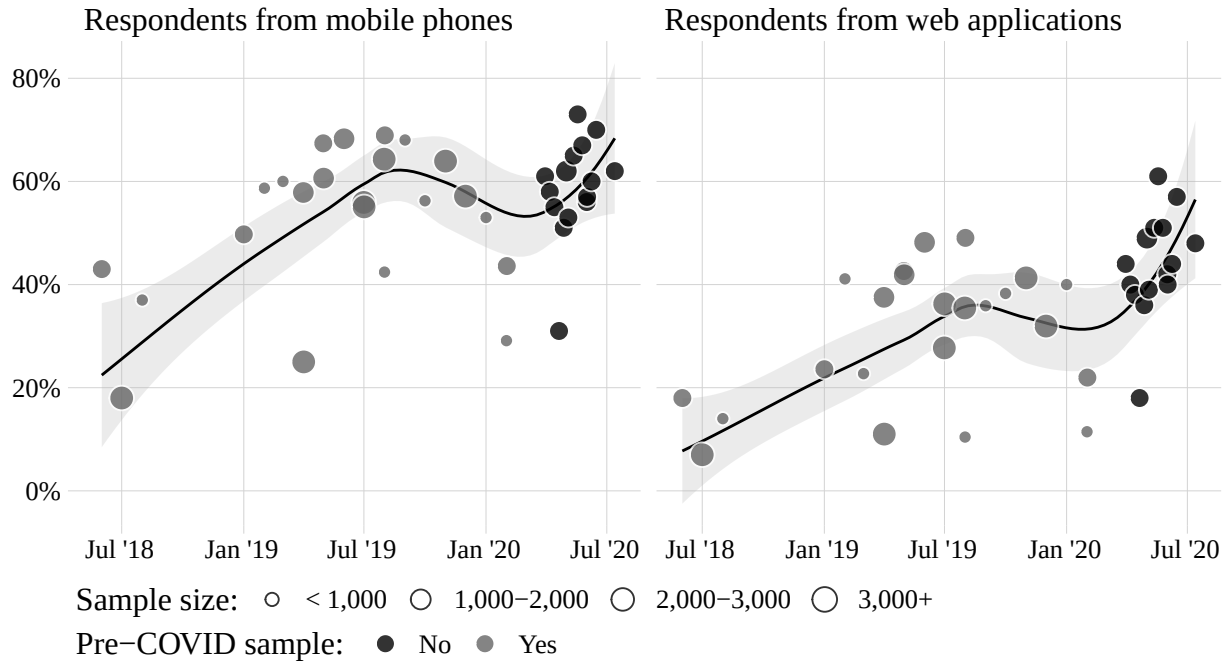
A second approach is to rely on individual-level meta-data. One possibility is that increased inattention during the COVID-era stems from an increasing proportion of respondents arriving to the survey environment from mobile games.[5] The "User Agent" string captured from the end user's browser by online 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 information: 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 participants coming from web-applications (rather than internet browsers) ranged from 0.19 to 0.61, and the proportion of participants coming from mobiles (rather than desktops or tablets) ranged from 0.31 to 0.73. Based on an additional sample of 63,245 respondents obtained from Lucid surveys fielded prior to March 2020, we observe that the proportion of participants from both web-applications and mobiles was trending upward prior to the COVID-19 outbreak. In the 2020 surveys, approximately 41% of respondents came from web applications (56% from mobiles), an increase from 33% (56% from mobiles) in 2019 and just 13% (33% from mobiles) in 2018. These results, reported in Figure 4, are consistent with the declining data quality documented by Aronow et al. (2020).

[4]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.

[5]In a separate Lucid survey 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, we asked respondents to self-report whether they were recruited to the survey from a 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. Interviews with Lucid representatives in 2021 suggest suppliers that have since been removed from the platform were providing these respondents from mobile games.

Figure 4: Respondents from mobile devices and web applications from Jun 2018 to Jul 2020



Notes: Points denote percentage of respondents in each Lucid survey, sized in proportion to sample size. Dark lines and 95% confidence bands from loess smoother. Estimates are based on individual-level browser meta-data and 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.

Pooling across the 13 surveys used for our replication studies, 97% of participants from web-applications were also on mobile devices, and 72% on mobile devices arrived from web-applications. Those from web-applications spent approximately 7 minutes less time completing surveys than those from web browsers, who spent an average of 21.5 minutes. Respondents 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.

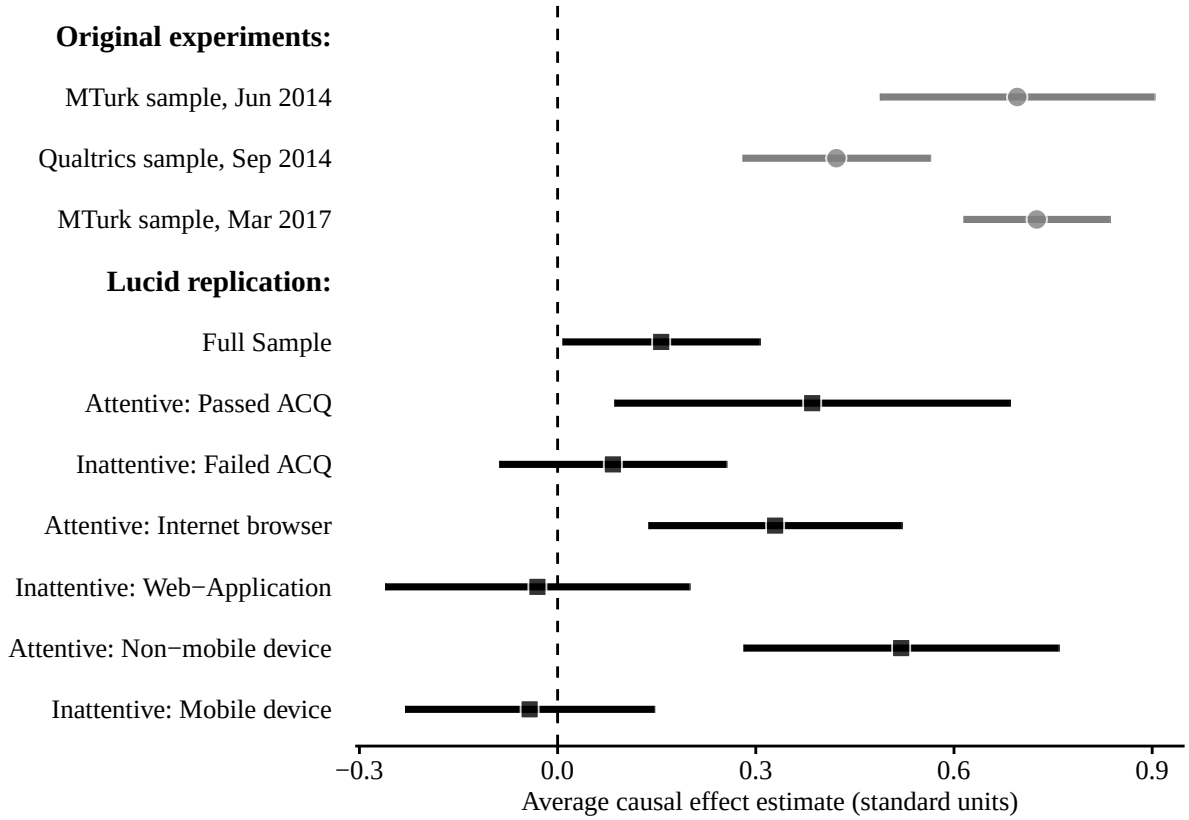
Table 2: ACQ pass rates by attentiveness and level of difficulty

Difficulty	Browser	Web-App	Difference	Non-Mobile	Mobile	Difference
Easy	0.66 (0.02)	0.55 (0.02)	0.11 (0.03)*	0.63 (0.02)	0.58 (0.02)	0.05 (0.03)
Medium	0.53 (0.02)	0.37 (0.02)	0.16 (0.03)*	0.50 (0.02)	0.42 (0.02)	0.08 (0.03)*
Hard	0.22 (0.02)	0.15 (0.01)	0.07 (0.02)*	0.24 (0.02)	0.16 (0.01)	0.08 (0.02)*

Notes: The "Easy" and "Medium" questions were novel ACQs that subjects completed after reading a short news article (see Fig. C.9-C.10). 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 Fig. C.11). Standard errors in parentheses. $p < 0.05^*$.

While most of our surveys did not include pre-treatment ACQs, we did include one in the direct replication of Peyton (2020), allowing for a direct comparison of estimates among the attentive using both the ACQ and metadata approaches. Figure 5 shows that regardless of the approach used to classify subjects as attentive, estimated treatment effects among the attentive are positive and significant. Among those who passed the ACQ, the average effect of treatment was 0.39 (SE: 0.15). Among those coming from browsers, the estimate was 0.33 (SE: 0.10), and among non-mobile users it was 0.52 (SE: 0.12). Among those who failed the ACQ (estimate: 0.08, SE: 0.09), came from a web-application (estimate: -0.03, SE: 0.12), or used a mobile phone (estimate: -0.04, SE: 0.10), estimated treatment effects were all close to zero and nonsignificant.

Figure 5: Reanalysis of estimated treatment effects on trust in government for Peyton (2020) replication



Notes: Estimates for the Lucid replication, fielded in May 2020, are presented for the full sample, and each sub-group partition created by three different methods for classifying attentive v. inattentive respondents. Pairwise correlations between these methods: 0.67 for Non-mobile and Browsers, 0.10 for Non-mobiles and ACQ pass, and 0.09 for Browsers and ACQ pass. Estimates for the original experiments come from the replication archive for Peyton (2020) at https://doi.org/10.7910/DVN/L3NT6P. The Qualtrics sample was a quota sample that, like our Lucid replication sample, approximates U.S. census margins on respondents demographics. In the MTurk sample from March 2017, Peyton (2020) Appendix S5.8 reports 87% of respondents passed their pre-treatment ACQ. In our direct replication, 19% of respondents passed this same ACQ.

A final approach is to reason about the plausible scope for inattention to explain the attenuated treatment effects we find. If we assume that treatments do not affect outcomes among inattentive respondents, we can recover the effect among attentive respondents by dividing the overall estimates by the proportion of attentive respondents. If this assumption is incorrect, either because treatment effects among the inattentive are non-zero or because certain treatments affect attention, then this adjustment could be biased upward or downward.

Pooling across all replication studies, we find effect sizes(s) were, on average, 73% their pre-COVID magnitude. If 73% of respondents were attentive and 26% inattentive, our replication estimate would, on average, match our pre-COVID benchmarks. Empirical estimates of inattention suggest this rate is not unreasonable. Aronow et al. (2020) report that approximately 70% of respondents passed ACQs in surveys fielded between March and July 2020. Dividing each of the COVID-era summary estimates that we obtained by 0.70 yields an estimated replication correspondence of 104% the pre-COVID effect sizes.

Of course, it is unlikely the case that pre-COVID samples were entirely attentive, in which case we would also need to adjust those estimates upward, implying that even accounting for growing inattention cannot fully explain smaller effect sizes during COVID-19. We emphasize that this does not constitute proof that inattention explains the attenuated effect sizes across our replications. Instead, we offer it as a partial explanation that is consistent with the evidence provided in Figure 5 and the broader trend of declining attentiveness among online survey respondents during the early onset of the pandemic.

4 Discussion

Our goal in this replication study was to understand whether online experiments conducted during the COVID-19 pandemic would generalize to other periods. We investigated by conducting 33 replications during the early COVID era of 12 published experimental studies. Overall, we find strong evidence that these experiments replicated in terms of sign and significance, but at reduced magnitudes. Because these pre-COVID experiments were, with one exception, successfully replicated we infer that the pandemic does not pose a fundamental threat to the generalizability of results from online experiments conducted during this period.

We considered two ways in which treatment effects might have been different during this period. First, we considered the possibility that the pandemic and the significant changes it brought to all aspects of daily life changed individuals' attitudes and patterns of information processing. If so, then the same experiment conducted on otherwise similar groups of people (online survey takers) before and during the pandemic might yield fundamentally different results. We do not find support for this idea. Second, it's possible that the pandemic changed, or occurred alongside changes to, the online survey respondent pool. Decreased attentiveness, as documented by others (Arechar and Rand, 2020; Aronow et al., 2020), provides some indication that the composition of online samples changed, at least during the early onset of the pandemic.

Using a variety of means – attention check questions, subject meta-data, and external estimates of attentiveness – we conclude that declining attentiveness in online samples may at least partly explain why the replication estimates presented here are, on average, smaller than pre-COVID benchmarks. Inattention can reduce respondents' compliance with experimental stimuli (shrinking average treatment effect estimates towards zero) and it adds nonrandom measurement error to outcome variables (decreasing the precision of treatment effect estimates). One important implication is that rising inattention in online samples can decrease the false positive rate at the cost of increasing the false negative rate. That is, inattention can bias estimates in the conservative direction such that small but substantively important treatment effects may be harder to detect.

Ultimately, answering the empirical question of whether treatment effects are equivalent for those who are attentive and those who are not requires a design that induces the inattentive to pay attention. Unfortunately, prior research has shown that inducing attentiveness among inattentive subjects is difficult (Berinsky, Margolis and Sances, 2016). We experienced these same challenges in our own work. In a separate Lucid study, we attempted to induce attentiveness by randomly assigning some respondents who failed an initial attention check to a condition in which we told them that they failed and gave them a second chance to pass. Only 17 of the 410 subjects (4%) in this treatment group passed when specifically reminded to re-read the prompt carefully because they had missed something, suggesting it is still difficult to induce attentiveness.

We therefore recommend that researchers concerned about inattention in online samples 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 inattentive respondents at the design stage, or retain them and use attention checks to compare estimates between attentive and inattentive subgroups at the analysis stage. We caution, however, that the former strategy necessarily restricts inferences to the sub-group of respondents that passed an attention check task, and these respondents may differ from those that failed on other dimensions as well (see Berinsky, Margolis and Sances, 2014, Online Appendix C).

In sum, we conclude that the pandemic does not pose a fundamental threat to the generalizability of online experiments, provided respondents pay attention to treatment content and outcome questions. People still prefer riskier options for responding to unusual disease outbreaks when in a loss frame – even when the real-world analogue could hardly be more

salient. People still believe misinformation, but can be corrected with 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, we find mostly ordinary responses to treatment in online samples.

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.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351(6280):1433–1436.

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. 2019. "Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents." *Research & Politics* 6(1):2053168018822174.

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.

Findley, Michael G, Kyosuke Kikuta and Michael Denly. 2020. "External Validity." *Annual Review of Political Science* .

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.

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.

IJzerman, Hans, Neil A Lewis, Andrew K Przybylski, Netta Weinstein, Lisa DeBruine, Stuart J Ritchie, Simine Vazire, Patrick S Forscher, Richard D Morey, James D Ivory et al. 2020. "Use caution when applying behavioural science to policy." *Nature Human Behaviour* 4(11):1092–1094.

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* .

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science* 349(6251).

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.

Rosenfeld, Daniel L, Emily Balcetis, Brock Bastian, Elliot Berkman, Jennifer Bosson, Tiffany Brannon, Anthony L Burrow, Daryl Cameron, CHEN Serena, Jonathan E Cook et al. forthcoming. "Conducting Social Psychological Research in the Wake of COVID-19." *Perspectives on Psychological Science.* .

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

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.