

# Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach

Alexander Coppock\*

July 26, 2017

Forthcoming at *Political Science Research and Methods*

## Abstract

To what extent do survey experimental treatment effect estimates generalize to other populations and contexts? Survey experiments conducted on convenience samples have often been criticized on the grounds that subjects are sufficiently different from the public at large to render the results of such experiments uninformative more broadly. In the presence of moderate treatment effect heterogeneity, however, such concerns may be allayed. I provide evidence from a series of 15 replication experiments that results derived from convenience samples like Amazon's Mechanical Turk are similar to those obtained from national samples. These results suggest that either the treatments deployed in these experiments cause similar responses for many subject types or convenience and national samples do not differ much with respect to treatment effect moderators. Using evidence of limited within-experiment heterogeneity, I show that the former is likely to be the case. Despite a wide diversity of background characteristics across samples, the effects uncovered in these experiments appear to be relatively homogeneous.

---

\*Alexander Coppock is Assistant Professor, Department of Political Science, Yale University. Portions of this research were funded by the Time-sharing Experiments in the Social Sciences (TESS) organization. This research was reviewed and approved by the Institutional Review Board of Columbia University (IRB-AAAP1312). The author is grateful to James N. Druckman and Donald P. Green for their guidance and support.

The generalizability of an experiment is the extent to which it generates knowledge about causal relationships that is applicable beyond the narrow confines of the research site. How much further beyond such confines results generalize is often a matter of great scientific importance. Vigorous discussion of the generalizability of findings occurs across all social scientific fields of inquiry (in economics: Levitt and List (2007); psychology: Sears (1986); Henrich, Heine and Norenzayan (2010), education: Tipton (2013); sociology: Lucas (2003); and political science: McDermott (2002)).

Concerns about the generalizability of findings usually fall into one of two categories: criticisms of the experimental context and criticisms of the experimental subjects. The first concern is a worry that an effect measured under experimental conditions is a poor guide to the effect in the “real world.” For example, a classic critique of survey experiments that investigate priming is that the effect, while “real” in the sense of being reliably reproducible in the survey context, is unimportant for the study of politics because primes have fleeting effects and political communication takes place in a competitive space where the marginal impact of a prime is likely to be canceled by a counterprime (Chong and Druckman 2010). Laboratory studies face a similar critique. Subjects may respond to the artificial context and obtrusive measurement in the lab in ways that do not generalize to the field.<sup>1</sup> Even well-designed field studies are sometimes accused of failing to generalize to the relevant political or policy decision because of some missing contextual feature.

A separate critique concerns the extent to which the study subjects are similar to the main population of interest. Because experimental subjects are very rarely drawn at random from any well-defined population, disagreements over whether treatment effect estimates obtained on a specific sample generalize to other places and times can be difficult to resolve. Within the social sciences, a point of some contention has been the increase in the use of online convenience samples of experimental subjects, particularly samples obtained via Amazon’s Mechanical Turk (MTurk).<sup>2</sup> MTurk respondents often complete dozens of academic surveys over the course of a week, leading to concerns that such “professional” subjects are particularly savvy or susceptible to demand effects (Chandler et al. 2015). These worries have been tempered somewhat by empirical studies that find that the MTurk population responds in ways similar to other populations (e.g., Berinsky, Huber and Lenz (2012)), but concerns remain that subjects on MTurk differ from other subjects in both measured and unmeasured ways (Huff and Tingley 2015).

In the present study, I contribute to the growing literature on the replicability of survey experiments across platforms, following in the footsteps of two closely-related studies. Mullinix

---

<sup>1</sup>See Coppock and Green (2015) for an analysis of the published record of paired lab and field studies that, contrary to the prediction of low lab generalizability, finds a strong correspondence across settings.

<sup>2</sup>As noted by Krupnikov and Levine (2014), criticisms of MTurk are often made on blogs rather than in academic journals. See, e.g., Kahan (2013).

et al. (2015) replicate 20 experiments and find a high degree of correspondence between estimates obtained on Mechanical Turk and on national probability samples, with a cross-sample correlation of 0.75.<sup>3</sup> Krupnikov and Levine (2014) find that when treatments are expected to have different effects by subgroup, cross-sample correspondence is weaker. The correlation between their MTurk and YouGov estimates is 0.41.<sup>4</sup> To preview the results presented below, I find a strong degree of correspondence between national probability samples and Mechanical Turk: the cross-sample correlation of 40 pairs of average treatment effect estimates derived from 12 pairs of studies is 0.85 (df = 38).

The remainder of this article will proceed as follows. First, I will present a theoretical framework that shows how the extent of treatment effect heterogeneity determines the generalizability of findings to other places and times. I will then present results from 15 replication studies, showing that in large part, original findings are replicated on both convenience and probability samples. I attribute this strong correspondence to the overall lack of treatment effect heterogeneity in these experiments; I conduct formal tests for unmodeled heterogeneity to bolster this claim.

## 1 Treatment Effect Heterogeneity and Generalizability

Findings from one site are generalizable to another site if the subjects, treatments, contexts, and outcome measures are the same – or would be the same – across both sites (Cronbach, Shapiro et al. 1982; Coppock and Green 2015). I define a *site* as the time, place, and group of units at and among which a causal process may hold. The most familiar kind of site is the research setting, with a well defined group of subjects, a given research protocol, and implementation team. Typically, the purpose of an experiment conducted at a one site is to generalize knowledge to other sites where no experiment has been conducted. For example, after an experiment conducted in one school district finds that a new curriculum is associated with large increases in student learning, we use that knowledge to infer both what would happen if the new curriculum were implemented in a different district, and what must have happened in the places where the curriculum was already in place.

The inferential target of a survey experiment conducted on a national probability sample of respondents is (often) the average treatment effect in the national population at a given moment in time, the Population Average Treatment Effect (PATE). The site of such an experiment might be an online survey administered to a random sample of adult Americans in 2012. The infer-

---

<sup>3</sup>This figure obtained from private correspondence with the authors.

<sup>4</sup>I am grateful to Yanna Krupnikov for sharing the replication data for this study from which this correlation was calculated. In the course of reanalyzing the data, it was discovered that the statistical routine originally used to compare samples overstated the confidence with which many of the YouGov/MTurk differences could be deemed significant.

ential target of the same experiment conducted on a convenience sample is a Sample Average Treatment Effect (SATE), where the sample in question is not drawn at random from the population. If a treatment engenders *heterogeneous* effects such that the distribution of treatment effects among those in the convenience sample is different from the distribution in the population, then the SATE will likely be different from the PATE.

When treatments, contexts, and outcome measures are held constant across sites, the generalizability of results obtained from one site to other sites depends on the degree and nature of treatment effect heterogeneity. If treatments have constant effects (that is, treatment effects are homogeneous), then the peculiarities of the experimental sample are irrelevant: what is learned from any subgroup can be generalized to any other population of interest. When treatments have heterogeneous effects, then the experimental sample might be very consequential. In order to assert that findings from one site are relevant for another site, a researcher must have direct or indirect knowledge of the distribution of treatment effects in both sites.

To illustrate the relationship between generalizability and heterogeneity, Figure 1 displays the potential outcomes and treatment effects for an entire population in two different scenarios. In the first scenario (represented in the left column of panels), treatment effects are heterogeneous, whereas in the second scenario, treatment effects are constant. The top row of panels displays the treated and untreated potential outcomes of the subjects and the bottom row displays the individual-level treatment effects (the difference between the treated and untreated potential outcomes).

An unobserved characteristic ( $U$ ) of subjects is plotted on the horizontal axes of Figure 1. This characteristic represents something about subjects that is related to both their willingness to participate in survey experiments and their political attitudes. In both scenarios,  $U$  is negatively related to the untreated potential outcome ( $Y_0$ ): higher values of  $U$  are associated with lower values of  $Y_0$ . This feature of the example represents how convenience samples may have different baseline political attitudes. Subjects on Mechanical Turk, for example, have been shown to hold more liberal attitudes than the public at large (Berinsky, Huber and Lenz 2012).

The untreated potential outcomes do not differ across scenarios, but the treated potential outcomes do. In scenario 1, effects are heterogeneous: higher values of  $U$  are associated with higher treatment effects, though the relationship depicted here is nonlinear. In scenario 2, treatment effects are homogeneous, so the unobserved characteristic  $U$  is independent of the differences in potential outcomes.

In both scenarios, imagine that two studies are conducted: one that samples from the entire population and a second that uses a convenience sample indicated by the shaded region. The population-level study targets the PATE, whereas the study conducted with the convenience sample targets a SATE. In scenario 1, the SATE and the PATE are different: generalizing from one research site to the other would lead to incorrect inferences. Note that this is a two-way street:

with only an estimate of the PATE in hand, a researcher would make poor inferences about the SATE and vice-versa. In scenario 2, the PATE and SATE are the same: generalizing from one site to another would be appropriate.

Figure 1: Implications of Treatment Effect Heterogeneity for Generalizability from Convenience Samples

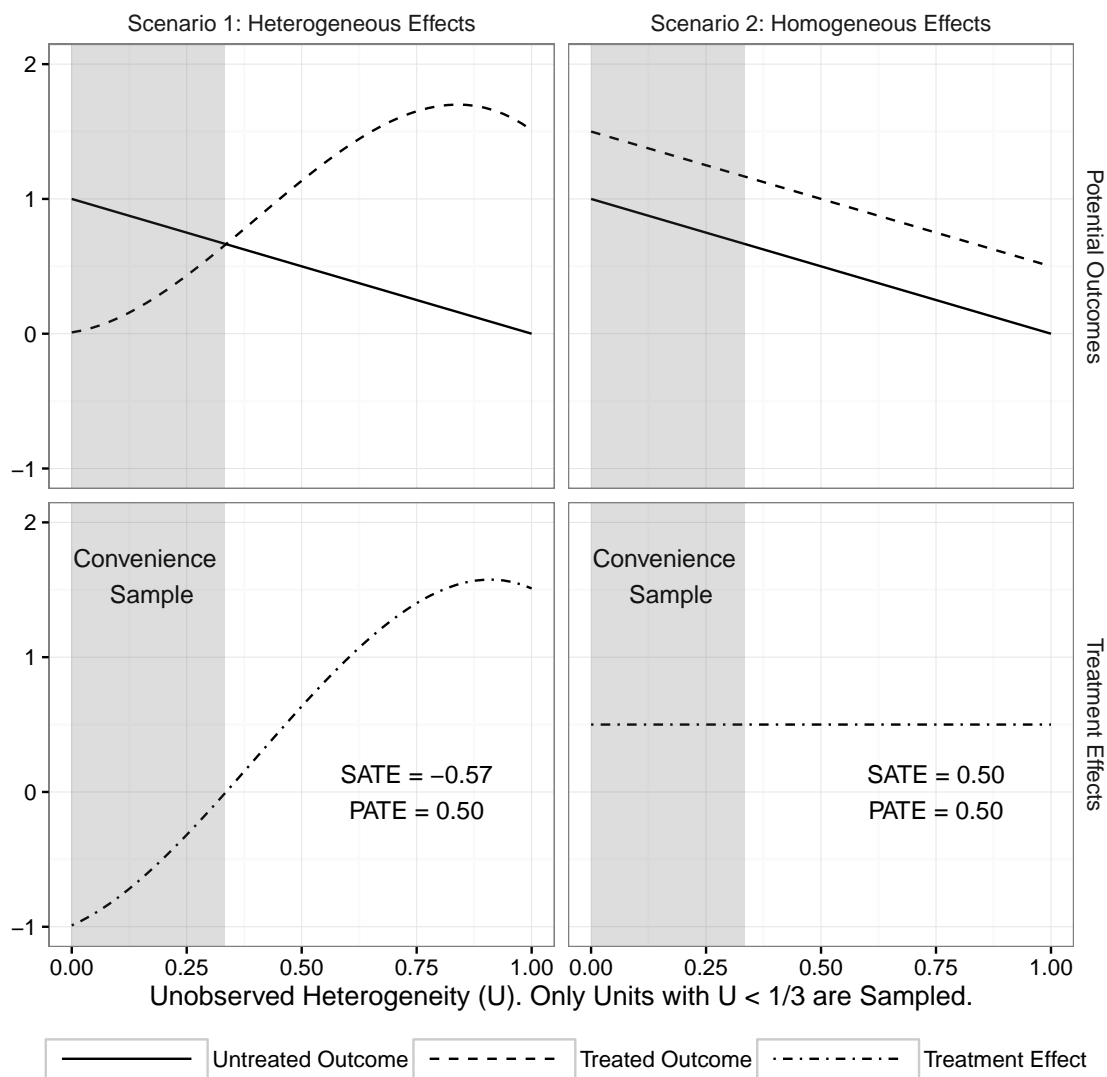


Figure 1 illustrates four points that are important to the study of generalizability. First, the fact that subjects may self-select into a study does not on its own mean that a study is not generalizable. Generalizability depends on whether treatment effect heterogeneity is independent of the (observed and unobserved) characteristics that determine self-selection. Second, even when

*baseline* outcomes ( $Y_0$ ) are different in a self-selected sample from some population, the study may still be generalizable. The relevant theoretical question concerns the differences between potential outcomes (the treatment effects), not their levels. Thirdly, when the characteristics that distinguish the population from that sample are unobserved, the PATE may not be informative about the SATE, i.e., experiments that target the PATE should not be privileged over those that use convenience samples unless the PATE is truly the target of inference. Finally, it is important to distinguish systematic heterogeneity from idiosyncratic heterogeneity. When treatment effects vary systematically with measurable subject characteristics, then the generalizability problem reduces to measuring these characteristics, estimating conditional average treatment effects, then reweighting these estimates by post-stratification. This reweighting can lead to estimates of either the SATE or the PATE, depending on the relevant target of inference. However, when treatment effects vary according to some unobserved characteristic of subjects (that also correlates with the probability of participation in a convenience sample), then no amount of post-stratification will license the generalization of convenience sample results to other sites.

## **2 Results I: Replications of 12 Survey Experiments**

The approach adopted here is to replicate survey experiments originally conducted on nationally-representative samples with MTurk subjects and, in three cases, with fresh nationally-representative samples. The experiments selected for replication came in two batches. The first set of five (Haider-Markel and Joslyn 2001; Peffley and Hurwitz 2007; Transue 2007; Chong and Druckman 2010; Nicholson 2012) were selected for three reasons. First, as evidenced by their placement in top journals, these studies addressed some of the most pertinent political science questions. Second, these studies all had replication data available, either posted online, available on request, or completely described in summaries published in the original article. Finally, they were all conducted on probability samples drawn from some well-defined population. As shown in Table 1, the target population was not always the U.S. national population. For example, in Haider-Markel and Joslyn (2001), the target of inference is the PATE among adult Kansans in 1999.

The second set of seven replications grew out of a collaboration with Time-Sharing Experiments for the Social Sciences (TESS), an NSF-supported organization that funds online survey experiments conducted on a national probability sample administered by GfK. These studies are of high quality due in part to the peer-review of study designs prior to data collection and to the TESS data-transparency procedures, by which raw data are posted one year after study completion. I selected seven studies, four of which (Brader 2005; Nicholson 2012; McGinty, Webster and Barry 2013; Craig and Richeson 2014) I replicated on Mechanical Turk, and three of which (Hiscox 2006; Hopkins and Mummolo 2017; Levendusky and Malhotra 2015) I replicated both

on Mechanical Turk and TESS/GfK. None of the seven I replicated were in the set of TESS studies replicated Mullinix et al. (2015). By and large, the replications were conducted with substantially larger samples than the original studies. All replications were conducted between January and September 2015. Mechanical Turk subjects were paid \$1.00 for 5-10 minutes of their time and were eligible to participate if they were U.S. residents and had a HIT acceptance rate above 95%.

Table 1: Study Manifest

Citation	Sampling Frame	N estimates	Original N	Replications N	
				MTurk	TESS/GfK
Haider-Markel and Joslyn (2001)	Kansans in 1999 (RDD)	1	518	1,009	
Brader (2005)	Americans in 2003 (TESS/KN)	2	281	1,709	
Peffley and Hurwitz (2007)	Black and White Americans in 2000-2001 (SRC)	1	905	1,285	
Transue (2007)	White Americans in the Twin Cities area in 1998 (MMIS)	2	345	367	
Chong and Druckman (2010)	Americans in 2009 (Bovitz)	9	1,302	1,887	
Nicholson (2012)	Americans in 2008 (YouGov)	2	1,491	1,249	
McGinty, Webster and Barry (2013)	Americans in 2012 (TESS/GfK)	6	2,935	2,487	
Craig and Richeson (2014)	Americans in 2012 (TESS/GfK)	2	611	709	
Johnston and Ballard (2016)	Americans in 2013 (TESS/GfK)	5	2,041	2,985	
Hiscox (2006)	Americans in 2003 (TESS/CSR)	4	1,610	2,972	2,084
Levendusky and Malhotra (2015)	Americans in 2012 (TESS/GfK)	2	1,041	1,987	1,411
Hopkins and Mummolo (2017)	Americans in 2011 (TESS/GfK)	4	3,269	2,972	3,189

The studies cover a wide range of issue areas – gun control, immigration, the death penalty, the Patriot Act, home foreclosures, mental illness, free trade, health care, and polarization – and generally employ framing, priming, or information treatments designed to persuade subjects to change their political attitudes.<sup>5</sup> The particulars of each study’s treatments and outcome measures are detailed in the online appendix. With very few (minor) exceptions, the stimulus mate-

<sup>5</sup>The substantive results of these experiments, while not the focus of the present study, indicate that subjects can be persuaded to change their political opinions in the direction of the appeal by approximately 0.2 standard deviations. For further information, see [citation withheld] or the online appendix.

rials and outcome measures were identical across original and replication versions of the experiments.

The number of treatment effect estimates in each study (reported in Table 1) is a function of the number of treatment arms and dependent variables. In most cases, the studies employ multiple treatment arms and consider effects on a single dependent variable. In some cases, however, a single treatment versus control comparison is considered with respect to a range of dependent variables. I limited the number of dependent variables analyzed in each study to two. Where possible, I followed the original authors' choices about which two dependent variables were most theoretically important; sometimes I had to use my best judgment to decide which were the two "main" outcome variables. I acknowledge that these choices introduce some subjectivity into the analysis. Mullinix et al. (2015) address this problem by focusing the analysis on the "first" dependent variable in each study, as determined by the temporal ordering of the dependent variables in the original TESS protocol. Their approach has the advantage of being automatically applicable across all studies but is no less subjective.

A wide range of analysis strategies were used in the original publications, including difference-in-means, difference-in-differences, ordinary least squares with covariate adjustment, subgroup analysis, and mediation analysis. To keep the analyses comparable, I reanalyzed all the original experiments using difference-in-means without conditioning on subgroups or adjusting for background characteristics.<sup>6</sup> Survey weights were incorporated where available. In all cases, I employed the Neyman variance estimator, equivalent to a standard variant of heteroskedasticity-robust standard errors (Samii and Aronow 2012). I standardized all dependent variables by dividing by the control group standard deviation in the original study. I used the identical specifications across all versions of a study. The sample sizes reported in Table 1 refer to the number of subjects included in the analyses here. In some cases, this number is smaller than the sample size reported in the original articles because I have omitted some treatment arms (e.g., placebo treatments that do not figure in the main comparison of interest).

The study-by-study results<sup>7</sup> are presented in Figure 2. On the horizontal axis of each facet, I have plotted the point estimate of the ATE with 95% confidence intervals. The scale of the horizontal axis is different for each study, for easy inspection of the within-study correspondence across samples. On the vertical axes, I have plotted each treatment versus control comparison, by study version. The top nine facets compare two versions of each study (original and MTurk), while the bottom three compare three versions (original, MTurk, and TESS/GfK). The coefficients are ordered by the magnitude of the original study effects from most negative to most positive. This plot gives a first indication of the overall success of the replications: in no cases do the

---

<sup>6</sup>In the case of factorial designs, I used OLS to obtain a single set of coefficients for each factor, averaging over the other margins. In one case, I adjusted for blocks to account for the randomization scheme.

<sup>7</sup>See Appendix Table A.1 for these results in a tabular format.



replications directly contradict one another, though there is some variation in the magnitudes of the estimated effects.

All together, I estimated 40 original-MTurk pairs of coefficients. Of the 25 coefficients that were originally significant, 18 were significant in the MTurk replications, all with the correct sign. Of the 15 coefficients that were not originally significant, 11 were not significant in the MTurk replications either, for an overall replication rate (narrowly defined) of  $(18 + 11)/(25 + 15) = 72.5\%$ . In zero cases did two versions of the same study return statistically significant coefficients with opposite signs. The match rate of the sign and statistical significance of coefficients across studies is a crude measure of correspondence, as it conflates the power of the studies with their correspondence. For example, if all pairs of studies had a power of 0.01, i.e., they only had a 1% chance of correctly rejecting a false null hypothesis of no average effect, then the match rate across pairs of studies would be close to 100%: nearly all coefficients would be deemed statistically insignificant. Measures of the replication rate that use the fraction of original studies within the 95% confidence interval of the replication (Open Science Collaboration 2015) or vice-versa (Gilbert et al. 2016) face a similar problem: lower powered studies appear to replicate at higher rates.

A better measure of the “replication rate” is the correlation of the standardized coefficients. Rather than relying on the artificial distinction between significant and nonsignificant, the correlation is a straightforward summary of the extent to which larger effects in the original studies are associated with larger effects in the replications. In the case of these 12 pairs of studies (40 coefficients), the correlation between the MTurk and original coefficients is 0.85 (df = 38). This raw correlation is likely an underestimate of the true correlation in treatment effects because all pairs of treatment effect estimates are measured with noise. Figure 3 plots each coefficient with 95% confidence intervals for both the original and MTurk versions.

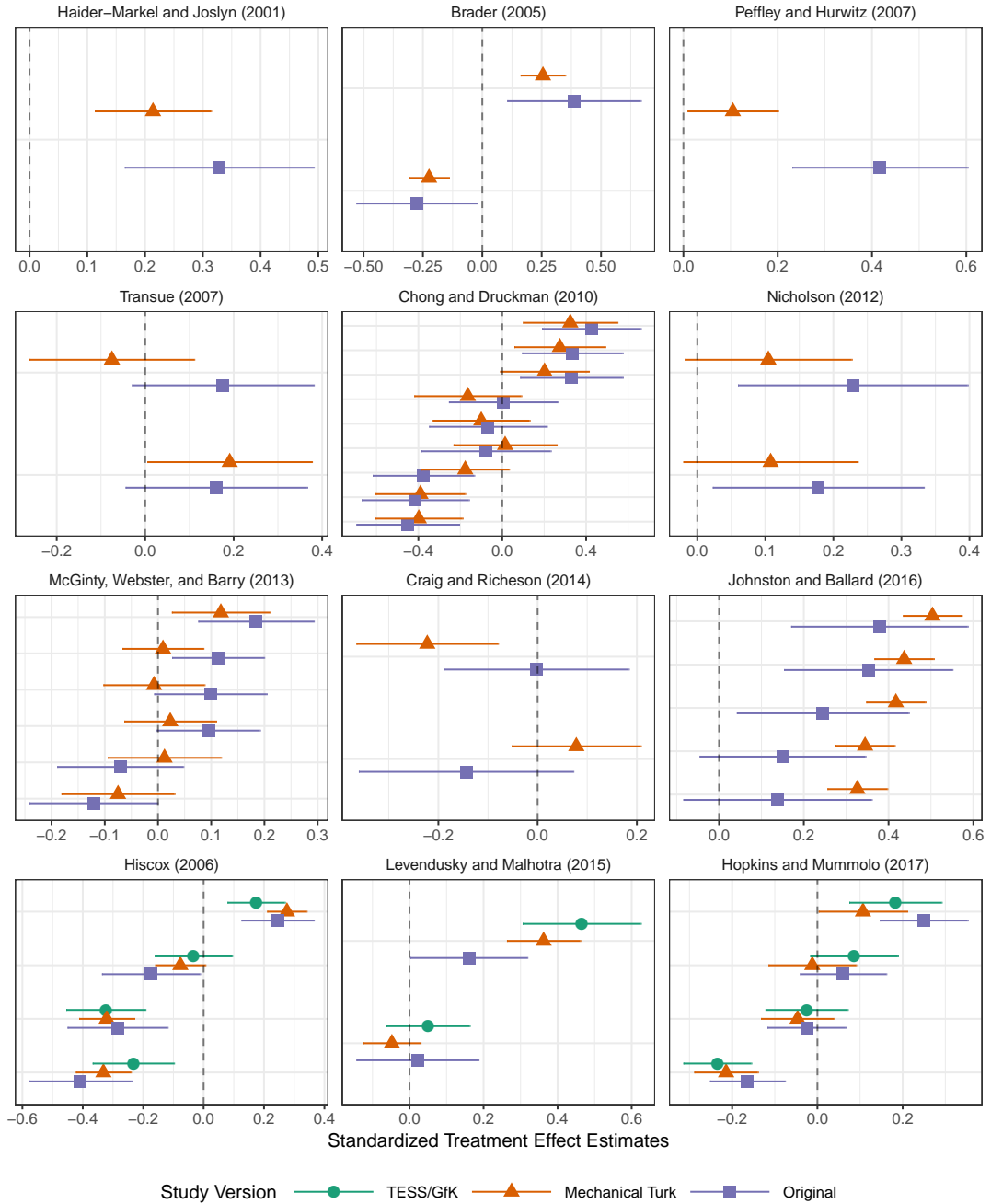
For the three studies (10 coefficients) replicated in parallel on MTurk and on fresh TESS/GfK samples (Hiscox 2006; Hopkins and Mummolo 2017; Levendusky and Malhotra 2015), the replication picture is even rosier. The correlation of the MTurk and original estimates is 0.90 (df = 8); TESS/GfK estimates with the MTurk estimates, 0.96 (df = 8); TESS/GfK with the original estimates, 0.85 (df = 8).

All in all, these results show a strong pattern of replication across samples, lending credence to the idea that at least for the sorts of experiments studied here, estimates of causal effects obtained on Mechanical Turk samples tend to be similar to those obtained on probability samples.

### **3 Results II: Testing the Null of Treatment Effect Homogeneity**

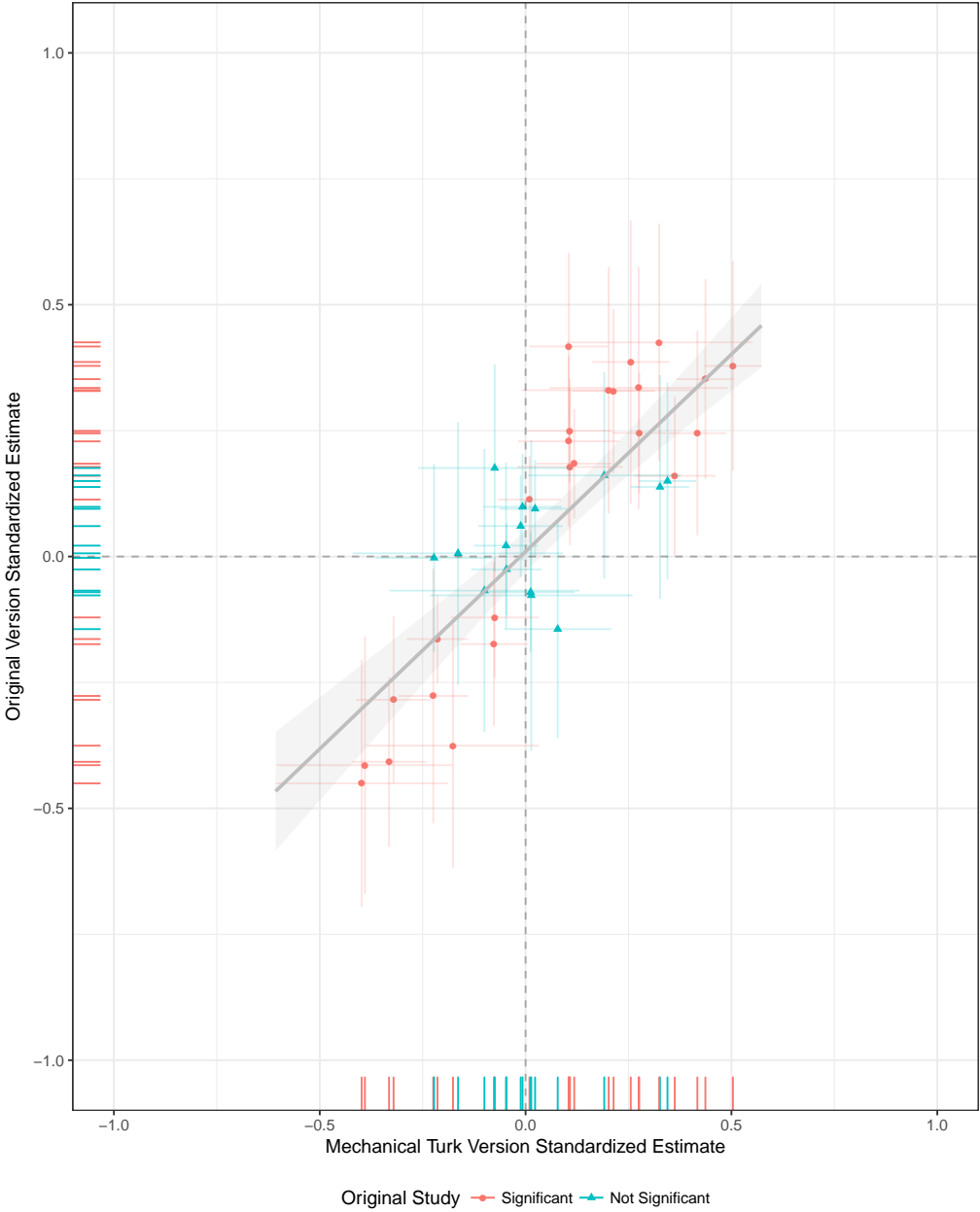
What can explain the strong degree of correspondence between the Mechanical Turk and probability sample estimates of average causal effects? It could be that effects are heterogeneous

Figure 2: Original and Replication Results of 12 Studies



Note: Each facet represents a study. Original treatment effect estimates are plotted with square points and replication treatment effect estimates are plotted with triangular (Mechanical Turk) or circular (TESS/GfK) points.

Figure 3: Original Coefficient Estimates Versus Estimates Obtained on Mechanical Turk



Note: Each point represents a standardized treatment effect estimate in both the original study and the Mechanical Turk replication study

within each sample – but this heterogeneity works out in such a way that the average effects across samples are approximately equal. This explanation is not out of the realm of possibility, and future work should consider whether the conditional average treatment effects (CATEs) estimated within well-defined demographic subgroups (e.g., race, gender, and partisanship) also correspond across samples.

In this section, I report the results of empirical tests of a second theoretical explanation for the generalizability of these results across research sites: the treatments explored in this series of experiments have constant effects. If effects are homogeneous across subjects, then the differences in estimates obtained from different samples would be entirely due to sampling variability.

Building on advances in Fisher permutation tests, Ding, Feller and Miratrix (2015) propose a test of treatment effect heterogeneity that compares treated and control outcomes with a modified Kolmogorov-Smirnov (KS) statistic. The traditional KS statistic is the maximum observed difference between two cumulative distribution functions (CDFs), and is useful for summarizing the overall difference between two distributions. The modified KS statistic compares the CDFs of the de-meaned treated and control outcomes, thereby removing the estimated average treatment effect from the difference between the distributions and increasing the sensitivity of the test statistic to treatment effect heterogeneity.

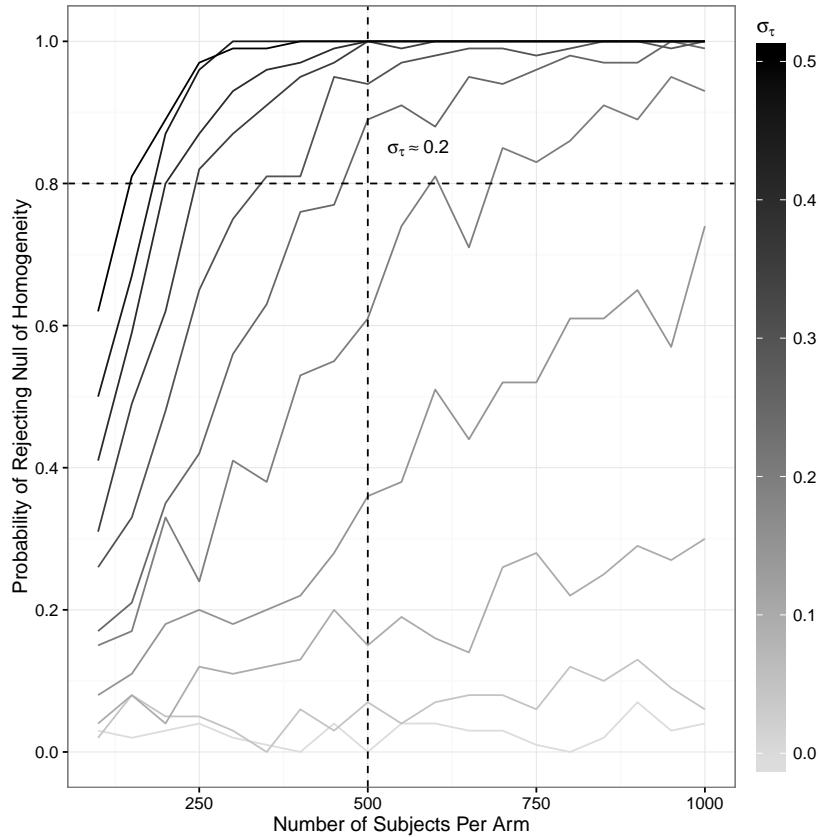
The permutation test compares the observed values of the modified KS statistic to a simulated null distribution. This null distribution is constructed by imputing the missing potential outcomes for each subject under the null of constant effects, then simulating the distribution of the modified KS statistic under a large number of possible random assignments. One difficulty is choosing *which* constant effect to use for imputation. A natural choice is to use the estimated ATE; however, as Ding, Feller and Miratrix (2015) show, this approach may lead to incorrect inferences. Instead, they advocate repeating the test for all plausible values of the constant ATE and reporting the maximum p-value. In practice, set of “all plausible” values of the ATE is approximated by the 99.99% confidence interval around the estimated ATE.

This test, like other tests for treatment effect heterogeneity (Gerber and Green 2012, pp. 293-294), can be low-powered. To gauge the probability of correctly rejecting a false null hypothesis, I conducted a small simulation study that varied two parameters: the number of subjects per treatment arm and the degree of treatment effect heterogeneity. Equation 1 shows the potential outcomes function for subject  $i$ , where  $Z_i$  is the treatment indicator and can take values of 0 or 1,  $\sigma_\tau$  is the standard deviation of the treatment effects, and  $X_i$  is an idiosyncratic characteristic, drawn from a standard normal distribution. The larger  $\sigma_\tau$ , the larger the extent of treatment effect heterogeneity, and the more likely the test is to reject the null of constant effects. To put  $\sigma_\tau$  in perspective, consider a treatment with enormous effect heterogeneity: large positive effects of 0.5 standard units for half the sample, but large negative effects of 0.5 standard units for the other half. In this case the standard deviation of the treatment effects would be equal to 0.5. While the

simple model of heterogeneity used for this simulation study does not necessarily reflect how the test would perform in other scenarios, it provides a first approximation of the sorts of heterogeneity typically envisioned by social scientists.

$$Y_i = 0 \cdot Z_i + \sigma_\tau \cdot X_i \cdot Z_i + X_i \tag{1}$$

Figure 4: Simulation Study: Power of the Randomization Test



The results of the simulation study are presented in Figure 4. The MTurk versions of the experiments studied here typically employ 500 subjects per treatment arm, suggesting that we would be well powered (power  $\approx 0.8$ ) to detect treatment effect heterogeneity on the scale of 0.2 standard deviations, and moderately powered for 0.1 standard deviations (power  $\approx 0.6$ ).

The main results of the heterogeneity test applied to the present set of 27 studies (12 original and 15 replications) are displayed in Table 2. Among the original 12 studies, just 1 of the 40 treatment versus control comparisons revealed evidence of effect heterogeneity. Among the Mechanical Turk replications, 8 of 40 treatments were shown to have heterogeneous effects. On

the TESS/GfK replications, 0 of 10 tests were significant. In only one case (Hiscox 2006), did the same treatment versus control comparisons return a significant test statistic across samples. In order to guard against drawing false conclusions due to conducting many tests, I present the number of significant tests after correcting the  $p$ -values by the Holm correction (Holm 1979) in the last column of Table 2.<sup>8</sup>

Table 2: Tests of Treatment Effect Heterogeneity

Study Site	N Comparisons	N Significant	N Significant (Holm)
Original	40	1	0
Mechanical Turk	40	8	7
TESS/GfK	10	1	0

Far more often than not, we fail to reject the null of treatment effect homogeneity. Because we are relatively well powered to detect politically-meaningful differences in treatment response, I conclude from this evidence that the treatment effect homogeneity explanation of the correspondence across experimental sites is plausible.

## 4 Discussion

Levitt and List (2007, p. 170) remind us that “Theory is the tool that permits us to take results from one environment to predict in another[.]” When the precise nature of treatments varies across sites, we need theory to distinguish the meaningful differences from the cosmetic ones. When the contexts differ across sites – public versus private interactions, field versus laboratory observations – theory is required to generalize from one to the other. When outcomes are measured differently, we rely on theory to predict how a causal process will express itself across sites.

In the results presented above, I have not made theoretical claims about the differences in treatments, contexts, or outcome measures across sites – they were all held constant by design. The only remaining impediment to the generalizability of the survey experimental findings from convenience samples to probability samples is the composition of the subject pools. If treatments have heterogeneous effects, researchers have to be careful not to generalize from a sample that has one distribution of treatment effects to populations that have different distributions of effects.

<sup>8</sup>The Holm correction controls the family-wise error rate under the same assumptions as the more familiar Bonferroni correction, but is strictly more powerful (Holm 1979). To implement the correction, order the  $m$  uncorrected  $p$ -values within a “family” from smallest to largest. Identify the smallest  $p$ -value for which the following condition holds:  $p_k \leq \frac{k}{m}\alpha$ , where  $k$  indexes the  $p$ -values, and  $\alpha$  is the target significance level. The smallest  $p$ -value that meets this condition is insignificant, as are all larger  $p$ -values. All smaller  $p$ -values are significant.

Before this replication exercise (and others like it such as Mullinix et al. (2015) and Krupnikov and Levine (2014)), social scientists had a limited empirical basis on which to develop theories of treatment effect heterogeneity for the sorts of treatments explored here, which by and large attempt to persuade subjects of policy positions. Both within and across samples, the treatments studied here have exhibited muted treatment effect heterogeneity. Whatever differences (measured and unmeasured) there may be between the Mechanical Turk population and the population at large, they do not appear to interact with the treatments employed in these experiments in politically meaningful ways. In my view, it is this lack of heterogeneity that explains the overall correspondence across samples.

I echo the concerns of Mullinix et al. (2015), who caution that the pattern of strong probability/convenience sample correspondence does not imply that all survey experiments can be conducted with opt-in Internet samples with no threats to inference more broadly. Indeed, the crucial question is the extent of treatment effect heterogeneity. Some treatments of course have different effects for different subgroups and in such cases, an estimate obtained from a convenience sample may not generalize well. Future disagreements about whether a convenience sample can serve as a useful database from which to draw general inferences should be adjudicated on the basis of rival theories concerning treatment effect heterogeneity (or large, well-powered empirical demonstrations of such theories). Crucially, simply noting that convenience and probability samples differ in terms of their background characteristics is not sufficient for dismissing the results of experiments conducted on convenience samples.

Moreover, in an age of 9% response rates (Keeter et al. 2017), even probability samples can only be considered representative of the population under the strong assumption that, after reweighting or poststratification, no important differences remain between those who respond to the survey and the population. Probability samples may also only generalize to the moment in time they were conducted. Future research into the generalizability of treatment effects should also compare the extent to which probability samples drawn at one time correspond with probability samples drawn years later. Similar concerns may also extend to convenience samples, as some work suggests that the composition of the Mechanical Turk subject pool has changed in important ways over time (Stewart et al. 2015).

Finally, it is worth reflecting on the remarkable robustness of the experiments replicated here. Concerns over *p*-hacking (Simonsohn, Nelson and Simmons 2014), fishing (Humphreys, Sanchez de la Sierra and van der Windt 2013), data snooping (White 2000), and publication bias (Franco, Malhotra and Simonovits 2014; Gerber et al. 2010) have lead many to express a great deal of skepticism over the reliability of results published in the scientific record (Ioannidis 2005). An effort to replicate 100 papers in psychology (Open Science Collaboration 2015) was largely viewed as a failure, with only one-third to one-half of papers replicating, depending on the measure. By contrast, the empirical results in this set of experiments were largely confirmed.

I speculate that this high replication rate may be explained in part by the procedure used to select studies for replication in the first place. I chose studies that had replication data available and whose treatment effect estimates were relatively precise. Further, the studies originally conducted on TESS underwent pre-implementation peer review, a process that might have either excluded theoretically tenuous studies or improved the design of chosen studies. Because these studies may have been of unusually high quality relative to the modal survey experiment in the social sciences, we should exercise caution when generalizing from this set of replications to all studies conducted on convenience samples.

## A Appendix

Table A.1: Original and Replication Average Treatment Effect Estimates

Study	Dependent Variable	Treatment	Original	Mechanical Turk	TESS/GfK
Haider-Markel and Joslyn (2001)	Support for Concealed Carry Law	Citizens' Rights Frame	0.33 (0.08)*	0.21 (0.05)*	
		Positive Frame	-0.28 (0.13)*	-0.22 (0.04)*	
Brader (2005)	Support for Immigration	Positive Frame	0.39 (0.14)*	0.26 (0.05)*	
		African Americans	0.42 (0.09)*	0.10 (0.05)*	
Peffley and Hurwitz (2007)	Willingness to Pay Tax	Other Americans	0.18 (0.10)	-0.08 (0.10)	
		Public Schools	0.16 (0.10)	0.19 (0.09)*	
Transue (2007)	Willingness to Pay Tax	Public Schools	0.16 (0.10)	0.19 (0.09)*	
		Both / Memory Based	-0.07 (0.14)	-0.10 (0.12)	
Chong and Druckman (2010)	Patriot Act Support	Both / No Processing	-0.08 (0.16)	0.01 (0.13)	
		Both / Online Processing	0.01 (0.13)	-0.16 (0.13)	
	Patriot Act Support	Con / Memory Based	-0.38 (0.12)*	-0.18 (0.11)	
		Con / No Processing	-0.41 (0.13)*	-0.39 (0.11)*	
	Patriot Act Support	Con / Online Processing	-0.45 (0.12)*	-0.40 (0.11)*	
		Pro / Memory Based	0.33 (0.12)*	0.20 (0.11)	
	Patriot Act Support	Pro / No Processing	0.43 (0.12)*	0.32 (0.11)*	
		Pro / Online Processing	0.33 (0.12)*	0.27 (0.11)*	
Nicholson (2012)	Support for Foreclosure Bill	In Party Cue	0.18 (0.08)*	0.11 (0.07)	
		In Party Cue	0.23 (0.09)*	0.10 (0.06)	
McGinty, Webster, and Barry (2013)	Magazines	LCM Ban	0.18 (0.06)*	0.12 (0.05)*	
		Mental Illness	0.10 (0.05)	-0.01 (0.05)	
	Magazines	News	0.11 (0.04)*	0.01 (0.04)	
		LCM Ban	-0.07 (0.06)	0.01 (0.05)	
	SMI Danger	Mental Illness	-0.12 (0.06)*	-0.08 (0.05)	
		News	0.09 (0.05)	0.02 (0.04)	
Craig and Richeson (2014)	Support for Immigration	Majority Minority	-0.00 (0.10)	-0.22 (0.07)*	
		Majority Minority	-0.14 (0.11)	0.08 (0.07)	
Johnston and Ballard (2016)	Way of Life	Expert Treatment	0.38 (0.11)*	0.50 (0.04)*	
		Expert Treatment	0.15 (0.10)	0.34 (0.04)*	
	Agree on Immigration	Expert Treatment	0.14 (0.11)	0.33 (0.04)*	
		Expert Treatment	0.35 (0.10)*	0.44 (0.04)*	
	Agree on Trade with China	Expert Treatment	0.24 (0.10)*	0.42 (0.04)*	
Hiscox (2006)	Support for Free Trade	Expert	0.25 (0.06)*	0.28 (0.03)*	0.17 (0.05)*
		Negative	-0.28 (0.08)*	-0.32 (0.05)*	-0.32 (0.07)*
	Support for Free Trade	Pos + Neg	-0.41 (0.09)*	-0.33 (0.05)*	-0.23 (0.07)*
		Positive	-0.17 (0.08)*	-0.08 (0.04)	-0.03 (0.07)
Levendusky and Malhotra (2015)	Extremity of Policy Views	Polarized Treatment	0.02 (0.08)	-0.05 (0.04)	0.05 (0.06)
		Polarized Treatment	0.16 (0.08)*	0.36 (0.05)*	0.47 (0.08)*
Hopkins and Mummolo (2017)	Support for Crime Spending	Crime Argument	0.06 (0.05)	-0.01 (0.05)	0.09 (0.05)
		Health Care Argument	-0.03 (0.05)	-0.05 (0.04)	-0.03 (0.05)
	Support for Stimulus Spending	Stimulus Argument	-0.16 (0.04)*	-0.21 (0.04)*	-0.23 (0.04)*
		Terrorism Argument	0.25 (0.05)*	0.11 (0.05)*	0.18 (0.06)*

\*p < .05

Robust standard errors are in parentheses.

Dependent variables measured in standard units.



## References

- 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.
- Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49(2):388–405.
- Chandler, Jesse, Gabriele Paolacci, Eyal Peer, Pam Mueller and Kate Ratliff. 2015. "Non-Naïve Participants Can Reduce Effect Sizes." *Psychological Science* 26:1131–1139.
- Chong, Dennis and James N. Druckman. 2010. "Dynamic Public Opinion: Communication Effects over Time." *American Political Science Review* 104(04):663–680.
- Coppock, Alexander and Donald P. Green. 2015. "Assessing the Correspondence Between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research." *Political Science Research and Methods* 3(1):113–131.
- Craig, Maureen A. and Jennifer A. Richeson. 2014. "More Diverse Yet Less Tolerant? How the Increasingly Diverse Racial Landscape Affects White Americans' Racial Attitudes." *Personality and Social Psychology Bulletin* 40(6):750–761.
- Cronbach, Lee J, Karen Shapiro et al. 1982. *Designing Evaluations of Educational and Social Programs*. Jossey-Bass.
- Ding, Peng, Avi Feller and Luke Miratrix. 2015. "Randomization Inference for Treatment Effect Variation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .
- Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203):1502–1505.
- Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.
- Gerber, Alan S., Neil Malhotra, Conor M. Dowling and David Doherty. 2010. "Publication Bias in Two Political Behavior Literatures." *American Politics Research* 38(4):591–613.
- Gilbert, Daniel T., Gary King, Stephen Pettigrew and Timothy D. Wilson. 2016. "Comment on "Estimating the reproducibility of psychological science"." *Science* 351(6277):1037–1037.
- Haider-Markel, Donald P. and Mark R. Joslyn. 2001. "Gun Policy, Opinion, Tragedy, and Blame Attribution: The Conditional Influence of Issue Frames." *The Journal of Politics* 63(2):520–543.
- Henrich, Joseph, Steven J. Heine and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33:61–83.
- Hiscox, Michael J. 2006. "Through a Glass and Darkly: Attitudes Toward International Trade and the Curious Effects of Issue Framing." *International Organization* 60(03):755–780.
- Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* pp. 65–70.
- Hopkins, Daniel J. and Jonathan Mummolo. 2017. "Assessing the Breadth of Framing Effects." *Quarterly Journal of Political Science* .
- Huff, Connor and Dustin Tingley. 2015. ""Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research & Politics* 2(3):2053168015604648.

- Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21(1):1–20.
- Ioannidis, John P. A. 2005. "Why Most Published Research Findings are False." *PLoS Medicine* 2(8):e124.
- Johnston, Christopher D. and Andrew O. Ballard. 2016. "Economists and Public Opinion: Expert Consensus and Economic Policy Judgments." *The Journal of Politics* 78(2):443–456.
- Kahan, Dan M. 2013. "Fooled Twice, Shame on Who? Problems with Mechanical Turk Study Samples, Part 2."
- Keeter, Scott, Nick Hatley, Courtney Kennedy and Arnold Lau. 2017. What Low Response Rates Mean for Telephone Surveys. Technical report Pew Research Center.
- Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross-sample Comparisons and External Validity." *Journal of Experimental Political Science* 1(01):59–80.
- Levendusky, Matthew and Neil Malhotra. 2015. "Does Media Coverage of Partisan Polarization Affect Political Attitudes?" *Political Communication* .
- Levitt, Steven D. and John A. List. 2007. "What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *The journal of economic perspectives* pp. 153–174.
- Lucas, Jeffrey W. 2003. "Theory-testing, Generalization, and The Problem of External Validity." *Sociological Theory* 21(3):236–253.
- McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10(4):325–342.
- McGinty, Emma E., Daniel W. Webster and Colleen L. Barry. 2013. "Effects of News Media Messages about Mass Shootings on Attitudes Toward Persons with Serious Mental Illness and Public Support for Gun Control Policies." *American Journal of Psychiatry* 170(5):494–501.
- 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.
- Nicholson, Stephen P. 2012. "Polarizing Cues." *American Journal of Political Science* 56(1):52–66.
- Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251).
- Peffley, Mark and Jon Hurwitz. 2007. "Persuasion and Resistance: Race and the Death Penalty in America." *American Journal of Political Science* 51(4):996–1012.
- Samii, Cyrus and Peter M. Aronow. 2012. "On Equivalencies Between Design-based and Regression-based Variance Estimators for Randomized Experiments." *Statistics & Probability Letters* 82(2):365–370.
- Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51(3):515–530.
- Simonsohn, Uri, Leif D. Nelson and Joseph P. Simmons. 2014. "P-curve: A Key to the File-drawer." *Journal of Experimental Psychology: General* 143(2):534–547.
- Stewart, Neil, Christoph Ungemach, Adam JL Harris, Daniel M Bartels, Ben R Newell, Gabriele Paolacci and Jesse Chandler. 2015. "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers." *Judgment and Decision Making* 10(5):479.
- Tipton, Elizabeth. 2013. "Improving Generalizations From Experiments Using Propensity Score Subclassification Assumptions, Properties, and Contexts." *Journal of Educational and Behavioral Statistics* 38(3):239–266.
- Transue, John E. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51(1):78–91.
- White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica* 68(5):1097–1126.