The Generalizability of Heterogeneous Treatment Effect Estimates Across Samples

Alexander Coppock, Thomas J. Leeper, Kevin J. Mullinix*

February 22, 2018
Draft prepared for presentation at the 2018 SPSP Annual Convention March 1 - 3.

Abstract

The extent to which studies conducted with non-representative convenience samples are generalizable to broader populations depends critically on the level of treatment effect heterogeneity. Recent inquiries (e.g., Mullinix et al. 2015; Coppock forthcoming) have found a strong correspondence between average treatment effects estimated in nationally-representative experiments and in replication studies conducted with convenience samples. In this paper, we consider three possible explanations: low levels of effect heterogeneity, high levels of effect heterogeneity that are unrelated to selection into the convenience sample, or just good luck. We reanalyze 26 original-replication study pairs (encompassing 98,425 individual survey responses) to assess the extent to which a model of heterogeneity in treatment response estimated on the original dataset predicts the heterogeneity in the replication experiment, and vice-versa. While there are exceptions, the overwhelming pattern that emerges is one of treatment effect homogeneity, providing a partial explanation for strong correspondence across both unconditional and conditional average treatment effect estimates.

*Alexander Coppock is Assistant Professor of Political Science, Yale University, Thomas J. Leeper is Associate Professor in Political Behaviour, London School of Economics and Political Science, Kevin J. Mullinix is Assistant Professor of Political Science, Appalachian State University.

Randomized experiments are increasingly employed across the social sciences to study beliefs, opinions, and behavioral intentions (Druckman et al. 2006, 2011). Experiments are nevertheless sometimes met with skepticism about the degree to which results *generalize* (Gerring 2012). Indeed, it is often said that experiments achieve better internal validity than they do external validity because of the non-representative samples typical used in experimental research (e.g., Sears 1986; Cronbach 1986; McDermott 2011, though see Aronow and Samii 2015 for a critique of the claimed external validity of some nonexperimental regression estimates).

To date, the social science community has generated only limited theory and evidence to guide expectations about when a convenience sample and a target population are sufficiently similar to justify generalizing from one to the other. Sometimes demographic differences between, say, a student sample and the national population of the US are taken as *prima facie* evidence that results obtained on the student sample are unlikely to generalize. By contrast, several recent empirical studies suggest that convenience samples, despite drastic demographic differences, frequently yield average treatment effect estimates that are not substantially different from those obtained through nationally representative samples (Mullinix et al. 2015; Coppock forthcoming; Krupnikov and Levine 2014; Weinberg, Freese and McElhattan 2014).

On the one hand, such findings suggest that even in the face of differences in sample composition, claims of strong external validity are justified. On the other hand, the rough equivalence of sample average treatment effects (SATEs) in these experiments could be the result of: (1) effect homogeneity across participants such that sample characteristics are irrelevant, (2) effect heterogeneity that is approximately orthogonal to selection, or (3) effect heterogeneity that is not orthogonal to selection but works out "by chance." Arbitrating between these three explanations for between-sample similarity of SATEs is critical to assessing whether experimental findings *in general* are likely to be externally valid.

Disentangling these three possibilities requires going beyond the generalizability of *average* treatment effects to consider treatment effect heterogeneity. The generalizability of an experimental finding depends critically on the degree of individual-level effect heterogeneity. In order to advance the state of knowledge on why convenience samples seem to generalize, we focus on the between-study correspondence in patterns of effect heterogeneity.

Our main finding is straightforward. In this set of experiments, average effect estimates are approximately the same in convenience and national samples because treatment effect heterogeneity is muted. Although subjects differ tremendously in their baseline social and political attitudes, their responses to treatment appear quite similar.

1 Linking Generalizability to Effect Heterogeneity

Generalizability or “external validity” refers to whether the conclusions drawn from a given study can usefully apply or be extrapolated to other places and times. The commonly applied “SUTO” framework (Cronbach 1986; Shadish, Cook and Campbell 2002) predicts that generalizability will be higher, the more similar are the settings, units, treatments, and outcomes used across research sites. Because the setting, treatments, and outcomes are features of a research design that are easily under the control of experimenters, the subjects (units) of the study typically face the most scrutiny.[1] Mutz (2011) argues that probability-based sampling provides the subjects of the highest quality and that studies conducted on nonprobability samples exhibit lower external validity (though see chapter 8 for a nuanced discussion of the supposed tension between internal and external validity). As we will show, the degree to which otherwise identical experiments conducted on two different samples yield the same results hinges upon the degree to which the effect of an experimental treatment is systematically heterogeneous across individuals. While such effect *moderation* is a common topic in introductory statistics lessons (Baron and Kenny 1986), in applied research, and in meta-analysis, its essential role in debate about generalizability is underappreciated.

Understanding the link between individual-level treatment effects and study-level generalizability requires some formalization. We presume a large population (say, all Americans over the age of 18 in 2016; though it could be any population), each member i of which is endowed with an untreated potential outcome $Y_i(0)$ and a treated potential outcome $Y_i(1)$. The individual-level treatment effect τ_i is defined as the difference between these potential outcomes $\tau_i = Y_i(1) - Y_i(0)$. A common population parameter of interest is thus the average of all individual level treatment effects, the Population Average Treatment Effect (PATE). Population-based survey experiments sample (at least approximately) at random from the population, then further randomly assign subjects to treatment, or more abstractly, assign subjects to reveal either their untreated or treated potential outcome, but not both (Holland 1986). Due to representative sampling from the population *and* random assignment to experimental condition, such population-based survey experiments allow the Sample Average Treatment Effect (SATE) to provide an unbiased estimate of the PATE (Mutz 2011; Imai, King and Stuart 2008).

However, many survey experiments are in practice conducted on sets of participants that are not sampled at random from any well-defined population. Instead, partici-

[1] The survey-experimental studies we analyze here all hold setting, treatment, and outcome constant. All are online survey experiments conducted in the United States using identical experimental protocols, and most were conducted at roughly the same point in time. Variation in results across studies should be due to sample characteristics alone, all else being held constant by design.

pants often constitute a *convenience sample* (e.g., university students or workers recruited from an online platform like Amazon Mechanical Turk). The SATE in these cases remains the average difference in potential outcomes for the subjects in the sample. The SATE from a convenience sample need not equal the SATE for a population-representative sample or the PATE. For this reason, the use of such convenience samples is often subject to criticism regarding external validity (see Sears 1986; McDermott 2011; Mutz 2011).

Perhaps surprisingly, some empirical evidence (Mullinix et al. 2015; Coppock forthcoming) finds — despite the lack of a design-based justification for claims of generalizability of the SATE in such cases — a close correspondence between convenience sample SATEs and representative samples SATEs. This correspondence can occur if one of the following conditions holds:

1. Individual-level treatment effects τ_i are homogeneous ($Var(\tau_i) = 0$) for this class of experiments,

2. Individual-level treatment effects τ_i are heterogeneous ($Var(\tau_i) > 0$), but the pattern of heterogeneity is similar in both samples and unrelated to sample selection, or

3. Individual-level treatment effects τ_i are heterogeneous ($Var(\tau_i) > 0$) and the pattern of heterogeneity is different in the population as a whole and in the "population" from which the convenience sample is drawn, but the SATEs are approximately the same in the two samples by good luck.

Understanding which of these three stories is primarily responsible for the observed similarity of effect estimates for this class of survey experiments is important. If the first explanation (treatment effect homogeneity) is correct, the credibility of results obtained on nonprobability convenience samples is strengthened.[2] If the second explanation (treatment effect heterogeneity is unrelated to selection into sample) is correct, then we are left to worry whether the next experiment will share this fortuitous characteristic. Finally, if the third explanation is correct, the observed correspondence across population-based and convenience samples is meaningless – the effects from convenience samples are not generalizable, just lucky in the experiments that have been replicated across samples.

[2] We hasten to add, however, that even if effects are homogeneous in these particular studies, effects may be heterogeneous in others; that is, we do not yet know the extent to which our study of generalizability is itself generalizable. We limit our ambition to make claims about generalizability for the common class of survey experiments under examination here.

Each of these scenarios depends on the degree of effect heterogeneity and whether the sources of that heterogeneity are similar in the convenience sample and in the population as a whole. Consider, for example, a stylized scenario wherein we can measure a single observable trait (taking values A or B) for each unit. Half the population are As and half are Bs. In addition, all subjects have a trait that is not even in principle observable, which can be "high" or "low." Half of both the As and Bs are high types. Finally, suppose that that As are twice as likely to select into a convenience sample as B types. The population-based survey experiment will be half As and half Bs, but the convenience sample will be two-thirds As and one-third Bs. A summary of these descriptions is given in Table 1.

Table 1: Hypothetical Distribution of Types

Observable Type	Unobservable Trait	Treatment Effect	Population Proportion	Sample Proportion
A	low	c_1	1/4	1/3
A	high	c_2	1/4	1/3
B	low	c_3	1/4	1/6
B	high	c_4	1/4	1/6

The PATE will be a weighted average of the Conditional Average Treatment Effect (CATE) among these four groups:

$$\begin{aligned} PATE &= \pi_{pop,A} CATE_A + \pi_{pop,B} CATE_B \\ &= \frac{1}{2}\left(\frac{1}{2}c_1 + \frac{1}{2}c_2\right) + \frac{1}{2}\left(\frac{1}{2}c_3 + \frac{1}{2}c_4\right) \\ &= \frac{1}{4}c_1 + \frac{1}{4}c_2 + \frac{1}{4}c_3 + \frac{1}{4}c_4 \end{aligned}$$

The SATE is an analogous weighted average of CATEs:

$$\begin{aligned} SATE &= \pi_{smp,A} CATE_A + \pi_{smp,B} CATE_B \\ &= \frac{2}{3}\left(\frac{1}{2}c_1 + \frac{1}{2}c_2\right) + \frac{1}{3}\left(\frac{1}{2}c_3 + \frac{1}{2}c_4\right) \\ &= \frac{1}{3}c_1 + \frac{1}{3}c_2 + \frac{1}{6}c_3 + \frac{1}{6}c_4 \end{aligned}$$

This example allows us to further formalize the three scenarios just discussed to illuminate how a PATE might be equal to a SATE:

Scenario 1: Treatment effect homogeneity
Under treatment effect homogeneity $c_1 = c_2 = c_3 = c_4$, so regardless of the relative proportions of each subject type in either the population or any sample, $PATE = SATE$.

Scenario 2: Treatment effect heterogeneity orthogonal to selection
If $c_1 = c_3$ and $c_2 = c_4$, then $PATE = SATE$.

Scenario 3: Treatment effect heterogeneity not orthogonal to selection
If, however, $c_1 \neq c_3$ or $c_2 \neq c_4$ then selection and treatment effect heterogeneity are not orthogonal. The PATE may nevertheless equal the SATE by "good luck" if the conditional average treatment effects for each subgroup work out so that the following equality holds:

$$\frac{1}{4}c_1 + \frac{1}{4}c_2 + \frac{1}{4}c_3 + \frac{1}{4}c_4 = \frac{1}{3}c_1 + \frac{1}{3}c_2 + \frac{1}{6}c_3 + \frac{1}{6}c_4$$

Scenario 3 shows that even if we observe PATEs and SATEs approximately equal to each other, we cannot be sure that the convenience sample estimate generalizes in the sense of being representative of the population.

1.1 Measuring Effect Heterogeneity

Knowing whether effects are homogeneous or heterogeneous and, if the latter, across what characteristics (if any) is not trivial. Due to the fundamental problem of causal inference, $Var(\tau_i)$ is unknowable because individual-level treatment effects cannot be observed. At a certain level, this implies that the three scenarios cannot be arbitrated between. But even lacking the ability to observe any τ_i, a relatively standard battery of mathematical procedures can allow us to assess whether there appears to be variation in *average* treatment effects across observable subgroups of experimental participants.

Drawing credible expectations about generalizability or replicability requires insight into sources of effect heterogeneity in the population and information about the distribution of those characteristics in a convenience sample. We therefore focus on apparent sources of treatment effect heterogeneity and consider the degree to which observable subgroups respond to treatment similarly, on average, regardless of recruitment method. We will measure the correlation in estimated CATEs both across and within studies conducted on nationally-representative and convenience samples.

2 Design

We aim to distinguish between Scenarios 1, 2, and 3 through two reanalyses of 26 original-replication pairs collected by Mullinix et al. (2015) and Coppock (forthcoming). This set of studies is useful for our purposes because they constitute a unique sample of direct study replications performed on convenience samples (namely Amazon Mechanical Turk) and nationally representative samples using identical experimental protocols. The analyses reported by Mullinix et al. (2015) and Coppock (forthcoming) focused narrowly on replication as assessed by the correspondence between SATEs in each study pair. Both papers found a high degree of correspondence. Our goal in the present study is to assess the degree of correspondence across original and replication studies within subgroups defined by subjects' pre-treatment background characteristics. Instead of comparing SATEs, we will compare the CATEs among 16 distinct subgroups. We will estimate the CATE in each group by difference-in-means.

Because of the varied experimental protocols for each of the 52 separate experiments reanalyzed here, the largest challenge we face is measuring subject characteristics in the same scale. While some studies measure a rich set of of demographic, psychological, and political attributes, others only measure a few. We have identified six attributes that are measured in nearly all studies: age, education, gender, ideology, partisanship, and race. These attributes are not always measured in the same way, so we have coarsened each to a maximum of three categories in order to maintain rough comparability across studies. The resulting covariate scales are presented in Table 2. We acknowledge that our covariate measures are *rough* and that many subtleties of scientific interest will unfortunately be masked. In particular, we regret the extreme coarsening of race and ethnicity into white/nonwhite, but smaller divisions left us with far too little data in some cases. We would argue that disaggregating our samples by our admittedly crude measures represents a large increase in the subtlety with which these datasets have previously been analyzed but we nevertheless recognize that some comparisons are simply out of reach with existing data.

Table 2: Coarsened Covariate Information

Age	Education	Gender	Ideology	Partisanship	Race
18-39	Less than College	Men	Liberal	Democrat	Nonwhite
40-59	College	Women	Moderate	Independent	White
60+	Graduate School		Conservative	Republican	

Once we have calculated all of the CATEs, we need to summarize them. Our first statistic of interest is the across-study correlation of CATEs, subgroup by subgroup. This statistic refers to how much knowing the CATE among, say, women on Mechanical Turk helps to predict the CATE among women in the national population. Our second statistic is the within-study correlation of CATEs: does knowing which subgroups had the highest CATEs in the Mechanical Turk version of a study help to predict which subgroups have high CATEs in the population version of the study?

2.1 The Studies

A complete description of each experiment and replication procedures are available in the original papers and their supplementary materials (Mullinix et al. 2015; Coppock forthcoming). The full list of studies, with the sample sizes used in the analyses reported here, is presented in Table 4. These studies are broadly representative of the sorts of framing, priming, and information survey experiments used by political scientists, psychologists, and sociologists. They do not include experiments used primarily for measurement, such as conjoint or list experiments. By and large, these experiments estimate the effects of stimuli on social and political attitudes and opinions.

3 Results

We present our across-study and within-study results separately.

3.1 Across-Study Results

Figure 1 displays scatterplots of the estimated CATEs subgroup by subgroup. The relationship between the conditional average treatments in the original and Mechanical Turk versions of the studies is unequivocally positive for all demographic subgroups. Whereas previous analyses of these datasets showed strong correspondence of *average* treatment effects, this analysis shows that the same pattern holds at every level of age, gender, race, education, ideology, and partisanship that we measure.

The figure also indicates whether the CATEs are statistically significantly different from each other. Out of 378 opportunities, the difference-in-CATEs is significant 58 times, or 15% of the time. In zero of 378 opportunities do the CATEs have different signs while both being statistically significant. Of the 156 CATEs that were significant in the original, 118 are significant in the MTurk version. Of the 222 CATEs that were insignificant in the original, 144 were insignificant in the MTurk version. The overall "significance match" rate is therefore 69%. We must be careful, however, not to overinterpret conclusions based on statistical significance, as they confounded by the power of the studies: If the studies were infinitely powered, all estimates would be significant and the match rate would be 100%. If all studies were infinitely underpowered, all estimates would be insignificant and the match rate would again be 100%. We prefer the correlation statistic since it operates on the estimates rather than on arbitrary significance levels.

The estimated correlations across CATEs are shown in Table 3. The correlations are all strongly positive, ranging from 0.53 to 0.90. The lowest correlation is observed for the independent category, perhaps owing to unmodeled heterogeneities in that group. The strength of the correlations is all the more impressive considering the large amount of measurement error: each CATE estimate is only an estimate, accompanied by sometimes large amounts of uncertainty, as indicated by the wide confidence intervals. While we could attempt to estimate the true correlation after accounting for measurement error, this exercise would only serve to increase the already high correlations. In this set of studies, the CATEs within demographic subgroups are strongly correlated across studies.

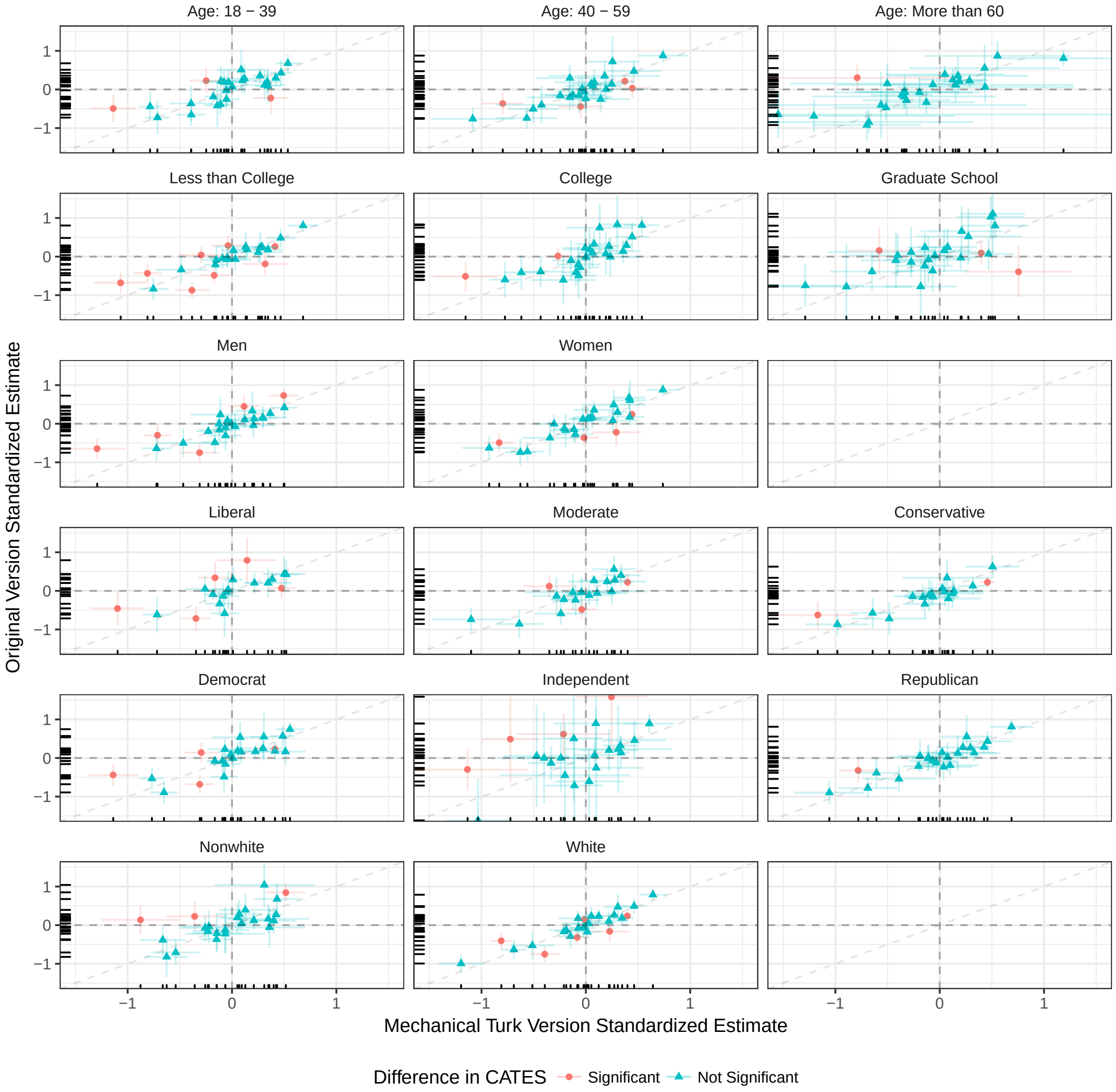
Table 3: Across-Study Correspondence of Conditional Average Treatment Effects

Covariate Class	Correlation	Slope	N Comparisons
Age: 18 - 39	0.74	0.69	26
Age: 40 - 59	0.82	0.81	26
Age: More than 60	0.80	0.65	26
Less than College	0.84	0.80	25
College	0.76	0.83	25
Graduate School	0.63	0.66	25
Men	0.81	0.74	25
Women	0.86	0.89	25
Liberal	0.68	0.67	19
Moderate	0.80	0.83	19
Conservative	0.89	0.75	19
Democrat	0.78	0.77	23
Independent	0.53	0.77	22
Republican	0.90	0.85	23
Nonwhite	0.66	0.74	24
White	0.90	0.88	26

3.2 Within-Study Results

We now have two basic findings to explain: average treatment effects are approximately the same in probability and nonprobability samples and so are conditional average

Figure 1: Across-Study Correspondence of Conditional Average Treatment Effects



treatment effects. Which of our three explanations (no heterogeneity, heterogeneity orthogonal to selection, or good luck) can account for both findings?

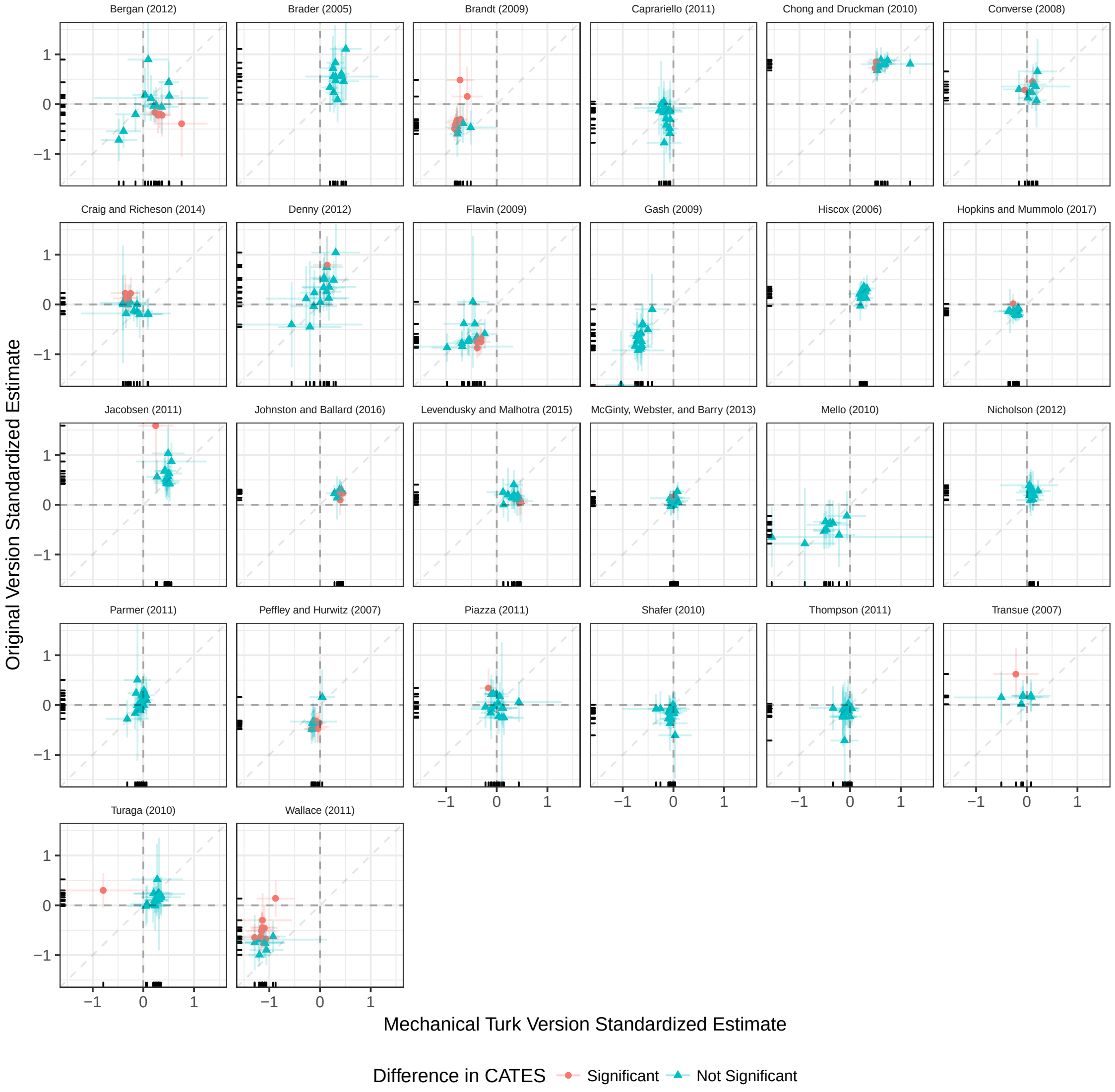
To arbitrate between these explanations, we turn to within-study comparisons. Within a given study, are the CATEs that were estimated to be high in the original study also high in the MTurk version? Figure 2 shows that the answer tends to be no. The CATEs in the original study are mostly uncorrelated with the CATEs in the MTurk versions. Table 4 confirms what the visual analysis suggests. We see correlations that are smaller than the across-study correlations and correlations of both signs.

An inspection of the CATEs themselves reveals why. Most of the CATEs are tightly clustered around the overall average treatment effect in each study version. Another way of putting it is, the treatment effects within each study version appear to be *mostly* homogeneous. We conclude from this preliminary analysis that the *main* reason why we observe strong correspondence in average treatment effects is low treatment effect heterogeneity.

Table 4: Within-Study Correspondence of Conditional Average Treatment Effects

Study	Original N	MTurk N	Correlation	Slope	N Comparisons
Bergan (2012)	1206	1913	0.33	0.40	16
Brader (2005)	280	1709	0.40	1.12	12
Brandt (2009)	1225	3131	0.30	0.95	13
Caprariello (2011)	825	2729	-0.36	-1.20	16
Chong and Druckman (2010)	958	1400	0.27	0.09	13
Converse (2008)	1019	1913	0.29	0.43	10
Craig and Richeson (2014)	608	847	-0.63	-0.59	16
Denny (2012)	1733	1913	0.79	1.41	16
Flavin (2009)	2015	2729	0.18	0.22	16
Gash (2009)	1022	3131	0.87	2.20	16
Hiscox (2006)	1610	2972	0.54	1.16	16
Hopkins and Mummolo (2017)	3266	2972	-0.21	-0.22	16
Jacobsen (2011)	1111	3171	-0.49	-1.81	16
Johnston and Ballard (2016)	2045	2985	0.24	0.33	16
Levendusky and Malhotra (2015)	1053	1987	-0.19	-0.17	16
McGinty, Webster, and Barry (2013)	2935	2985	0.36	0.53	16
Mello (2010)	2112	3131	0.65	0.26	10
Nicholson (2012)	781	1099	-0.13	-0.24	12
Parmer (2011)	521	3277	0.42	0.87	16
Peffley and Hurwitz (2007)	905	1285	0.62	1.62	13
Piazza (2011)	1135	3171	-0.24	-0.29	16
Shafer (2010)	2592	2729	-0.29	-0.44	16
Thompson (2011)	591	3277	0.13	0.25	16
Transue (2007)	345	367	-0.16	-0.15	7
Turaga (2010)	774	3277	-0.11	-0.06	16
Wallace (2011)	2929	2729	0.48	1.15	16

Figure 2: Within-Study Correspondence of Conditional Average Treatment Effects



4 Discussion

We have suggested that different samples will yield similar SATEs when either: (1) there is no effect heterogeneity, (2) any effect heterogeneity is orthogonal to sample selection, or (3) by some good luck. Arbitrating between these possibilities requires knowledge of the $Var(\tau_i)$. Claims about the between-sample comparability of experimental results ("generalizability") depend upon the degree of individual-level effect heterogeneity in each sample.

Drawing on a fine-grained analysis of 26 pairs of survey experiments conducted on representative and non-representative samples and various methods of assessing the pattern of effect heterogeneity in each study, we have shown that effect heterogeneity is typically limited. The convenience samples we analyze therefore provide useful estimates not only of the PATE but also of subgroup CATEs.

Our results indicate that even descriptively unrepresentative samples constructed with no design-based justification for generalizability still tend to produce estimates not just of the SATE but also of subgroup CATEs that generalize quite well. Yet important caveats are in order. First, we have not considered all possible survey experiments, let alone all possible experiments in other modes or settings. Our pairs of studies were limited to those conducted in an online mode on samples of United States residents. However, this set of studies is also quite comprehensive, drawing from multiple social science disciplines, utilizing a variety of experimental stimuli and outcome question formats. The studies are also drawn not just from published research (which we might expect to be subject publication biases) but from a large sample of all experiments conducted by Time-Sharing Experiments for the Social Sciences.

Second, because we can never perfectly know the variation in treatment effects, our analysis of heterogeneity is limited by both the set of covariates that are available for direct comparison between samples and any measurement error in those covariates. We made several decisions about coarsening of covariates (for example, comparing whites to members of all other racial and ethnic groups) that reflected the need for minimum sample sizes at the expense of measurement precision. Accordingly, our results may mask possible moderators of treatment effects (though we would note that the low levels of heterogeneity according to the covariates we were able to measure leads us to be skeptical of predictions of high levels of unmodeled effect heterogeneity). Our reliance on existing studies as the basis for our empirics is important because it means that we are evaluating the degree and pattern of effect heterogeneity using the types of samples and set of covariates typically used in survey-experimental research. Additional and more precisely measured covariates might have allowed for detection of more complex patterns of effect heterogeneity, but survey-experimental research rarely offers such detail.

Finally, the subgroup samples we analyzed were relatively small. While we may be well-powered to estimate an SATE, these studies were not necessarily designed to detect any particular source of effect heterogeneity. Larger sample sizes, oversampling of rare populations, and more precise measurement of covariates would have allowed the detection of smaller variations in effect sizes across groups, but researchers rarely have access to larger samples than those used here.

Perhaps the most controversial conclusion that could be drawn from the present research is that we should be much more suspect of extant claims of effect moderation. A common post-hoc data analysis procedure is to examine whether subgroups (often one at a time) differ in their apparent response to treatment. We find only limited evidence that such moderation ever occurs and when it does, the differences in effect sizes across groups are small. The response to this evidence should not be that any convenience sample can be used to study any treatment without concern about generalizability (see, for example, Deaton and Cartwright 2016) but rather that debates about generalizability and replication must focus on the underlying causes of replication and nonreplication, among these most importantly, the variation in treatment effects across experimental units.

References

Aronow, Peter M. and Cyrus Samii. 2015. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* .

Baron, Reuben M. and David A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51(6):1173–1182.

Coppock, Alexander. forthcoming. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* .

Cronbach, Lee J. 1986. "Social Inquiry by and for Earthlings." *Metatheory in Social Science: Pluralisms and Subjectivities* pp. 83–107.

Deaton, Angus and Nancy Cartwright. 2016. Understanding and isunderstanding randomized controlled trials. Technical report National Bureau of Economic Research.

Druckman, James N., Donald P. Green, James H. Kuklinski and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4):627–635.

Druckman, James N., Donald P. Green, James H. Kuklinski and Arthur Lupia. 2011. *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.

Gerring, John. 2012. *Social Science Methodology: A Unified Framework*. Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2):481–502.

Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1(March):59–80.

McDermott, Rose. 2011. "New Directions for Experimental Work in International Relations." *International Studies Quarterly* 55(2):503–520.

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.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton University Press.

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.

Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton, Mifflin and Company.

Weinberg, Jill D., Jeremy Freese and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample." *Sociological Science* 1:292–310.