The Generalizability of Heterogeneous Treatment Effect Estimates Across Samples

Alexander Coppock[1] Thomas J. Leeper[2] Kevin J. Mullinix[3]

[1]Department of Political Science, Yale University
[2]Department of Government, London School of Economics and Political Science
[3]Department of Government and Justice Studies, Appalachian State University

Prepared for presentation at MPSA 2018

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 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 27 original-replication study pairs (encompassing 101,745 individual survey responses) to assess the extent to which subgroup effect estimates generalize. 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.

Introduction

Randomized experiments are increasingly employed across the social sciences to study beliefs, opinions, and behavioral intentions (*1, 2*). Experiments are nevertheless met with skepticism about the degree to which results *generalize* (*3*). 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., (*4–6*), though see (*7*) 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 (*8–11*).

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: (A) effect homogeneity across participants such that sample characteristics are irrelevant, (B) effect heterogeneity that is approximately orthogonal to selection, or (C) 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.

We aim to distinguish between Scenarios A, B, and C through two reanalyses of 27 original-

replication pairs collected by (*8*) and (*9*). 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 (*8*) and (*9*) 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 conditional average treatment effect (CATE) among 16 distinct subgroups. We will estimate the CATE in each group by difference-in-means. Our main statistics of interest will be the within-study and across-study correlations across CATEs.

Because of the varied experimental protocols for each of the 54 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 1. 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 imprecise 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 1: 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	

A complete description of each experiment and replication procedures are available in the original papers and their supplementary materials (*8, 9*). The full list of studies, with the sample sizes used in the analyses reported here, is presented in Table 3. 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.

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 394 opportunities, the difference-in-CATEs is significant 59 times, or 15% of the time. In zero of 394 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 238 CATEs that were insignificant in the original, 159 were insignificant in the MTurk version. The overall "significance match" rate is therefore 70%. 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 2. The correlations are all strongly positive, ranging from 0.40 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.

We now have two basic findings to explain: average treatment effects are approximately the same in probability and nonprobability samples and so too are conditional average 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 3 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 clus-

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

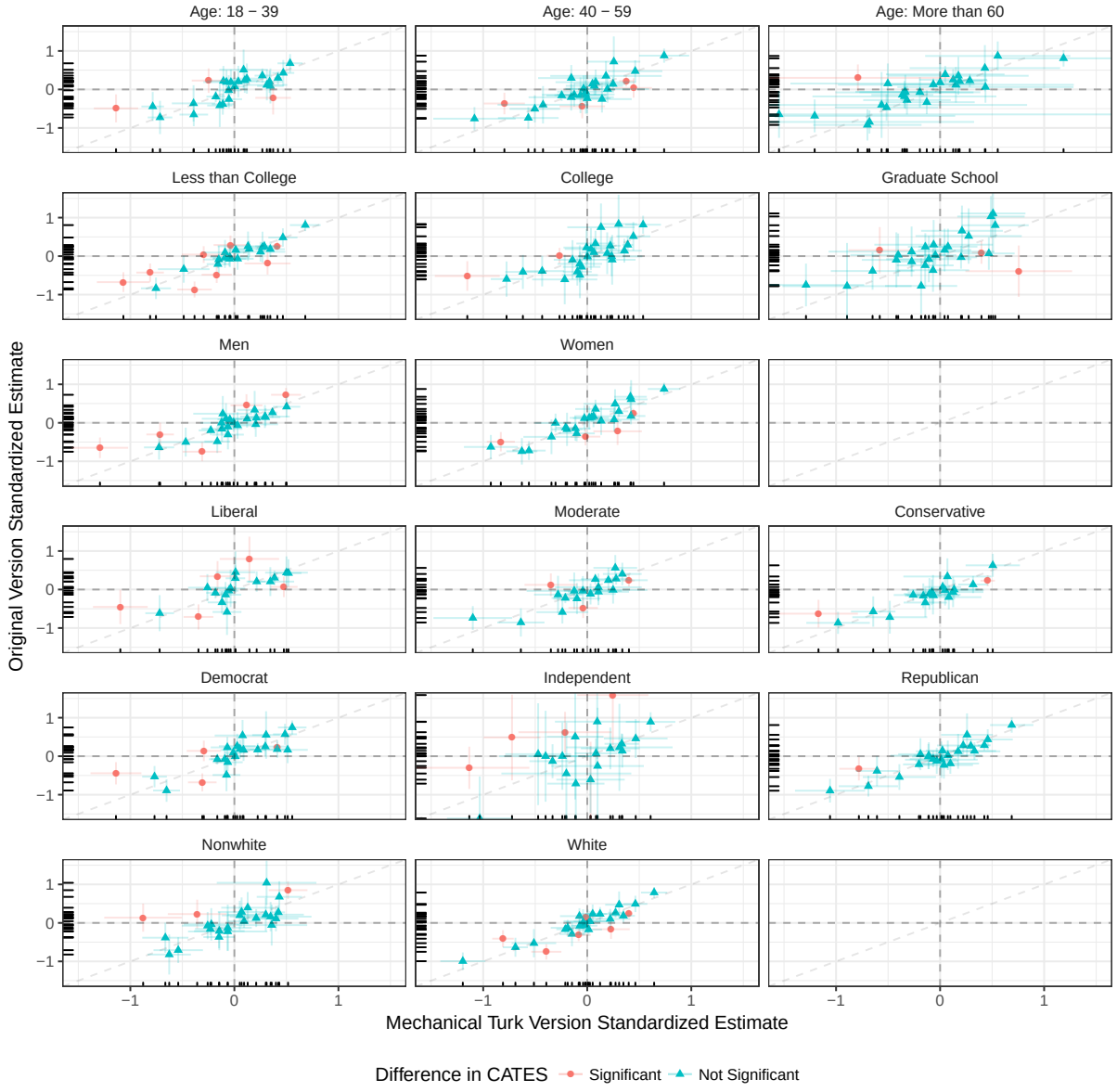

Table 2: Across-Study Correspondence of Conditional Average Treatment Effects

Covariate Class	Correlation	Slope	N Comparisons
Age: 18 - 39	0.74	0.69	27
Age: 40 - 59	0.82	0.81	27
Age: More than 60	0.80	0.65	27
Less than College	0.84	0.80	26
College	0.74	0.81	26
Graduate School	0.63	0.66	26
Men	0.81	0.74	26
Women	0.86	0.88	26
Liberal	0.67	0.68	20
Moderate	0.80	0.83	20
Conservative	0.89	0.75	20
Democrat	0.78	0.77	24
Independent	0.40	0.70	23
Republican	0.90	0.85	24
Nonwhite	0.66	0.73	25
White	0.90	0.88	27

tered 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.

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. Drawing on a fine-grained analysis of 27 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

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

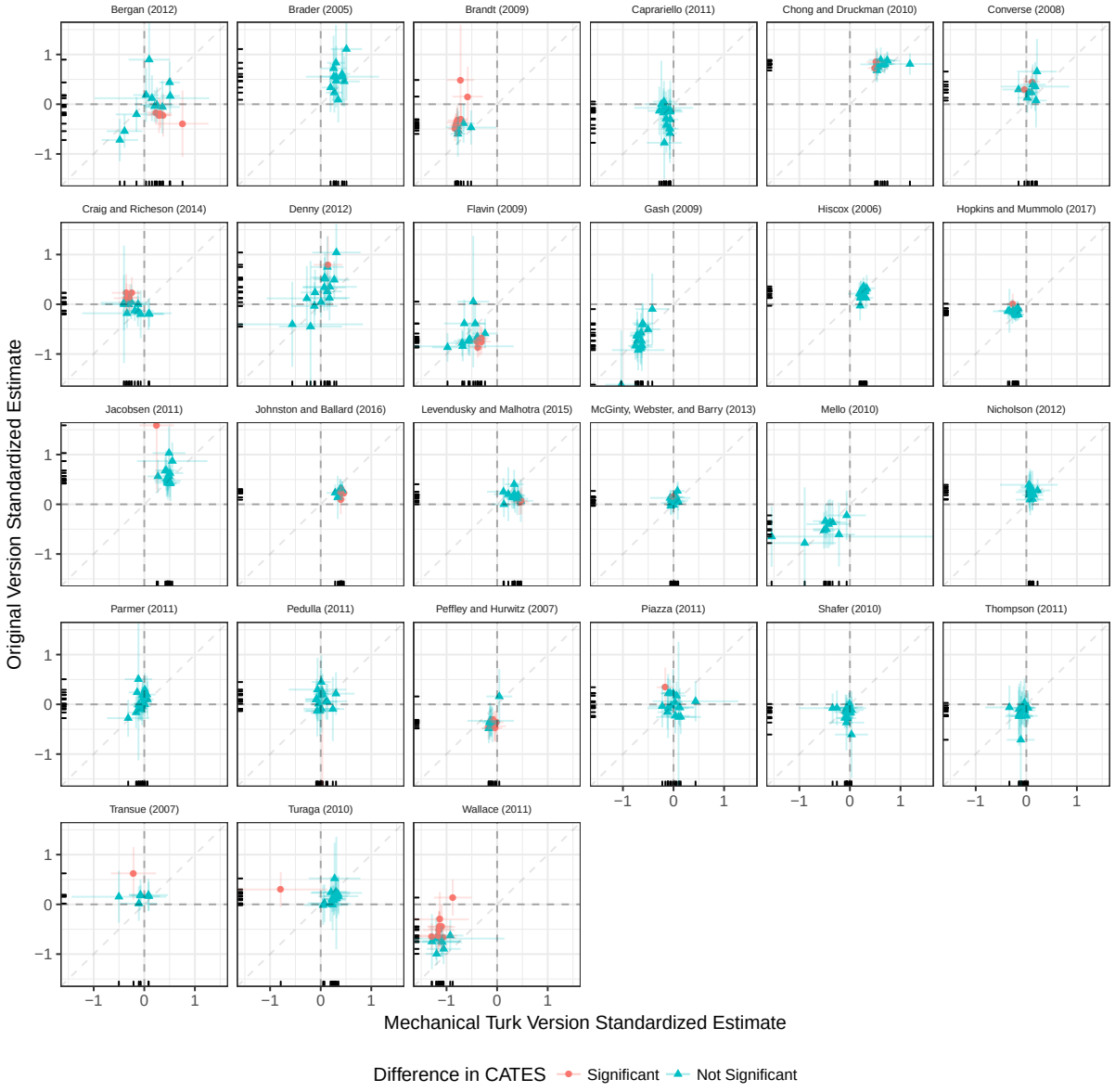

Table 3: Within-Study Correspondence of Conditional Average Treatment Effects

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
Pedulla (2011)	1407	1913	0.00	0.01	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

tend to produce estimates not just of the SATE but also of subgroup CATEs that generalize quite well.

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 also from the set of 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 a minimum level 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 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, (*12*)) 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 and Notes

1. J. N. Druckman, D. P. Green, J. H. Kuklinski, A. Lupia, *American Political Science Review* **100**, 627 (2006).

2. J. N. Druckman, D. P. Green, J. H. Kuklinski, A. Lupia, *Cambridge Handbook of Experimental Political Science* (Cambridge University Press, New York, 2011).

3. J. Gerring, *Social Science Methodology: A Unified Framework* (Cambridge University Press, 2012).

4. D. O. Sears, *Journal of Personality and Social Psychology* **51**, 515 (1986).

5. L. J. Cronbach, *Metatheory in Social Science: Pluralisms and Subjectivities* pp. 83–107 (1986).

6. R. McDermott, *International Studies Quarterly* **55**, 503 (2011).

7. P. M. Aronow, C. Samii, *American Journal of Political Science* (2015).

8. K. J. Mullinix, T. J. Leeper, J. N. Druckman, J. Freese, *Journal of Experimental Political Science* **2**, 109 (2015).

9. A. Coppock, *Political Science Research and Methods* (forthcoming).

10. Y. Krupnikov, A. S. Levine, *Journal of Experimental Political Science* **1**, 59 (2014).

11. J. D. Weinberg, J. Freese, D. McElhattan, *Sociological Science* **1**, 292 (2014).

12. A. Deaton, N. Cartwright, Understanding and isunderstanding randomized controlled trials, *Tech. rep.*, National Bureau of Economic Research (2016).