INSTITUTE for REPLICATION

No. 10
I4R DISCUSSION PAPER SERIES

Replication of "Re-Assessing Elite-Public Gaps in Political Behavior" by Joshua Kertzer

Eric Guntermann

Gabriel S. Lenz

November 2022

I4R DISCUSSION PAPER SERIES

I4R DP No. 10

Replication of "Re-Assessing Elite-Public Gaps in Political Behavior" by Joshua Kertzer

Eric Guntermann[1], Gabriel S. Lenz[1]

[1]Travers Department of Political Science, University of California, Berkeley CA/USA

NOVEMBER 2022

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Replication of "Re-Assessing Elite-Public Gaps in Political Behavior" by Joshua Kertzer

Eric Guntermann* Gabriel S. Lenz†

2022-10-17

Abstract

Kertzer (2022) conducts a meta-analysis of parallel experiments on samples of political elites and ordinary citizens. He examines whether the average treatment effect for elites is significantly different from the average treatment effect for citizens, finding that only 19 of 162 (11.7%) difference-in-difference estimates are statistically significant after adjusting for the false discovery rate. He also finds that elites and masses hold similar foreign policy attitudes after controlling for their demographic characteristics. In this replication report, we begin by running robustness and heterogeneity tests for the first claim. We find that the results survive many robustness tests. We also find, however, that only a small number of the these treatments significantly affected masses (N=28) or elites (N=30). This low rate suggests the possibility that almost all of these experiments failed to successfully manipulate either masses or elites. If so, we may not be able to conclude that masses and elites respond similarly to experiments with confidence until political scientists produce more experiments with actual treatment effects or with successful manipulation checks in cases of null effects. In the second part of this replication report, we conceptually replicate the second Kertzer analysis, finding a strong correlation between elite and mass political decisions and attitudes, thus confirming Kertzer's analysis.

1 Introduction

The article presents a replication of Kertzer's (2022) meta-analysis of parallel decision-making experiments conducted on elites and masses in different countries. It also presents a conceptual replication of Kertzer's (2022) analysis of the political attitudes of elites and masses on foreign policy. For the meta-analysis, Kertzer collected results from studies that included paired experiments on political elites and mass samples that contained:

1. an experiment where the treatments are randomly assigned by an experimenter, and
2. the same experiment is fielded both on a sample of political elites (current or former politicians, civil servants, military officers, etc.) and a mass public or convenience sample (Supplementary Appendix, page 2).

His main analysis of the experimental data relies on assessing the percentage of studies that have significantly different effects, both before and after controlling for the false discovery rate (FDR) using the Benjamini-Hochberg procedure.

*ericguntermann@berkeley.edu. Travers Department of Political Science, University of California, Berkeley. 210 Social Science Building, Berkeley, CA 94720-1950

†glenz@berkeley.edu. Travers Department of Political Science, University of California, Berkeley. 210 Social Science Building, Berkeley, CA 94720-1950

The author finds that decision-making is only significantly different in a small minority of cases: 39 of 162 cases (24.0%) or 19 of 162 treatments (11.7%) after controlling for the false discovery rate. He concludes that "the treatment effects recovered in the elite samples included in this analysis do not significantly differ in magnitude from those recovered from mass samples 88% of the time" (Claim 1, page 7).

For the study of political attitudes, the author uses parallel surveys of "foreign policy leaders" and mass public surveys of "nationally representative samples of the American public, originally fielded by telephone by Gallup, and eventually online by YouGov" (8). He uses data on "1504 individually-matched foreign-policy questions, from 26220 respondents (5741 foreign policy elites, and 20479 members of the public across twelve waves from 1975 to 2018" (8).

Kertzer calculates raw correlations between elite and mass attitudes on each foreign policy issue, both before and after adjusting attitudes for demographic characteristics. He reports that "the correlation between elite and public attitudes markedly improves after adjustment, ranging from r=0.76 in 1998 to r=0.90 in 1986", compared to the range of r=0.38 in 2016 to r=0.80 in 1986 in unadjusted data (Claim 2, pages 9 and 12).

We conduct a robustness replication of Claim 1 and a brief conceptual replication of Claim 2.[1] We pre-registered our analysis and code. For our robustness replication of Claim 1, we evaluated the decision to include studies and study treatments in the meta-analysis given the criteria Kertzer lays out and whether the results are robust to other reasonable decisions about inclusion. We also considered whether pooling results from similar treatment effects could improve power. Finally, we investigated whether the findings held across a range of attributes, including whether the experiments had manipulation checks, conducted the studies in person, etc. We were particularly interested in how much the twin problems of low power and a failure to successfully manipulate respondents lies behind the absence of treatment effect differences between masses and elites. Experiments are chronically underpowered in the social sciences and in political science in particular. One recent meta analysis concluded that the average power in political science studies is just 11% percent (Arel-Bundock et al. 2022), a notch below the 18% found in economics (Ioannidis et al. 2017). Even when studies are well powered, successful manipulation may be rare. Respondents often aren't paying much attention and, when they are, may be unmoved by the experimental treatment.

We focus the replication on Claim 1 because we think it's more important and because of the effort the replication entailed. We nevertheless also conduct a conceptual replication of Claim 2, leveraging data on the responses of elite and mass samples in the control conditions in the experiments Kertzer uses to assess Claim 1. Although Kertzer's analysis for Claim 1 focused on mass-elite responses to treatment, the data from the control group samples in these studies involve political decisions and political attitudes (uninfluenced by experimental interventions). So, we investigate whether elites and masses make similar decisions and express similar attitudes in the control groups. Importantly, the studies cover topics beyond international relations, whereas Kertzer's original analysis of Claim 2 examine data only on political attitudes in that domain. In our conceptual replication, we find correlations using attitudes unadjusted for demographics that are at the high end of the range Kertzer finds in his analysis of attitudes adjusted for demographics. Whereas Kertzer found correlations ranging from 0.76 to 0.90, we find a correlation of r=0.87 when including all treatment effects and r=0.86 when excluding placebo tests, non-political outcomes, and uncontroversial treatments.

[1] We obtained the replication data and code from https://doi.org/10.7910/DVN/LHOTOK.

2 Reproducibility

The study was already successfully reproduced by Institute for Replication collaborators.

3 Replication and Pre-registration

Replication code and data will be made available with the posting of this replication report. We pre-registered the replication plan and code here: https://doi.org/10.17605/OSF.IO/CEDTJ.

3.1 Robustness Replication of Meta-Analysis of Mass-Elite Experiments (Claim 1)

We begin with the robustness replication of Claim 1, focusing first on evaluating whether all 162 treatment effects should be included in the meta-analysis. The authors read each study and evaluated whether each treatment included was

- a real treatment rather than a placebo,
- political,
- not a manipulation check.

We determined that a small number, 23 of 162, did not fall into these categories, including all treatments from three studies (out of the 26 total studies, see Appendix A for an explanation of each case).

We also judged that of the 139 remaining treatment effects examined, the treatment and outcomes were sufficiently similar that we could pool in 49 (see the 'poolid' variable in the replication data). Here, we only pool treatments from the same study. Given the small elite sample sizes in many of the studies, pooling when possible seems important. We pool by calculating the precision-weighted averages of the poolable effects (Raudenbush and Bryk 2002).

In making these judgments, we recognize that reasonable people may disagree and encourage interested scholars to evaluate the studies themselves. We will post our coding and replication data with this report.

Does excluding these cases and pooling materially change the results of the meta-analysis? To answer this question, we replicate Kertzer's (2022) Figure 1 in our Figure 1(a). Each estimate shows the absolute value of the difference between the mass and the elite Average Treatment Effects (ATEs) with 95 percent confidence intervals. As in Kertzer (2022), black dots indicate non-significant effects, grey dots indicate effects that are significant prior to controlling for the false discovery rate, and white dots are significant after controlling for the false discovery rate (using the Benjamini-Hochberg procedure as in Kertzer 2022).

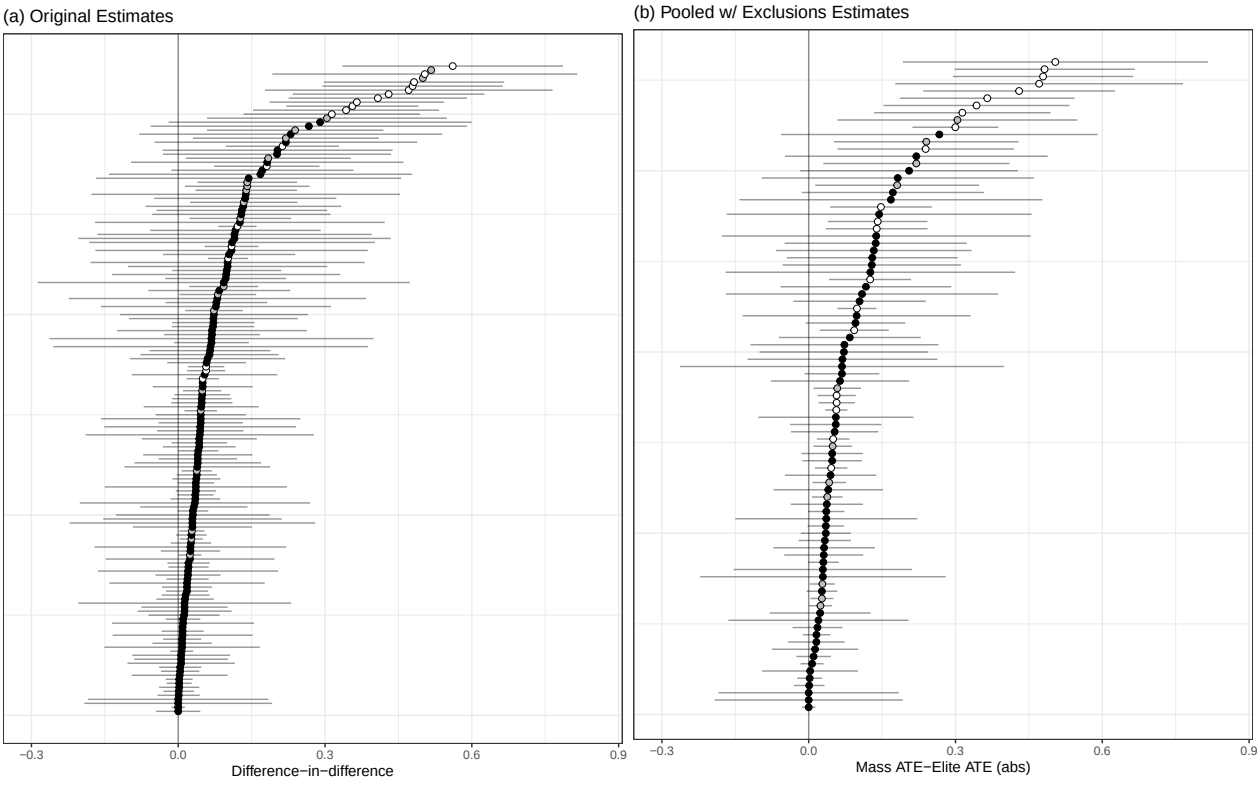

Figure 1: Replication of Kertzer Figure 1 in panel (a) and our replication in panel (b) excluding placebo, non-political, and uncontroversial treatment effects, as well as pooling study estimates where appropriate (because they had similar treatments and multiple samples or similar DVs). Black dots indicate non-significant effects, grey dots indicate effects that are significant prior to controlling for the false discovery rate, and white dots are significant after controlling for the false discovery rate. 95 percent confidence intervals.

Figure 1(b) reveals a broadly similar pattern to the original study with a small increase in the percent significant at conventional levels. The percentage of significant effects increases from 24.0 percent in the original analysis to 35.6 percent. After controlling for the false discovery rate, the percentage of significant effects increases from 11.7 percent to 23.3 percent.

To summarize these results differently, Figure 2 shows the precision-weighted averages of these estimates. The top estimate shows it for all studies. Below that, the estimate is for all studies after excluding the 23 treatment effects we just discussed, and below that the estimate is after pooling. It shows estimates along with 95% confidence intervals. It also shows the numbers and percentages of effects that are significant before and after controlling for the false discovery rate. The top few estimates show that the precision-weighted average remains largely unchanged with the exclusions, though the percent significant does increase (only the latter is affected by pooling).[2]

[2]We found three minor typos in the pre-registered R Markdown code that produces Figure 2. They are clearly indicated in the source code for this document.

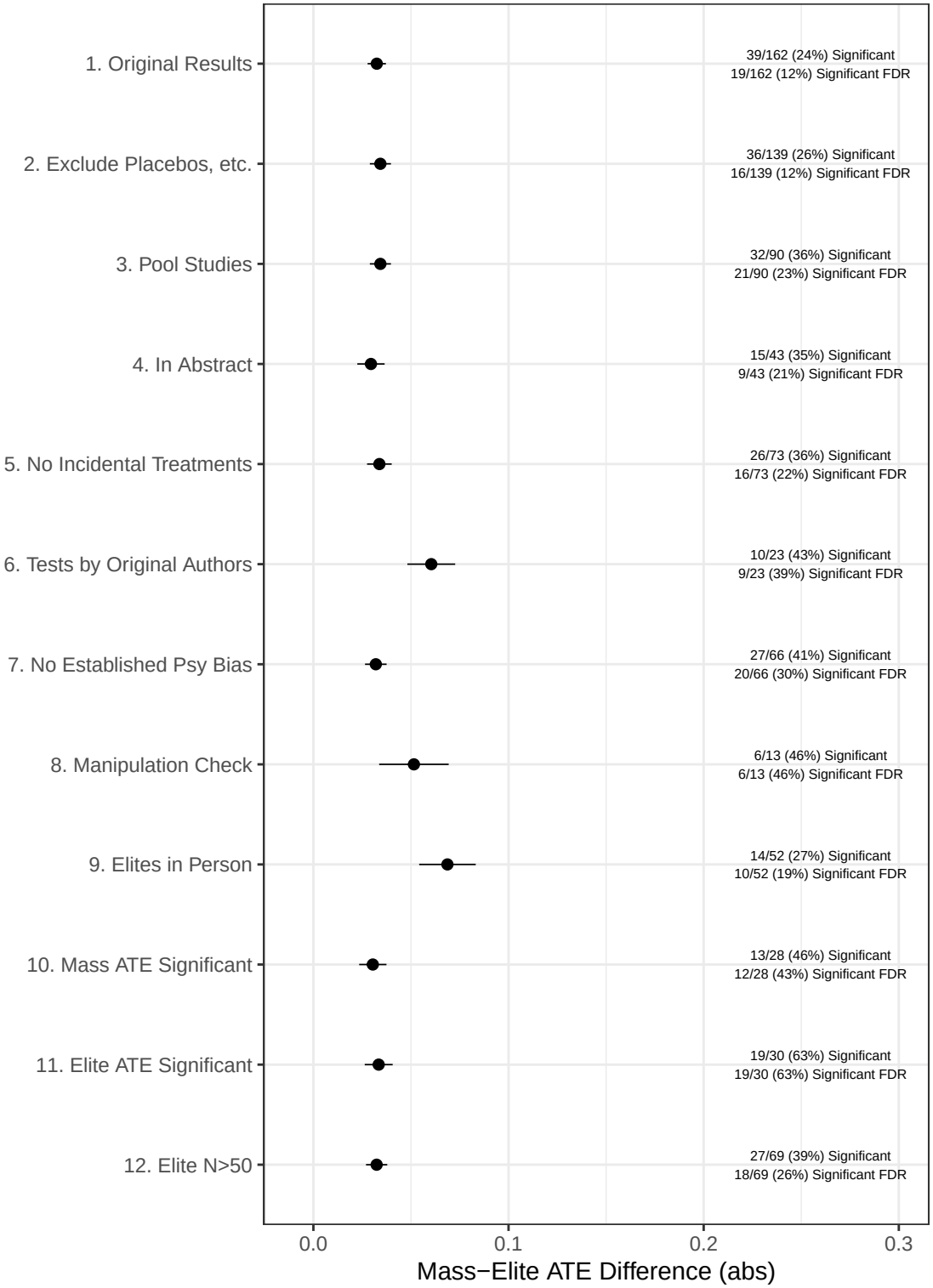

Figure 2: Meta-analytic estimates of the differences between mass and elite reactions to treatments. The figure shows the average mass-elite difference in the ATEs in the original study and then for various categories of studies we coded. The figure uses precision-weighted averages. All estimates after the third exclude placebo treatments, non-political decisions, and uncontroversial treatments, as well as pool studies. 95 percent confidence intervals.

We were surprised that the percent significant could change with the exclusions without much change in the precision-weighted average. After inspecting the data further, the explanation is in part the presence of several highly powered studies with treatment effect estimates near zero. Figure 4 shows scatterplots with the estimated absolute differences in treatment effects on the horizontal axis and the precision on the vertical axis (1/standard error^2). Given these studies' high power, their results disproportionally influenced the precision-weighted averages. High powered studies are of course more informative and published studies are chronically underpowered. Nevertheless, since each study is capturing an effect of a different treatment, we might not want to overweight a handful of studies to quite this degree. We therefore also present a version of Figure 2 that shows the mean of estimates rather than the precision-weighted means. Figure 3 shows the mean estimate by each category. We also present a version of Figure 2 without the extremely precise estimate from Friedman et al. (2017) in Appendix B. Note that Figures 3, 4 and Figure A1 in Appendix B are the only parts of our analysis that were not pre-registered. The numbers and percentages of treatment effects that are significantly different are not affected, but the mean estimates are larger than the precision-weighted mean estimates from Figure 2. We exclude confidence intervals from this figure since calculating the average confidence interval doesn't have a clear rationale.

Overall, we conclude that Kertzer's Claim 1 survives an evaluation of whether each study should be included and survives the pooling of similar treatments.

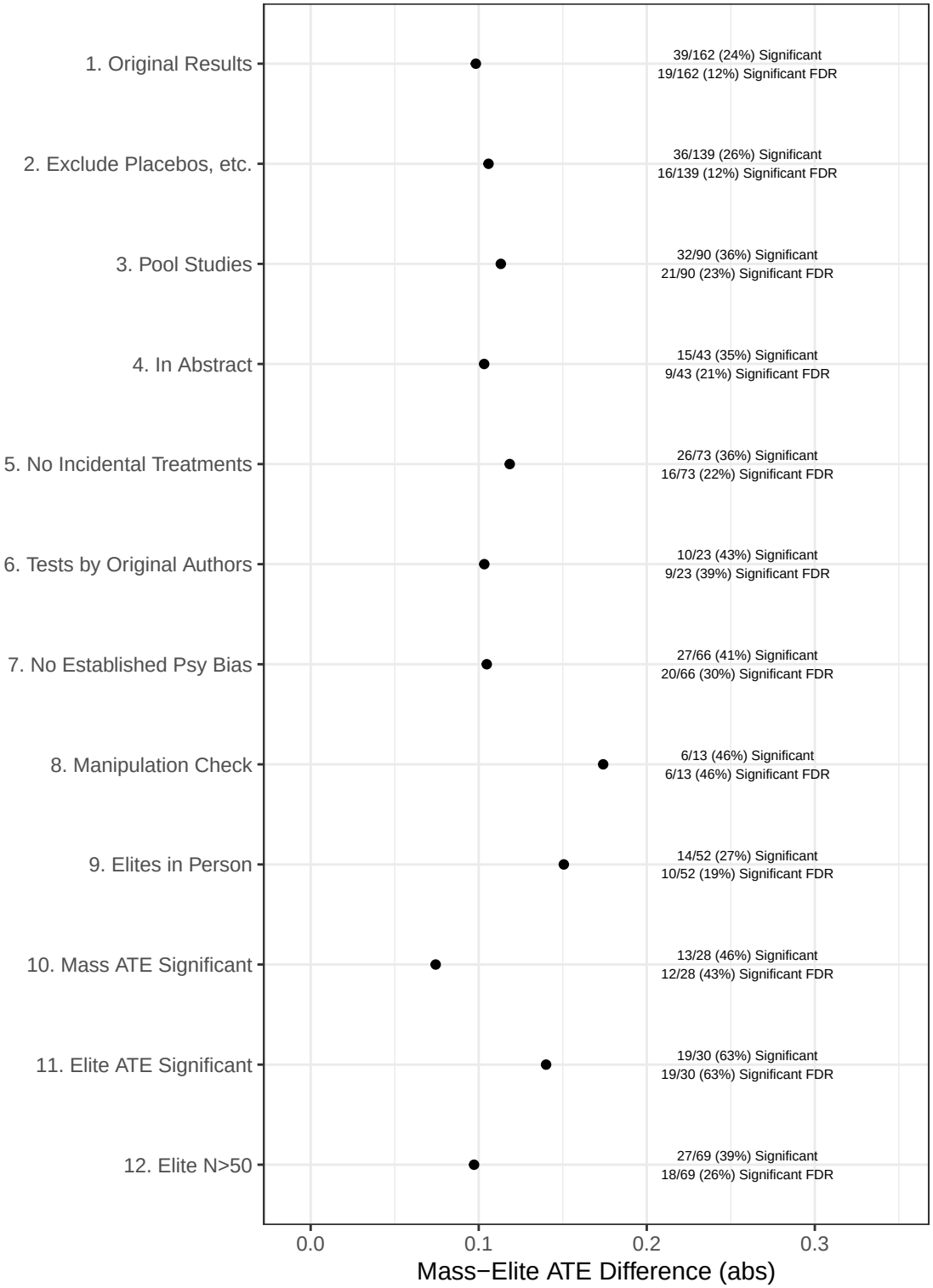

Figure 3: Meta-analytic estimates of the differences between mass and elite reactions to treatments. The figure shows the average estimate in the original study and then for various categories of studies we coded. The figure uses precision unweighted averages. All estimates after the third exclude placebo treatments, non-political decisions, and uncontroversial treatments as well as pool studies. Note that this figure was not pre-registered.

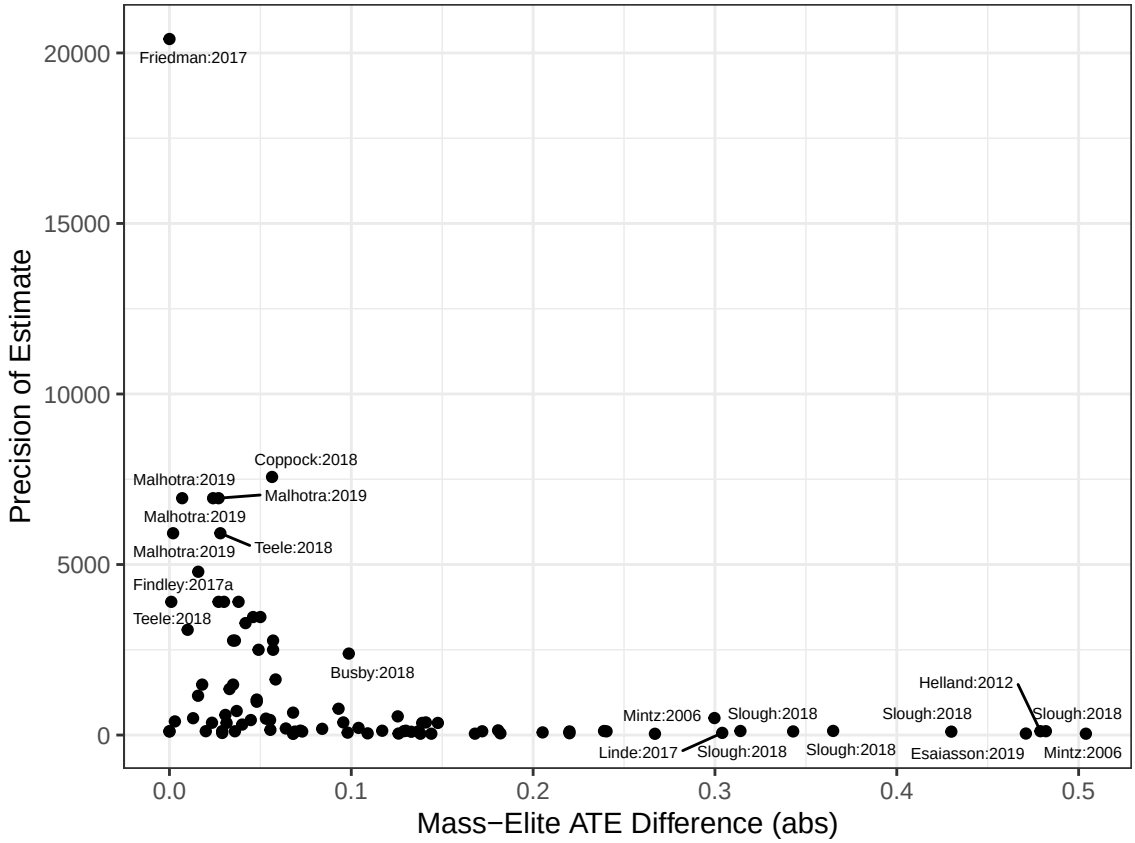

Figure 4: Mass-Elite ATE difference by the precision of the estimates for each treatment effect (1/Standard Error^2)

We now examine how much the overall finding is driven by studies of a particular type. Returning to Figure 2 and Figure 3, the remaining estimates in these figures show the precision-weighted average and the simple average for each of our coding categories. The first set focuses on whether the absence of mass-elite differences in treatment effects holds when the treatments examined were central to the original article, which may be the kinds of treatment effects that most interest researchers. Estimate 4 examines those where the author (or authors) mentioned the treatment effect in the abstract. Somewhat surprisingly, the treatment effects are not noticeably larger or more frequently statistically significant in this category.

Estimate 5 excludes incidental treatments, such as variations of secondary characteristics in conjoint experiments. The average effects increase slightly when we exclude these, but not noticeably.

Estimate 6 investigates treatment effects where authors conducted tests of statistical significance in their articles, which we take as a sign of the importance of the treatment effect to the research endeavor. We find that the percentage of significant results increases, especially when controlling for the FDR.

Estimate 7 studies whether the treatment effect examined well-established psychological biases, biases that are already known to affect masses and elites. In perusing the studies, we noticed that quite a few fell into this category, so we wondered how much these contributed to the overall conclusions in Kertzer. Our expectation was that these studies would find minimal mass-elite

differences, and the figure shows that the percent significant increases when excluding these cases, from 23% to 30% with the FDR correction.

The next set of estimates these figures present tackle the twin problems of underpowered studies and experimental manipulation failures. We were interested in how these two factors drove the overall findings. The manipulation check category, Estimate 8, examines the only 13 of 139 treatment effects where researchers reported a manipulation check in their study. In these 13, almost half of the mass-elite differences are statistically significant and the unweighted average effect is noticeably larger. This is an intriguing result but 13 is just too small of a number to draw conclusions. The bigger takeaway is that researchers should be conducting manipulation checks to see whether they successfully manipulated their conceptual variables.

In Estimate 9, we examine the instances where the elite studies were conducted in person, since such studies can be more certain that respondents were actually elites (as opposed to say a staff member taking the study) and since in-person experiments can increase elite attention to the study and, therefore, lead to successful manipulations of variables of interest. Surprisingly, the estimates don't change much when we examine these cases.

In Estimate 10, we examine cases where the mass treatment effect was statistically significant. Statistically significant mass treatment effects suggest successful manipulation and researchers can only of course draw conclusions about mass-elite responses to experimental manipulations when we know these manipulations are successful. Given pervasive inattention on surveys and numerous other barriers to manipulation, successful manipulation is in no way guaranteed. Of the 139 treatment effects, only 28 had statistically significant effects in the mass samples, a strikingly small number. In these 28, the percent of mass-elite treatment effects that are statistically significant rises to 43% with the multiple testing correction. Although 43% is much higher than the original 12% reported for all studies in Kertzer's original analysis, the number of studies seems too small to draw strong conclusions. Nevertheless, it suggests that future meta-analyses of this sort should be attentive to such cases. The low number of statistically significant effects in the mass samples could be consistent with a broad failure to successfully manipulate the conceptual variable in the mass samples. Overall, an array of findings here suggest that we may be learning less from this meta-analysis than hoped because so few of these treatments appear to have influenced masses and elites. Without successful manipulation, we can't learn much about the difference in mass and elite responses.

Next, Estimate 11 shows the averages for the only 30 of 139 cases where the treatment effects in the elite sample were statistically significant. In these 30, the mass-elite average difference is slightly larger and the percent significant reaches 63% with the multiple test correction. Again, it is striking that so few of the treatment effects had significant effects in the elite samples, only two more than in the mass sample, a result that could be consistent with a broad failure to successfully experimentally manipulate. The high percent significant is also striking, though again the number of treatment effects is so small that it limits the strength of any inference. Finally, Estimate 12 shows the results when elites samples were larger than 50, which occurred in 69 of the 139 cases. Consistent with the nice analysis in Kertzer's appendix, the sample size of elites does not seem to greatly affect the estimates.

As we indicated in our pre-Analysis plan, since we find notable increases in the percentage of significant treatment effects when either mass or elite treatment effects are significant, we combine these conditions. We consider how many differences are significant when neither mass nor elite effects are significant and how many are significant when both are significant. When neither mass nor elite treatment effects are significant, 7 percent of treatment effects are significantly different (7

percent when adjusting for FDR). When both mass and elite effects are significant, 63 percent are significant (63 percent when adjusting for FDR).

We now turn to scatterplots of the mass treatment effects and elite treatment effects, which help us assess the overall findings. Figure 5 presents these for the original sample (a) and after we exclude placebos, etc (b). Kertzer's appendix presented a version of (a). Since so many treatment effects are close to zero, we adopt a modulus transformation to better visualize values close to zero. Figure 6 shows the same data as Figure 5(b), but with labels for each study. We omit confidence intervals from these plots because they add too much visual complexity. These plots reveal that, while many of the points are close to the 45-degree line, some are not. They also make clear just how many are close to the origin, that is, no mass or elite treatment effects.

Some of the studies off the 45-degree line are the most interesting treatment effects in the meta-analysis. For example, Rosenzweig (2021) examines mass and elite perceptions about political violence. He finds that the masses think candidates who support political violence will be repudiated at the polls, whereas elites if anything think it might be helpful. If true, this seems like an enormously consequential mass-elite difference, one that might be contributing to elites fostering of political violence because they misperceive mass attitudes about violence. These and other studies off the 45-degree line suggest that while few of these treatments affected masses and elites differently—they mostly didn't influence either—at times and on important topics the treatments have noticeably different effects on masses and elites.

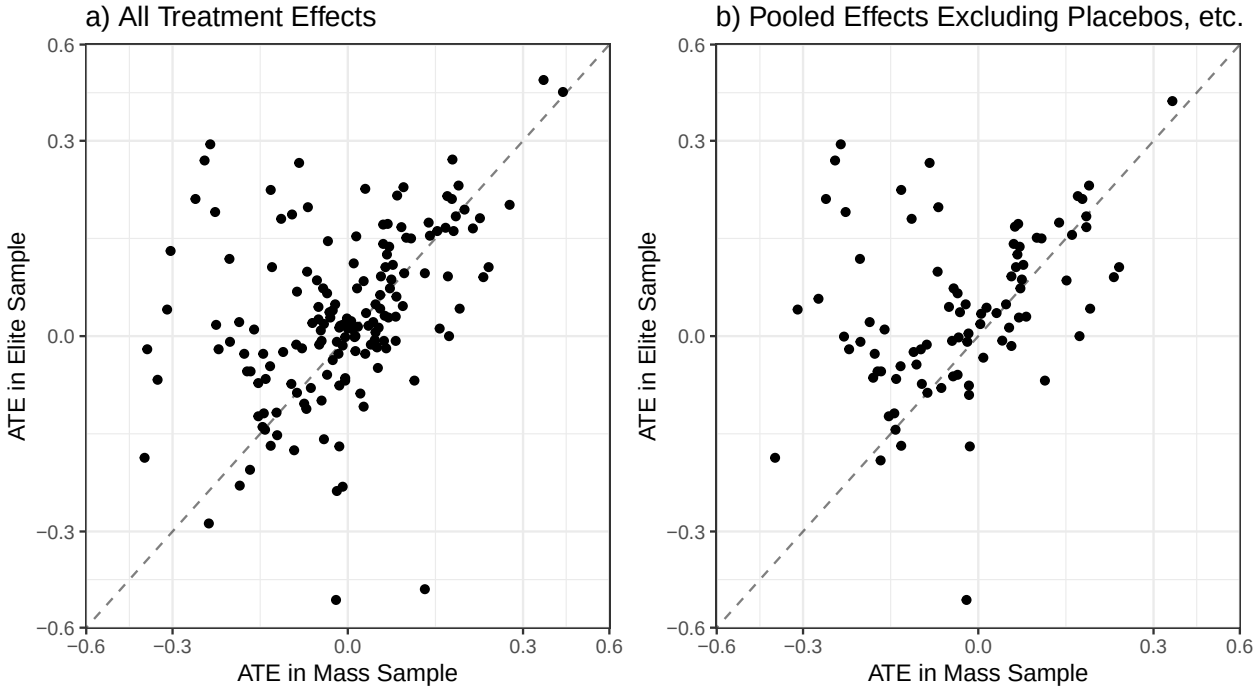

Figure 5: Scatterplots of Elite and Mass ATEs

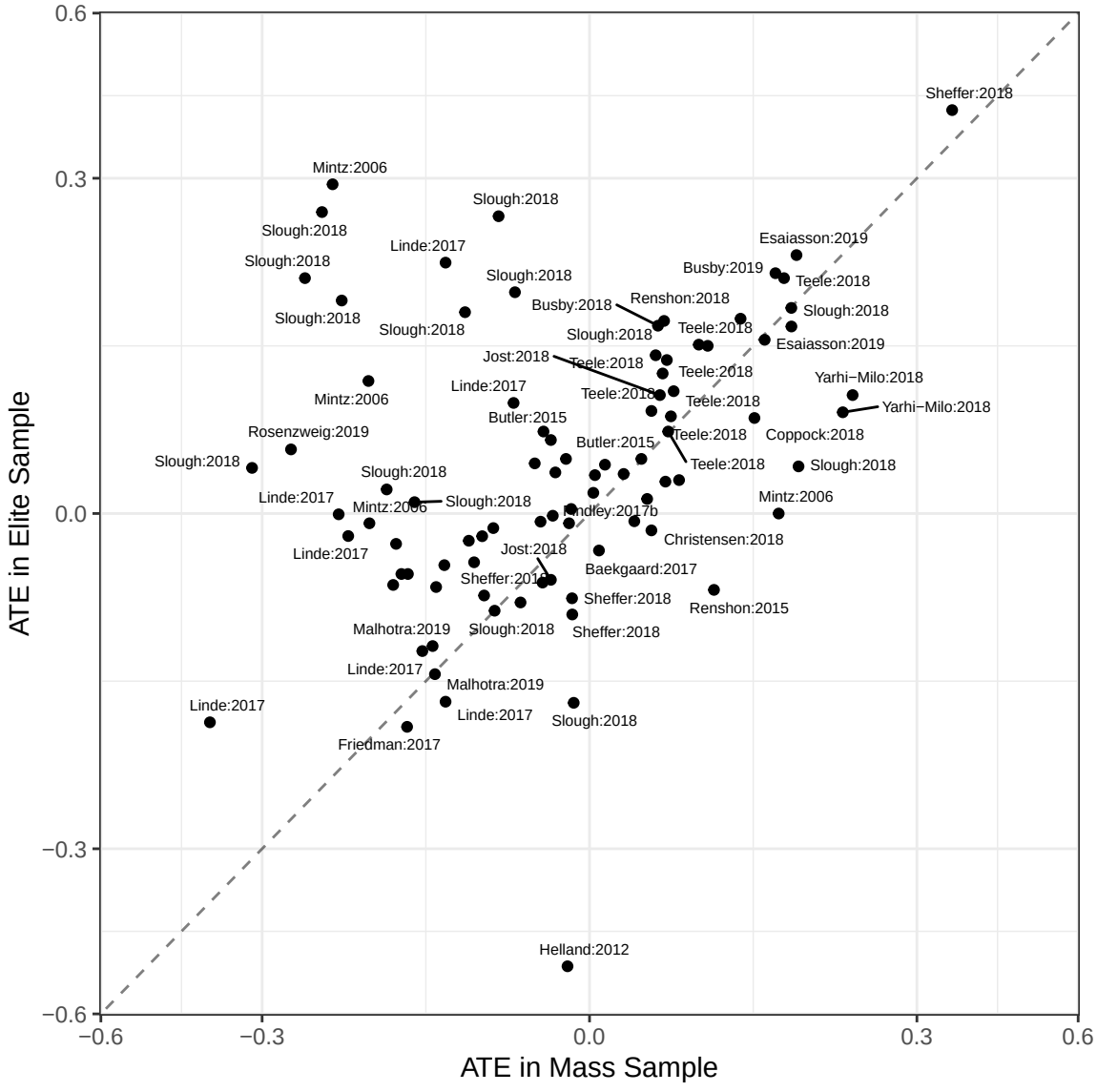

Figure 6: Scatterplots of Elite and Mass ATEs with Point Labels

3.2 Conceptual Replication of Public-Opinion Analysis (Claim 2)

Although our focus is on the meta-analysis of mass-elite reactions to treatments, we provide a conceptual replication of the second component of Kertzer's analysis using data from the meta-analysis. In particular, we examine whether masses and elites respond similarly in the control groups of these studies. For many of the studies, this is reasonably close to what Kertzer examines in the public opinion data section of his article: do elites and masses have similar political attitudes (when not exposed to treatments). Although the experiments more often concern decisions, many concern attitudes and, of course, won't reflect treatment effects in the control group. Only some of the studies deal with foreign policy decisions, so this conceptual replication shows us how close elite and mass preferences are outside the pure foreign-policy context. Figure 7 shows the mean response in the control groups for elites on the vertical axis and masses on the horizontal axis. Figure 7 (a) shows all effects before excluding placebos, non-political outcomes, and uncontroversial treatments,

while Figure 7 (b) excludes such treatments. Both also include correlations between elite and mass treatment effects.

Both figures show that elite and mass decisions/attitudes are strongly correlated. The correlation coefficients are close to the upper limit of the range of correlations Kertzer finds in his analysis of foreign policy attitudes, even after controlling for demographics (r=0.76 in 1998 to r=0.90 in 1986). Thus, his finding of generally similar opinions between elites and masses holds outside the domain of foreign policy. Although the correlations are high, it's worth noting that sometimes attitudes are different. Mintz, Redd, and Vedlitz (2006) provides an important example. He examines mass and elite responses to a foreign situation and concludes that they made very different decisions.

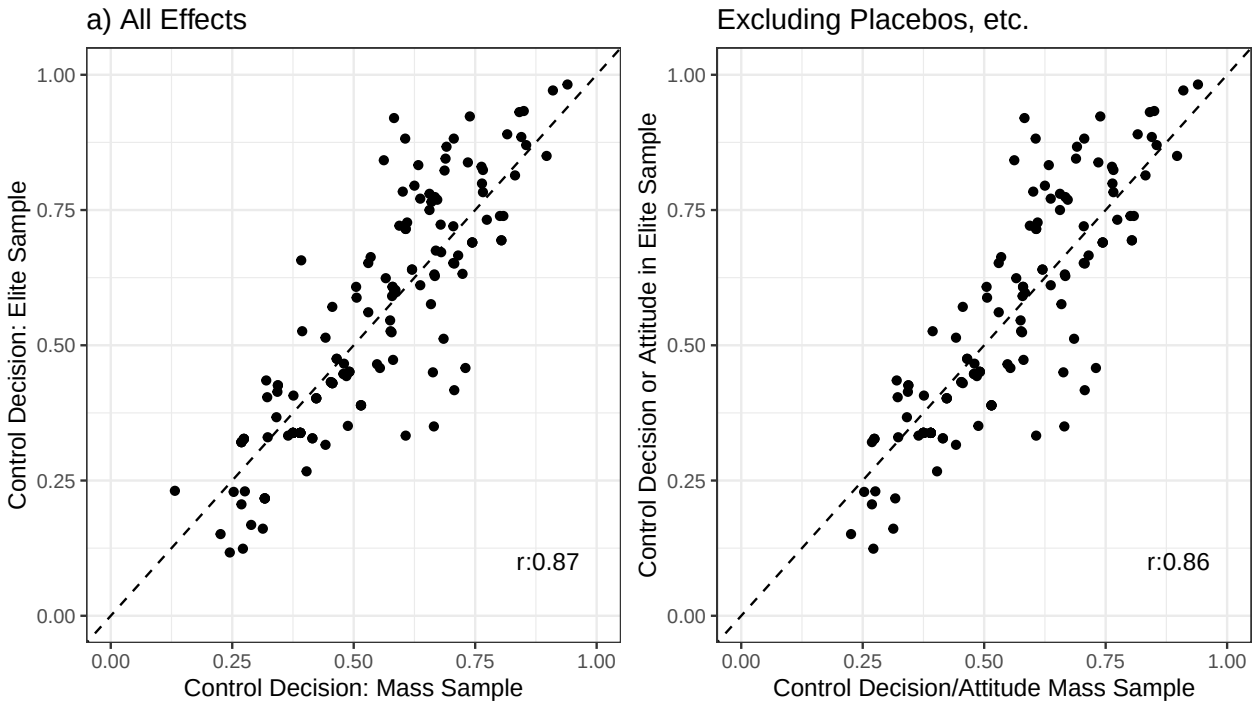

Figure 7: Scatterplots of Elite and Mass Decisions in the Control Groups

Figure 8 labels the data points.

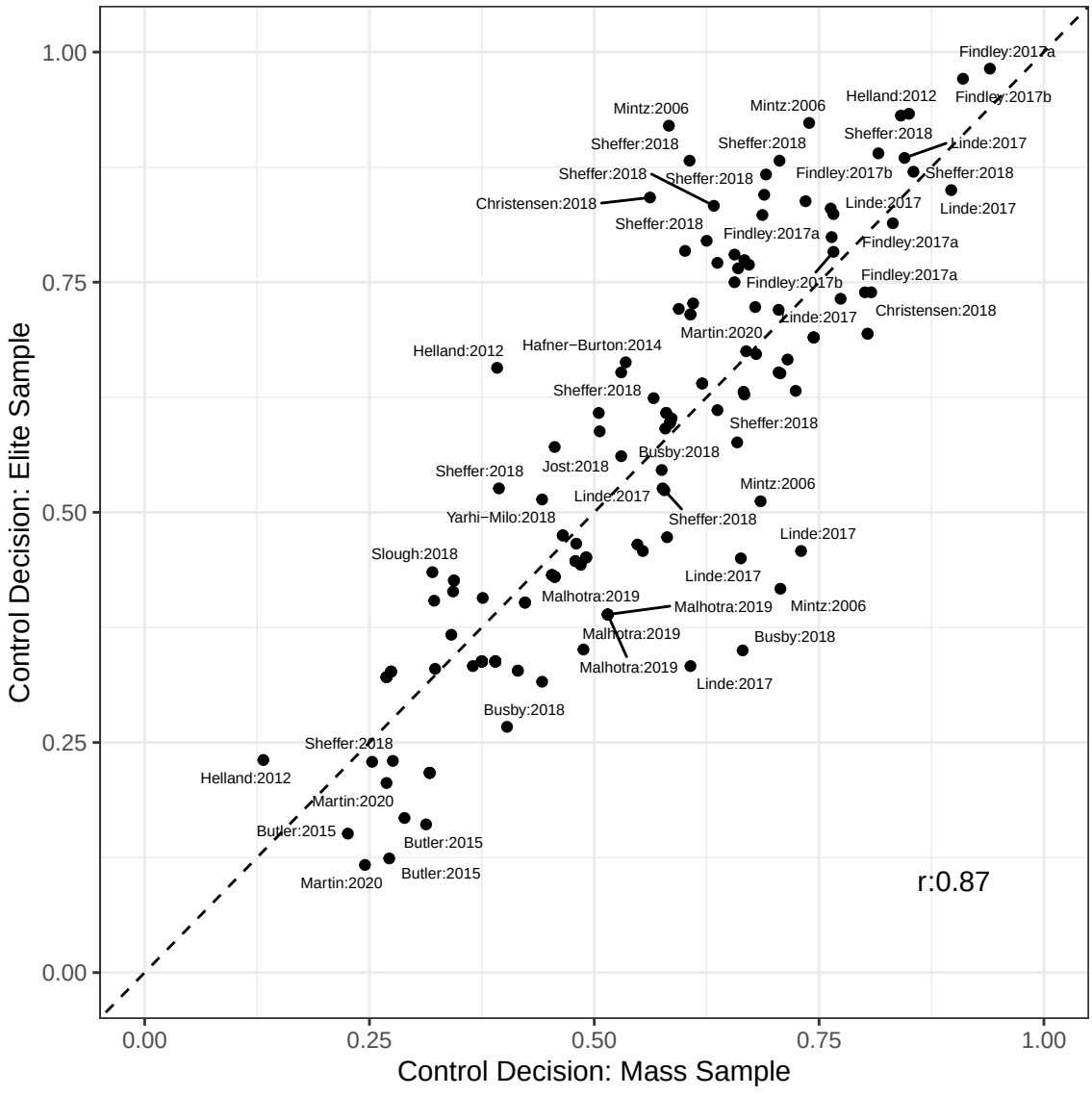

Figure 8: Scatterplots of Elite and Mass Decisions in the Control Groups

4 Conclusion

We have two main results. 1) Gaps between elite and mass decision-making are larger in studies whose manipulations appear to have worked. Only a small number of the these 162 treatments significantly affected masses (N=28) or elites (N=30). This low rate suggests the possibility that almost all of these experiments failed to successfully manipulate either masses or elites. If so, we may not be able to conclude that masses and elites respond similarly to experiments with confidence until political scientists produce more experiments with actual treatment effects or with successful manipulation checks in cases of null effects. More studies with successful manipulations are necessary to get a better answer. 2) Gaps in political attitudes between elites and masses are smaller than those found by Kertzer for foreign-policy issue even without controlling for demographics.

Appendix A

For the analysis of Claim 1, we excluded the following 23 treatment effects from the replication analysis. The replication files provide details on exactly which treatment effects we are excluding. We also recognize that reasonable people could disagree about these decisions. These decisions don't appear to have been consequential to the replication.

- We exclude 12 of 16 treatment effects from Coppock et al. (2018) because the authors only expected an effect of op-eds on the issue in the op-ed and these 12 were different issues and considered placebos by the authors.
- We exclude two of four treatment effects from Helland, Monkerud, and Løyning (2018) because in the first case the authors expected costly signals following black balls and in the second the authors expected B2 decisions following costly signals. So, these two were considered placebos by the authors.
- We exclude all six treatment effects from Martin and Raffler (2021) because the dependent variables in the study appeared to be manipulation checks in our view.
- We exclude both treatment effects from Naurin and Öhberg (2021) because it examined a nonpolitical outcome: what masses and elites thought about experiments on elites.
- We exclude the one treatment effect from Öhberg and Medeiros (2019) because it also evaluates a nonpolitical outcome: what people thought of a questionnaire.

Appendix B

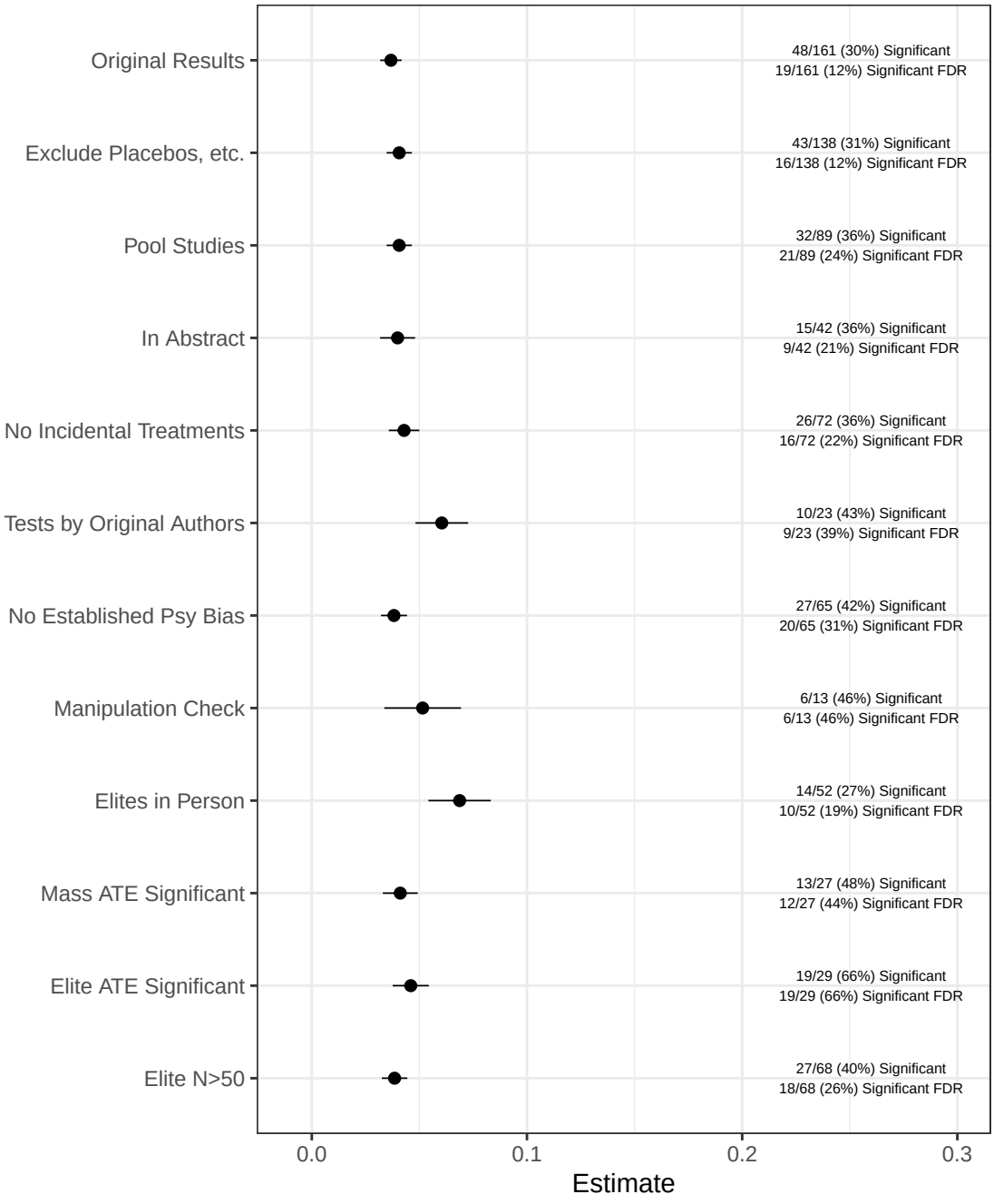

Figure A1: Meta-analytic estimate of the difference between mass and elite reaction to treatments. The figure shows the average estimate in the original study and then for various categories of studies we coded. The figure uses precision weighted averages. All estimates after the third exclude placebo treatments, non-political decisions, and uncontroversial treatments, and pool studies. All estimates exclude the extremely precise estimate from Friedman (2017).

Citations

Arel-Bundock, Vincent, Ryan Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and TD Stanley. 2022. "Quantitative Political Science Research Is Greatly Underpowered."

Coppock, Alexander, Emily Ekins, David Kirby, et al. 2018. "The Long-Lasting Effects of Newspaper Op-Eds on Public Opinion." *Quarterly Journal of Political Science* 13 (1): 59–87.

Helland, Leif, Lars Chr Monkerud, and Gjermund Løyning. 2018. "Chapter 14: Seasoned Parliamentarians Perform Worse Than Students in a Lobbying Experiment." In *At the Forefront, Looking Ahead: Research-Based Answers to Contemporary Uncertainties of Management*, 231–50. Universitetsforlaget Oslo.

Ioannidis, John PA, TD Stanley, Hristos Doucouliagos, et al. 2017. "The Power of Bias in Economics Research." *Economic Journal* 127 (605): 236–65.

Kertzer, Joshua D. 2022. "Re-Assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* 66 (3): 539–53.

Martin, Lucy, and Pia J Raffler. 2021. "Fault Lines: The Effects of Bureaucratic Power on Electoral Accountability." *American Journal of Political Science* 65 (1): 210–24.

Mintz, Alex, Steven B Redd, and Arnold Vedlitz. 2006. "Can We Generalize from Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?" *Journal of Conflict Resolution* 50 (5): 757–76.

Naurin, Elin, and Patrik Öhberg. 2021. "Ethics in Elite Experiments: A Perspective of Officials and Voters." *British Journal of Political Science* 51 (2): 890–98.

Öhberg, Patrik, and Mike Medeiros. 2019. "A Sensitive Question? The Effect of an Ethnic Background Question in Surveys." *Ethnicities* 19 (2): 370–89.

Raudenbush, Stephen W, and Anthony S Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Vol. 1. sage.

Rosenzweig, Steven C. 2021. "Dangerous Disconnect: Voter Backlash, Elite Misperception, and the Costs of Violence as an Electoral Tactic." *Political Behavior* 43 (4): 1731–54.