How conditioning on post-treatment variables can ruin your experiment and what to do about it

Jacob M. Montgomery
Dept. of Political Science
Washington University in St. Louis
jacob.montgomery@wustl.edu

Brendan Nyhan
Dept. of Government
Dartmouth College
nyhan@dartmouth.edu

Michelle Torres
Dept. of Political Science
Washington University in St Louis
smtorres@wustl.edu

November 4, 2016

Abstract

In principle, experiments offer a straightforward method for social scientists to accurately estimate causal effects. However, scholars often unwittingly distort treatment effect estimates by conditioning on variables that could be affected by their experimental manipulation. Typical examples include controlling for post-treatment variables in statistical models or eliminating observations based on post-treatment criteria. Though these modeling choices are intended to address common problems encountered when conducting experiments, they can bias estimates of causal effects. Moreover, problems associated with conditioning on post-treatment variables remain largely unrecognized in the field, which we show frequently publishes experimental studies using these practices in our discipline's most prestigious journals. We demonstrate the severity of experimental post-treatment bias analytically and document the magnitude of the potential distortions it induces using visualizations and reanalyses of real-world data. We conclude by providing applied researchers with recommendations for best practice.

Authors are listed in alphabetical order. We thank David Broockman, Daniel Butler, Eric S. Dickson, Sanford Gordon, and Gregory Huber for sharing replication data and Ryden Butler, Lindsay Keare, Jake McNichol, Ramtin Rahmani, Rebecca Rodriguez, Erin Rossiter, and Caroline Sohr for research assistance. We are also grateful to Dan Butler, Jake Bowers, Scott Clifford, Eric S. Dickson, D.J. Flynn, Sanford Gordon, Gregory Huber, Jonathan Ladd, David Nickerson, Efrén O. Pérez, Julian Schuessler, and Molly Roberts for helpful comments. All errors are our own.

1. INTRODUCTION

Experiments have become increasingly common in political science because they allow scholars to obtain unbiased estimates of causal effects without identifying and measuring all potential confounders or engaging in complex statistical modeling. Under randomization, the difference between the average outcome of observations that received a particular treatment and the average outcome of those who did not is an unbiased estimate of the true causal effect. Experimental research designs thus offer a powerful framework in which to test theories and evaluate causal claims.

However, the practice of experimental research can be far messier than this idealized description. Scholars often face difficult challenges in designing studies and analyzing results. Some participants may not pay attention to an experimental stimulus or otherwise fail to receive the treatment they were assigned. In other cases, researchers may wish to understand the mechanism that produces an experimental effect or to rule out possible alternative explanations for their findings.

Many scholars attempt to address these problems (or concerns raised about them during peer review) using practices such as dropping participants who fail to pass manipulation checks, controlling for variables measured after the treatments are administered, or including proposed mediators in the list of covariates. Though the dangers of post-treatment bias have previously been recognized in statistics, econometrics, and political science (e.g., Rosenbaum 1984; Wooldridge 2005; King and Zeng 2006; Elwert and Winship 2014; Acharya, Blackwell, and Sen 2016), many applied researchers seem unaware that these practices undermine the inferential value of randomization and distort causal effect estimates. In simpler terms, conditioning on post-treatment variables can ruin your experiment.

In this article, we provide the most comprehensive and accessible account to date of the sources, magnitude, and frequency of post-treatment bias in experimental political science research. We first identify and describe the practices that contribute to this problem and document that it continues to appear regularly in articles published in the field's top journals.

We then provide analytical and simulation evidence demonstrating how post-treatment bias contaminates experimental analyses. Third, we present reanalyses of two published studies showing how conditioning on post-treatment variables can have significant consequences for treatment effect estimates using real-world data. Finally, we conclude by identifying best practices for addressing practical challenges in experimental research without inducing post-treatment bias.

2. THE PREVALENCE OF POST-TREATMENT CONDITIONING IN CONTEMPORARY EXPERIMENTAL RESEARCH

Before turning to our main discussion of the dangers of conditioning on post-treatment variables, it is important to first address the notion that these concerns are already well known and understood. After all, published works in political science identified these issues (in passing) as problematic a decade ago (King and Zeng 2006) and more recent work has amplified these points in the context of observational research (Blackwell 2013; Acharya, Blackwell, and Sen 2016). Some readers may wonder if this exercise is needed given the increasingly widespread understanding of causal analysis in the discipline. In this section, we show that the dangers of post-treatment conditioning in experiments are either not fully understood or are being willfully ignored — our review of the published literature shows that they remain quite widespread in practice.

To demonstrate the prevalence of post-treatment conditioning in contemporary experimental research in political science, we analyzed all articles published in the *American Political Science Review* (APSR), the *American Journal of Political Science* (AJPS), and *Journal of Politics* (JOP) that included one or more survey, field, laboratory, or lab-in-the-field experiment from 2012 to 2014 ($n = 75$). Two or more of the authors of this article coded each article for whether the authors dropped cases based on potentially post-treatment criteria; controlled for or interacted their treatment variable with any variables that could plausibly be affected by the treatment (e.g., not race or gender when these were irrelevant to the study); or conditioned on variables that the original authors themselves identified as

Table 1: Post-treatment conditioning in experimental studies

Category	Prevalence
Engages in post-treatment conditioning	45.3%
Controls for/interacts with a post-treatment variable	*20.0%*
Omits cases based on post-treatment criteria	*16.0%*
Both types of post-treatment conditioning present	*9.3%*
No conditioning on post-treatment variables	53.3%
Insufficient information to code	1.3%

Sample: 2012–2014 articles in the *American Political Science Review*, the *American Journal of Political Science*, and *Journal of Politics* including a survey, field, laboratory, or lab-in-the-field experiment ($n = 75$).

outcome variables affected by the treatment without using appropriate mediation tests.[1]

Table 1 presents a summary of our results. Overall, we find that 45.3% of the experimental studies published in APSR, AJPS, and JOP from 2012 to 2014 engaged in post-treatment conditioning (34 of 75 studies). Specifically, more than one in three studies engaged in one of these practices — 20.0% (15 of 75) controlled for a post-treatment covariate in a statistical model and 16.0% of studies dropped cases based on potential post-treatment criteria (12 of 75 studies reviewed) — and almost one in ten engaged in both (9.3%, 7 studies). While some of these problems were the result of panel attrition (six studies; 8.0% of studies reviewed), the others subsetted their samples or dropped cases as a function of failed manipulation checks, noncompliance, attention screeners, or other post-treatment variables. Most strikingly, 12.0% of the total study sample conditioned on an outcome variable from a model contained in the same article (9 of 75). By contrast, no evidence was found of post-treatment conditioning for 53.3% of articles (40 of 75 studies) and 1.3% (one study) could not be coded using publicly available information.

These results indicate that almost half of the experimental studies published in our discipline's most prestigious journals during this period raise concerns about post-treatment bias. Nearly one in four drop cases based on potentially post-treatment criteria and over a quarter include post-treatment variables as covariates. Most telling, *nearly one in eight*

[1] Additional details on these coding procedures are provided in the Online Appendix.

articles directly condition on variables that the authors themselves identify as being an outcome of the experiment — an unambiguous indicator of a fundamental lack of understanding among researchers, reviewers, and editors that conditioning on post-treatment variables can invalidate results from randomized experiments. Empirically, then, the answer to the question of whether the discipline already understands post-treatment bias in experiments is clear: it does not.

3. THE INFERENTIAL PROBLEMS CREATED BY POST-TREATMENT BIAS

The pervasiveness of post-treatment conditioning in the experimental political science literature likely has many causes. However, we believe one contributing factor is a lack of clarity among applied analysts as to the source and nature of post-treatment bias. To be sure, the subjects has been covered extensively in technical work in statistics and econometrics dating back to at least to Rosenbaum (e.g., 1984). What the literature lacks, however, is a treatment of this subject that is both rigorous and accessible to non-technical readers. Indeed, in many popular textbooks, the inferential problems that come from conditioning on post-treatment covariates are discussed only briefly (Gelman and Hill 2006, Section 9.7; Angrist and Pischke 2014, pp. 214-17). Even when the subject is discussed more fully (e.g., Gerber and Green 2012), it is often dispersed among discussions of various issues such as attrition, mediation, and covariate balance. For this reason, we believe that providing a rigorous but approachable explication of the origins and consequences of post-treatment bias will help improve experimental designs and analysis in political science.

3.1. *Why experiments generate unbiased causal effects*

To understand how conditioning on post-treatment variables can distort estimates of causal effects, it is helpful to consider why experiments are so useful in the first place. Informally, a treatment can be understood to affect an outcome when its presence causes a different result than when it is absent (all else equal). In other words, we want to compare the potential outcomes for a given individual i when she receives a treatment, $Y_{[i,T=1]}$, with the outcome when she does not receive it, $Y_{[i,T=0]}$.

Though we cannot observe both potential outcomes for every individual, we can estimate the average treatment effect (ATE), which we denote as $\tau = \mathbb{E}(Y_{[T=1]} - Y_{[T=0]}) = \mathbb{E}(Y_{[T=1]}) - \mathbb{E}(Y_{[T=0]})$. Specifically, we can obtain an unbiased estimate of the ATE by comparing the mean outcome among individuals that received a treatment with the mean outcome among those who did not. We denote this quantity, which is the difference in conditional means, as follow:

$$\Delta = \mathbb{E}(Y_{[T=1]}|T=1) - \mathbb{E}(Y_{[T=0]}|T=0). \tag{1}$$

A key condition[2] that allows Δ to be equal to τ — the condition that allow us to estimate the causal effect of interest using our experiment — is that treatment assignment is unrelated to potential outcomes conditional on the observed covariates (X):

$$\textit{Assumption 1}: (Y_{[T=1]}, Y_{[T=0]}) \perp\!\!\!\perp T|X.$$

To understand why, consider a graphical causal model where Y is a linear function of a randomly assigned treatment T, covariate X, and unmeasured confounder u.[3] Further, assume that $X \perp\!\!\!\perp T$ and that $X \in \{0,1\}$. Figure 1 presents an example of a system of equations that meets these assumptions visually where C is a threshold constant and $\mathbb{1}(\cdot)$ is an indicator function.[4]

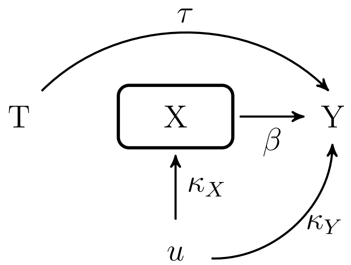
$$\begin{aligned} Y_i &= \alpha_Y + \tau T_i + \beta X_i + \kappa_Y u_i \\ X_i &= \mathbb{1}(\alpha_X + \kappa_X u_i > C), \end{aligned} \tag{2}$$

[2] Estimating a causal effect from an experiment requires other assumptions as well. We focus on the assumption of interest for our purposes here but see, e.g., Gerber and Green (2012).

[3] Pearl (2009) shows that that the graphical causal model approach is equivalent to the potential outcomes framework we use above. It is often especially helpful in clarifying which research designs can accurately recover causal estimates, which is why we employ it here.

[4] For the sake of expositional clarity, and without loss of generality, we assume that all variables are observed without error.

Figure 1: Causal graph when the covariate is unaffected by the treatment



Substituting into Equation (1), we can show the following:

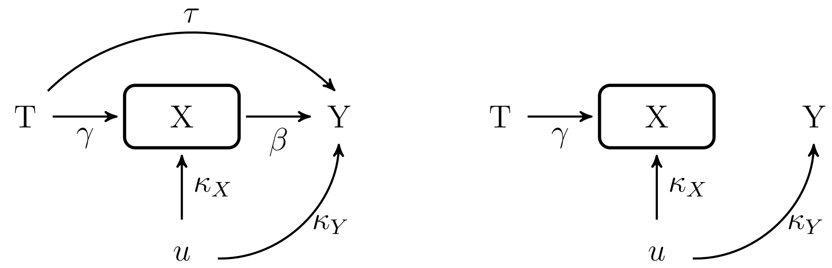
$$\begin{aligned}\Delta &= \mathbb{E}(\alpha_Y + \tau T + \beta X + \kappa_Y u | T = 1) - \mathbb{E}(\alpha_Y + \beta X + \tau T + \kappa_Y u | T = 0) \\ &= \alpha_Y + \tau \mathbb{E}(T|T = 1) + \beta \mathbb{E}(X|T = 1) + \kappa_Y \mathbb{E}(u|T = 1) \\ &\quad -\alpha_Y - \tau \mathbb{E}(T|T = 0) - \beta \mathbb{E}(X|T = 0) - \kappa_Y \mathbb{E}(u|T = 0)\end{aligned}$$

Canceling terms and recalling that $\mathbb{E}(T|T = 1) = 1$ and $\mathbb{E}(T|T = 0) = 0$, these equations can be expressed as follows:

$$\underbrace{\Delta}_{\text{Difference in means}} = \underbrace{\tau}_{\text{ATE}} + \underbrace{\kappa_Y \Big(\mathbb{E}(u|T = 1) - \mathbb{E}(u|T = 0)\Big)}_{\text{Bias from imbalance in } u} + \underbrace{\beta \Big(\mathbb{E}(X|T = 1) - \mathbb{E}(X|T = 0)\Big)}_{\text{Bias from imbalance in } X} \qquad (3)$$

Examining Equation (3), it is clear that both of the terms on the right must be zero in expectation for Δ to be an unbiased estimator of the true causal effect τ. In theory, that is precisely what experimental designs achieve. So long as we do not condition on a post-treatment factor, randomization guarantees that Assumption (1) is satisfied and both quantities go to zero. First, Assumption (1) implies that individuals in the treatment and control conditions will be balanced with respect to unobserved confounders in expectation. In mathematical terms, we can assume that $\mathbb{E}(u|T = 1) = \mathbb{E}(u|T = 0)$, which implies that the expected bias from a lack of balance in unmeasured confounders is zero. Second, Assumption (1) requires that X is not causally related to T. Thus, $\mathbb{E}(X|T = 1) = \mathbb{E}(X|T = 0)$, which means that the second term is also exactly zero in expectation. More generally, experimental analyses of data like the system represented in Figure 1 will satisfy Assumption (1) and the difference in conditional means (Δ) will be an unbiased estimate of the true ATE (τ).

Figure 2: Causal graph when covariate is a post-treatment variable



(a) Researcher-assumed causal model (b) Actual causal model

3.2. *The problem with conditioning on post-treatment variables*

While the advantages of experiments are now commonly understood in political science, the bias that can be induced by conditioning on post-treatment variables is less widely known. Conditioning on any variable affected by the treatment either directly or indirectly violates Assumption (1) and thereby invalidates any estimate of the ATE (τ) even if the experiment is otherwise perfect. More technically, Δ will be inconsistent. That is, Δ will provide an inaccurate estimate of τ *even with an infinite sample size.* Moreover, as we illustrate below, this *bias can be of almost any size and in any direction*, preventing us from characterizing its potential effects.

To understand the problem of post-treatment bias in a more specific context, we focus on a simplified example below. We assume that the researcher is attempting to estimate a model where the covariate X is assumed to have a direct effect on Y and that X is now partially a function of treatment assignment as depicted in Figure 2a. As a result, this covariate is now affected by the treatment and is thereby "post-treatment." The assumed model can be written formally as:

$$\begin{aligned} Y_i &= \alpha_Y + \tau T + \beta X + \kappa_Y u_i \\ X_i &= \mathbb{1}(\alpha_X + \gamma T_i + \kappa_X u_i > C) \end{aligned} \tag{4}$$

However, to illustrate our argument, we assume that the true causal model is such that neither the treatment nor the covariate has an effect on the outcome ($\beta = \tau = 0$). This situation is depicted in Figure 2b and can be written formally as:

$$\begin{aligned} Y_i &= \alpha_Y + \kappa_Y u_i \\ X_i &= \mathbb{1}(\alpha_X + \gamma T_i + \kappa_X u_i > C) \end{aligned} \tag{5}$$

Under these circumstances, it may seem harmless to condition on the post-treatment covariate X — after all, X has no direct effect on Y.[5] As we show below, however, this intuition is incorrect. Even in this very favorable context, conditioning on X leads to inconsistent estimates because the post-treatment covariate (X) and the outcome (Y) share an unmeasured cause (u). As a consequence, conditioning on X "unblocks" a path between T and u, which unbalances the experiment with respect to u and makes accurately estimating the causal effect of T impossible without further assumptions (Elwert and Winship 2014).[6]

Conceptually, there are two ways that researchers may mistakenly condition on post-treatment variables: dropping observations based on post-treatment criteria or controlling for post-treatment variables. We consider each practice below. Despite typically being motivated by different concerns, they both unblock the linkage between unmeasured confounders and the outcome variable and thereby bias our estimates of the average treatment effect.[7]

Dropping or selecting observations based on criteria influenced by the treatment: First, scholars may drop or select observations (either intentionally or inadvertently) as a function of some variable affected by the treatment. One possibility is that the experimental randomization will cause some types of respondents to be more likely to be omitted from the study sample than others. For instance, Malesky, Schuler, and Tran (2012) find that delegates to the national legislature in Vietnam who were randomly selected to have websites built for them on a

[5] If we instead allow X to have a direct effect on Y in the true model, the biases we describe below still hold, but the calculations involved are more complex. We make this simplifying assumption so that we can focus our exposition on the post-treatment bias that arises from unblocking the path from u to Y.

[6] In the language of Pearl (2009), this error is called "conditioning on a collider."

[7] In this portion of the manuscript, we provide specific examples of contemporary research employing practices that can lead to post-treatment bias to make our points more concrete. We wish to emphasize that we are not implying that the cited studies are themselves invalid or that the conclusions they reach are incorrect. As we illustrate below, one of the pernicious qualities of post-treatment bias is that we often cannot know how it affects reported results.

national news website were less likely to be re-nominated (Table 7.1.1). As a result, analyses of the effect of this treatment only among legislators who were re-nominated inadvertently condition on a post-treatment variable.[8]

In other instances, scholars who are concerned about respondents receiving the treatment will drop those who fail a post-treatment manipulation check or another measure of attention or compliance.[9] However, passage rates on these measures may also be affected by the treatment. Healy and Lenz (2014, 37), for instance, exclude all observations in a survey experiment where respondents failed to answer questions that were part of the treatment.

Finally, researchers may sometimes wish to study a subgroup but do not consider that the measure they use to define the subgroup was administered after the experimental intervention. For instance, Großer, Reuben, and Tymula (2013) analyze subsets of respondents based on the tax system selected by the group (Tables 2 and 3), which the authors show to be partially determined by treatment assignment (see result 2 on page 589).

In any of these cases, selecting the sample based on post-treatment variables can inadvertently unbalance the treatment and control conditions with respect to u, biasing estimates of the ATE per Equation 3. For instance, consider data generated using Model (5) above. If we calculate the difference in conditional means (Δ) using only cases that are selected into the sample based on a post-treatment variable ($X = 1$ in Equation 5 above), we obtain the following estimate:

$$\begin{aligned} \Delta &= \mathbb{E}(Y|T=1,X=1) - \mathbb{E}(Y|T=0,X=1) \\ &= \mathbb{E}(\alpha_Y + \tau T_i + \beta X_i + \kappa_Y u_i | T=1, X=1) \\ &\quad -\mathbb{E}(\alpha_Y + \tau T_i + \beta X_i + \kappa_Y u_i | T=0, X=1) \\ &= \tau + \underbrace{\kappa_Y\big(\mathbb{E}(u_i|T=1,X=1) - \mathbb{E}(u_i|T=0,X=1)\big)}_{\text{Bias from imbalance in u}} \end{aligned} \tag{6}$$

In general, this bias will *never* be zero. The reason is that the value of u_i must on average

[8] Similarly, Zhou and Fishbach (2016) show that many online experiments experience significant differential attrition by experimental condition.

[9] The negative consequences of this decision are also discussed in Aronow, Baron, and Pinson (N.d.).

be higher for observations in the control group ($T = 0$) who also meet the selection criteria ($X = 1$) under the assumed data-generating process for X. In other words, units in the control group need higher values of u_i to exceed the threshold C.[10] By selecting based on a criterion that is partially a function of unobserved covariates and the treatment, we have inadvertently created imbalance in the treatment and control conditions with respect to u.

To illustrate this problem, assume that $\alpha_x = 0$, u is distributed normally, $C > 0$, and $\gamma > 0$. The expected distribution of u among the units that did not received the treatment ($T = 0$) under these assumptions is presented the shaded region shown in the left panel of Figure 3. This distribution is clearly unbalanced by comparison with the expected distribution of u among units in the treated condition, which is presented in the right panel of Figure 3. The difference in the distributions is represented by the cross-hatching in the right panel. By selecting units to include in the study based on a post-treatment variable, we have invalidated the randomization, unbalanced the experiment with respect to an unmeasured confounder, and (as we show below) biased our estimate of the causal effect.[11]

More formally, each of these shaded regions are truncated normal distributions with $\mu = 0$ and standard deviation σ_u. Thus, we know that:

$$\begin{aligned}
\mathbb{E}(u|T=1, X=1) &= \sigma_u(\phi(\tfrac{C-\gamma}{\sigma_u}))/(1-\Phi(\tfrac{C-\gamma}{\sigma_u})), \text{ and} \\
\mathbb{E}(u|T=0, X=1) &= \sigma_u(\phi(\tfrac{C}{\sigma_u}))/(1-\Phi(\tfrac{C}{\sigma_u})),
\end{aligned} \tag{7}$$

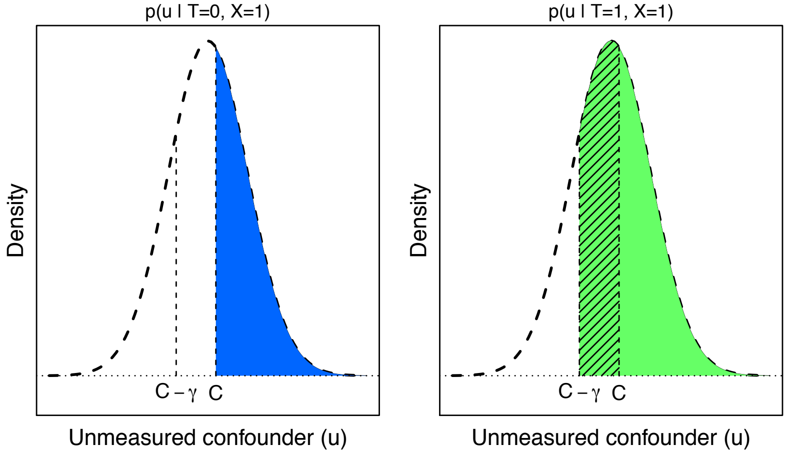
where $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and CDF for the standard normal distribution, respectively. In general, these quantities will not be equivalent unless the treatment T has no effect on the covariate X used in selection (i.e., $\gamma = 0$).

Selecting only cases where $X = 0$ creates a different pair of truncated normal distributions, but the problem is the same. The expected values of the unmeasured confounder u in

[10] We assume the treatment effect is constant across values of X (i.e., $\mathbb{E}(\tau|X = 1) = \mathbb{E}(\tau|X = 0)$) and make similar assumptions for the parameters κ_Y, κ_X, and β. Without these assumptions, the resulting bias would be *worse*. However, bias persists even if these assumptions hold.

[11] In this example, u_i will on average be higher in the control condition, which means that endogenous selection bias will be negative. However, the bias can be in any direction and of any size depending on the specific distribution of u and the values of κ_Y, κ_X, γ, and C that are chosen.

Figure 3: How conditioning on a post-treatment variable unbalances randomization



Expected distributions of an unmeasured confounder u for control (left panel) and treatment groups (right panel) when the population is selected based on post-treatment criteria (X=1) under the data-generating process in Equation 5. We assume $\alpha_x = 0$, $C > 0$, $\gamma > 0$, and that u is distributed normally.

this case would be as follows:

$$\begin{aligned} \mathbb{E}(u|T=1, X=0) &= -\sigma_u(\phi(\tfrac{C-\gamma}{\sigma_u}))/(\Phi(\tfrac{C-\gamma}{\sigma_u})), \text{ and} \\ \mathbb{E}(u|T=0, X=0) &= -\sigma_u(\phi(\tfrac{C}{\sigma_u}))/(\Phi(\tfrac{C}{\sigma_u})). \end{aligned} \tag{8}$$

As before, these quantities will generally not be equivalent unless $\gamma = 0$.

Including post-treatment variables as covariates: A closely related practice is to explicitly control for one or more post-treatment covariates in a statistical model. In some cases, well-intentioned scholars may engage in this practice in a mistaken effort to prevent omitted variable bias (which is not a concern in experiments).

In other cases, covariates may be included simply to improve the precision of the estimated treatment effect. Druckman, Fein, and Leeper (2012), for example, report an analysis of the effect of various framing manipulations on subjects' tendency to search for additional information and their expressed opinions. However, two models reported in the study (Table 4) control for measures of search behavior in previous stages of the experiment that are explicitly post-treatment (Figure 7).

A related issue is that researchers may measure a moderator after their experimental manipulation. Following the recommendations of Brambor, Clark, and Golder (2006), researchers now typically estimate these models including all lower-order terms as well as the interaction term of interest (e.g., $Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 T \times X$). For these models to be valid, the moderator X must not be affected by the experimental randomization. Any estimate of heterogeneous treatment effects that uses post-treatment moderators will otherwise be biased for the same reasons we provide above for the ATE (see, e.g., Gerber and Green 2012, 297).[12]

In both cases, scholars may wish to measure a potential control variable or moderator after the intervention under the assumption that is fixed (i.e., independent of potential outcomes). While we have engaged in this practice in our own research (citation redacted for peer review), it is potentially problematic as well. Even variables that seem likely to remain fixed when measured after treatment such as measures of racial or partisan identification can still be endogenous (e.g., Antman and Duncan 2015; Weiner 2015). Similar concerns apply to the practice of measuring strongly held attitudes like racial resentment after related interventions, which creates a risk of spillover effects (e.g., Transue, Lee, and Aldrich 2009).

Researchers may also control for post-treatment variables to try to account for noncompliance. For instance, Arceneaux (2012) hypothesizes that persuasive messages that evoke fear or anxiety will have a greater effect on attitudes. The study therefore measures subjects' level of anxiety in response to a manipulation and includes the anxiety measure and an interaction between it and the treatment in a model of issue opinion.

Another reason why post-treatment variables are included in models is to try to address complex questions about causal mechanisms (e.g., mediation). For example, Corazzini et al. (2014) studies the effect of electoral contributions on campaign promises as well on the generosity of candidates once elected (benevolence). The study shows that electoral institutions lead to more campaign promises (585), but later includes this "promise" variable

[12] In conceptual terms, post-treatment variables are not actually moderators because the potential outcomes within each possible value of the moderator fail to meet Assumption (1) above.

as a covariate — along with the treatment — in a model explaining levels of benevolence (Table 4). Because the effect of the treatment diminishes in the presence of this control, the study concludes that the effect of campaigns on benevolence "seems to be driven by the less generous promises in the absence of electoral competition" (587).

Again, including post-treatment variables as covariates for any of these reasons will bias causal effect estimates by creating imbalance with respect to u. To illustrate this point, we perform the following calculations:[13]

$$\begin{aligned}
\Delta &= \mathbb{E}(Y|T=1,X) - \mathbb{E}(Y|T=0,X) \\
&= \tau + \underbrace{\kappa_Y\big(\mathbb{E}(u_i|T=1,X) - \mathbb{E}(u_i|T=0,X)\big)}_{\text{Bias from imbalance in u}} \\
&= \tau + \kappa_y\Big[\Pr(X=0)\big[\mathbb{E}(u|T=1,X=0) - \mathbb{E}(u|T=0,X=0)\big] \\
&\qquad + \Pr(X=1)\big[\mathbb{E}(u|T=1,X=1) - \mathbb{E}(u|T=0,X=1)\big]\Big] \\
&= \tau + \kappa_Y\Big[\big(\Pr(X=0|T=1)\mathbb{E}(u|T=1,X=0)\big) - \big(\Pr(X=0|T=0)\mathbb{E}(u|T=0,X=0)\big) \\
&\qquad + \big(\Pr(X=1|T=1)\mathbb{E}(u|T=1,X=1)\big) - \big(\Pr(X=1|T=0)\mathbb{E}(u|T=0,X=1)\big)\Big]
\end{aligned} \tag{9}$$

The quantities inside the expectations are the same as those derived in Equations (7) and (8). The remaining conditional probabilities of X given T (e.g., $\Pr(X=0|T=1)$) can be directly calculated under our assumption that u_i is normally distributed.[14]

Intuitively, this result shows that controlling for a post-treatment variable creates a weighted average of the same biases that arise from dropping cases based on a post-treatment covariate. In other words, the bias induced by controlling for a post-treatment variable is a combination of the biased estimates we would get from selecting only cases where $X = 1$

[13] To simplify exposition, we focus here only on the bias resulting from the imbalance in u induced by controlling for the post-treatment variable X. As shown in Equation 3, however, bias can also arise from imbalance in observed covariates when controlling for X ($\beta(\mathbb{E}(X|T=1) - \mathbb{E}(X|T=0))$). While bias from imbalance in unobservables is even more problematic (it cannot be diagnosed and resolved directly), it is also not possible to eliminate bias from imbalance in observables without additional assumptions (see, e.g., Baum et al. N.d.).

[14] Adding the further assumption that $\kappa_X > 0$ and substituting relevant values, we obtain the following:

$$\begin{aligned}
\Pr(X=1|T=1) &= \Pr(\alpha_X + \gamma + \kappa_X u_i > C) \\
&= \Pr(u_i > \tfrac{C-\alpha_X-\gamma}{\kappa_X}) \\
&= \Phi(\tfrac{C-\alpha_X-\gamma}{\kappa_X\sigma_u}).
\end{aligned}$$

Further, $\Pr(X=1|T=1) = 1 - \Pr(X=0|T=1)$. Via similar calculations, $\Pr(X=1|T=0) = 1 - \Pr(X=0|T=0) = \Phi(\frac{C-\alpha_X}{\kappa_X\sigma_u})$. (Note: Assuming $\kappa_X < 0$ slightly alters these results, but does not meaningfully change our conclusions.)

and the estimates from selecting only cases where $X = 0$. In practice, these biases will rarely cancel out. As a result, we will be unable to correctly estimate the actual treatment effect τ.

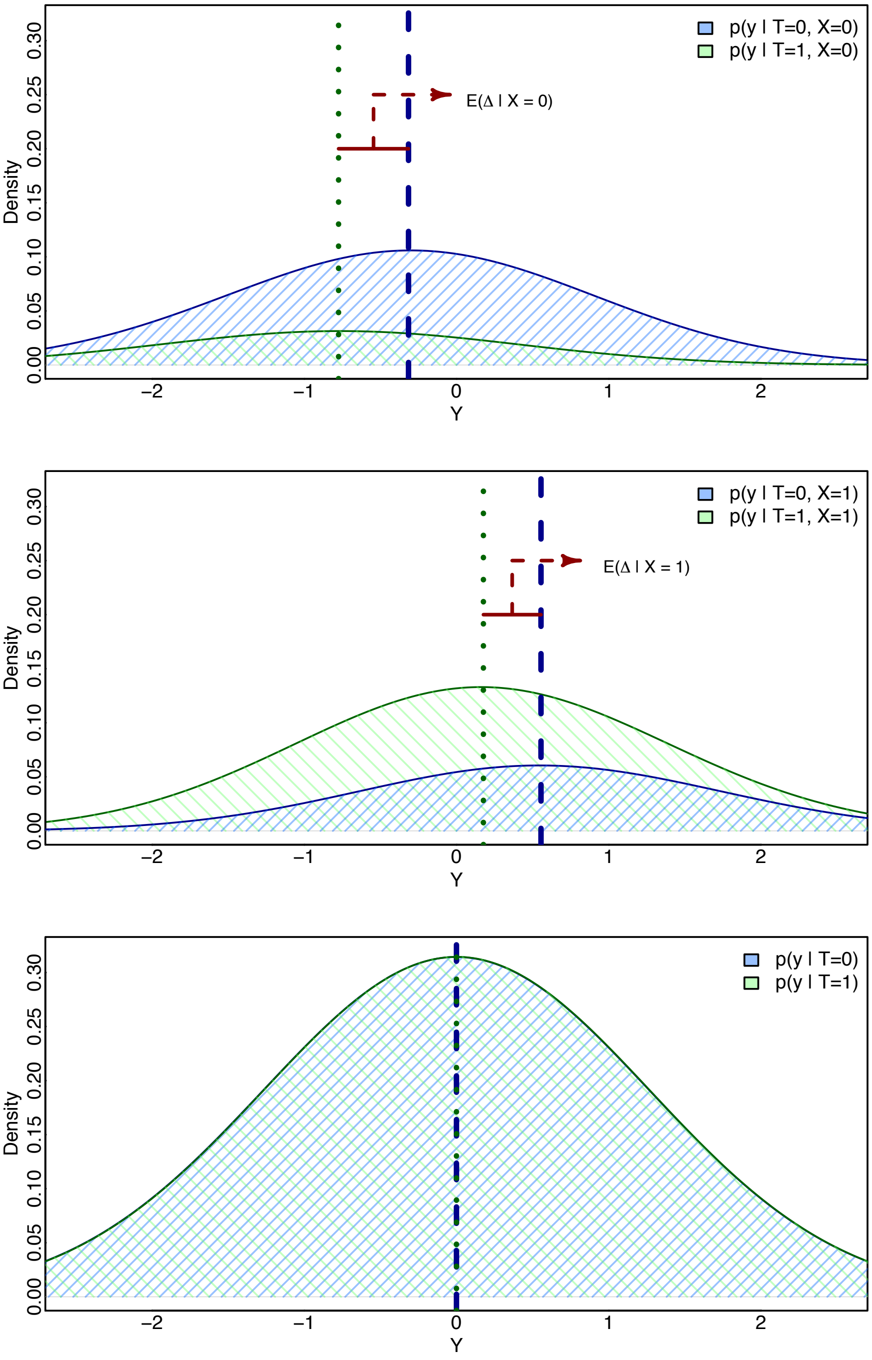
3.3. *Simulation evidence of post-treatment bias*

To further demonstrate the pernicious effects of conditioning on post-treatment variables, the appendix presents results from a series of simulations illustrating how dropping cases based on post-treatment criteria and including post-treatment variables in statistical models can severely bias experimental effect estimates. First, we show that the bias can in theory be of any size and in any direction, creating spurious effects even when the treatment *and* the post-treatment covariate have no direct effect on the outcome. Second, the magnitude of the potential bias depends on the relationship between the unmeasured confounder u and the outcome Y, which cannot be estimated by the researcher. In combination, these features make post-treatment bias a serious hazard that is difficult or even impossible to diagnose (as we discuss further below).

3.4. *Visualizing the inferential consequences of post-treatment bias*

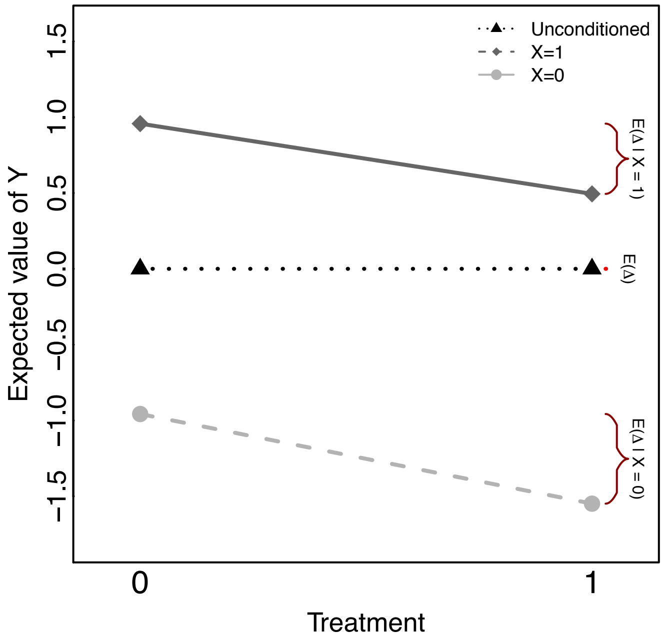
We next present two visualizations to help illustrate how inappropriately conditioning on a post-treatment variable can bias our estimated treatment effect. First, Figure 4 shows how imbalance in u induced by post-treatment conditioning can lead to mistaken inferences. Specifically, the plot shows how the distributions of the outcome Y when selecting on or controlling for X can differ systematically in the control ($T = 0$) and treatment ($T = 1$) conditions even when the unconditional (marginal) distribution of Y is unaffected by the treatment ($\tau = 0$). In this case, the effect of T appears to be negative both when $X = 0$ (top panel) and when $X = 1$ (middle panel). However, the true (marginal) effect of T is zero as shown in the bottom panel. That is, $\mathbb{E}(u|T = 1, X = 1) < \mathbb{E}(u|T = 0, X = 1)$ *and* $\mathbb{E}(u|T = 1, X = 0) < \mathbb{E}(u|T = 0, X = 0)$, which means that the treatment effect will (falsely) appear to be negative in both cases when we select on the post-treatment variable X as will their weighted combination (when we control for X).

Figure 4: How post-treatment conditioning can create spurious treatment effect estimates



The plot shows an example of how the distribution of the outcome (Y) can differ as a function of the treatment assignment (T) when conditioning on a post-treatment variable (X) even when the treatment has no effect. The vertical lines show the expected value of Y for each distribution. For this plot we use the data-generating process specified in Equation 5 and assume that $\tau = 0$, $\alpha_x = 0$, $C > 0$, $\gamma > 0$, and that u is distributed normally. Note that the effect of T appears to be negative when conditioning on X (top two panels), but the actual distribution of Y is unaffected by treatment assignment (bottom panel).

Figure 5: How conditioning on a post-treatment variable can bias treatment effect estimates



The plot shows the expected value of the outcome (Y) for different combinations of X and T. Note that within each unique value of X, the treatment appears to have a negative effect. However, the actual effect of the treatment for this data is zero, as shown by the relationship between the outcome and treatment when not conditioning on X. For this plot we use the data-generating process specified in Equation 5 and assume that $\alpha_x = 0$, $C > 0$, $\gamma > 0$, and u distributed normally.

Similarly, Figure 5 shows the expected value of the outcome for each possible combination of Y and X in this example, which again assumes that the true treatment effect (τ) is zero. The figure shows that the treatment appears to have a negative effect on the outcome when holding constant the value of X. However, the effect of the treatment is actually zero, as shown by the unconditional relationship between T and Y. Conditioning on the post-treatment variable X by dropping cases based on it or directly including the covariate in a statistical model will thus lead to a false conclusion about the treatment effect.

3.5. *The inadequacy of empirical tests for post-treatment bias*

Finally, it is important to note that post-treatment bias cannot be easily diagnosed or remedied empirically. A common belief apparent in the literature is that researchers can rule out post-treatment bias by conducting some kind of hypothesis test about the balance of

X between the treatment and control conditions. Scholars might, for instance, conduct a bivariate regression where X is a dependent variable and T is the explanatory variable. The argument is that if we fail to reject the null hypothesis $H_0 : E(X|T = 0) = E(X|T = 1)$, post-treatment bias should not bias any analyses that condition on X.[15]

Even in this simple example, however, the bias will not be eliminated unless γ is precisely zero – something that cannot be established using traditional hypothesis testing. Failing to reject the null hypothesis is not itself direct evidence that supports the null hypothesis (Gill 1999). To see the underlying problem, note that in Equations (6) and (9) the bias is not only a function of the imbalance in the confounder u but also a function of κ_Y, the effect of the unmeasured covariate on the outcome. Thus, the resulting bias is explicitly a function of a parameter that researchers have no ability to estimate. When κ_Y is large, even a small degree of imbalance in u can result in significant bias. The failure to reject the null hypothesis that X is not related to T is thus not sufficient evidence to rule out post-treatment bias.[16] We return to this point below.

4. HOW POST-TREATMENT BIAS CAN CONTAMINATE REAL-WORLD DATA

4.1. *Reanalysis 1: Dickson, Gordon, and Huber (2015)*

In order to illustrate the consequences that post-treatment practices may have on real-world inferences, we replicate and reanalyze Dickson, Gordon, and Huber (2015) (henceforward DGH), a lab experiment that manipulate rules and information to assess their effect on citizens' propensity to support or hinder authorities.

Participants were assigned to groups of five people. Within each group, one individual was randomly assigned to be an authority and the rest acted as citizens. Each group played multiple sessions in which citizens first decide whether they wanted to contribute to a

[15] We acknowledge that researchers may wish to trade off the possibility of mild bias for large efficiency gains from controlling for X. Our point here is only that such tests cannot eliminate the possibility of post-treatment bias.

[16] It is important to note that we have made several simplifying assumptions including that treatment effects are constant across values of X and that u is distributed normally. In a more general setting, it will often not be possible to provide reasonable bounds for the potential bias from conditioning on a post-treatment variable without further assumptions (Aronow, Baron, and Pinson N.d.).

common pot, which each citizen and the authority receive a share of later. After observing the contributions, the authority decides whether to target a group member for enforcement for failing to contribute to the pot. If a member was penalized in this way, citizens were given the option to help or hinder the authority (with a cost) and then everyone observes these actions and whether enforcement was successful.

A 2 × 2 design varies the institutional environment of each group. One dimension was related to the way authorities were compensated: fixed wage (*salary*) versus compensation based on penalties collected (*appropriations*). The other dimension, transparency, involved two treatments related to the amount of information citizens received regarding the actions of other players: either knowing only that someone had been targeted but not knowing anyone's contribution choices (*limited information*) versus fully observing contribution choices and target selection (*full information*).

DGH explores the propensity of citizens to hinder or help the authority across the different treatment groups. However, in an attempt to test certain theoretical claims and to diminish the effects of "perverse" participant behavior, the study follows two common approaches in the literature on experimental economics and behavioral games that raise concerns about post-treatment bias. First, DGH exclude cases in which they observe what they call "perverse" targeting of a contributor when at least one group member did not contribute.[17] Intuitively, dropping cases of perverse targeting might seem to allow researchers to focus on the effect of the treatment among individuals where the researchers believe *ex ante* that the treatment of interest is operative—i.e., cases where the subjects seem to have correctly understood the incentives. However, perverse targeting is itself a post-treatment behavior given the expected effect of institutions and information on citizen and authority behavior. For example, Table 1 of the original study reports that the difference in predation levels between compensation treatments is significant. It is therefore incorrect to exclude cases in which contributors are targeted, as DGH does in Table 2. That is, the restricted

[17] The other types are "predatory" targeting of a citizen when all citizens contribute and "resolute" targeting of non-contributors.

sample will *not* provide an estimate of the effect of the treatment among the subset of individuals who understand the incentives correctly, but rather a biased estimate that does not correspond to a causal estimand of theoretical interest.

Second, DGH controls for lagged measures of average contributions, average resoluteness, and perverse or predatory targeting in some analyses.[18] The goal of including these control variables is to account for the dynamic nature of the data-generating process and ensure that the effects of the treatments at time t are not fully mediated by the events and choices of players in previous periods. Although this approach is again a standard practice in the literature, it does not provide an estimate of the treatment effect that is unmediated by behavior in previous rounds. Instead, the result is a biased treatment effect estimate that does not correspond to a meaningful causal estimand.

We are interested in assessing the differences between the study's conclusions in Tables 2 and 4 about the causes of *net assistance to the authority* (which we successfully replicated) and those that would be obtained if post-treatment conditioning were avoided. We first test whether these variables, which are used both as control variables (the lagged measures) and to restrict the sample (excluding cases of contributor targeting), differ between treatment groups — a possible indication of post-treatment bias.[19] Figure A5 in the Online Appendix provides one indication that these measures are affected by the experimental randomization — the appropriations treatment has a significant effect on each of the lagged behavioral measures among groups in the low information condition.[20]

Table 2 demonstrates that post-treatment conditioning induces substantial differences

[18] See, however, columns 1 and 3 in Table 2 and columns 1, 3, and 5 in Table 4, which exclude controls. More generally, our concerns do not apply to all of the findings in the original DGH analysis.

[19] While failing to reject the null hypothesis in such tests does not rule out the possibility of post-treatment bias for the reasons discussed in Section 3.5, rejecting the null hypothesis can be considered evidence that a covariate is post-treatment.

[20] This finding does not mean that post-treatment conditioning will otherwise not induce bias. As we demonstrate, null results on balance tests do not preclude the existence of post-treatment bias.

Table 2: Treatment effect differences by post-treatment conditioning

	Full sample (1)	Lagged controls (2)	Drop cases (3)	Drop/controls (4)
Appropriations effect — full information (versus salary/full information)	-1.055*** (0.438)	-1.053*** (0.344)	-0.657* (0.366)	-0.790*** (0.299)
Appropriations effect — limited information (versus salary/limited information)	-0.368 (0.347)	-0.183 (0.490)	-0.789 (0.571)	-0.915 (0.564)
Limited information effect — salary (versus salary/full information)	-0.575 (0.369)	-0.529 (0.322)	-0.742* (0.409)	-0.719** (0.347)
Limited information effect — appropriations (versus appropriations/full information)	0.112 (0.416)	0.341 (0.47)	-0.874 (0.537)	-0.844 (0.528)
Period indicators	Yes	Yes	Yes	Yes

*p < .1; **p < .05; ***p < .01. Data from Dickson, Gordon, and Huber (2015). The models reported in columns 3 and 4 exclude groups with any targeting of contributors as in the original study.

in the significance and magnitude of the estimated effects of the treatments.[21] The first column, which omits any post-treatment controls or conditioning, shows that the appropriations treatment is significant only in the full information condition. By contrast, the effect of appropriations among groups with limited information and the effect of limited information in either compensation group are not distinguishable from zero. These results are largely unchanged when we include lagged behavioral controls in the second column. However, when we instead drop cases based of contributor targeting in the third column, the limited information treatment becomes significant at the $p < .10$ level in the salary condition. This effect becomes significant at the $p < .05$ level in the fourth column when we drop cases *and* include lagged controls. Moreover, the magnitude of the effect estimates varies substantially when we condition on post-treatment variables in the third and fourth columns and can even change direction. Most notably, the magnitude of the estimated effect of the appropriations treatment in the limited information condition more than doubles in magnitude and becomes nearly statistically significant ($p < .11$).

[21] These estimates correspond to the treatment effect estimates reported in Tables 2 and 4 of Dickson, Gordon, and Huber (2015) (which we replicated successfully), though they differ slightly due to the fact that the period effects in the original study were estimated using only subsets of the data (details available upon request). See the Online Appendix for full model results.

These findings offer new insight into the results in Dickson, Gordon, and Huber (2015). Our results replicate the appropriations treatment effect for the full information groups reported in Table 2 of the original study. However, our analysis raises concerns about post-treatment bias for both the limited information effect in the salary condition and the appropriation effect in the limited information condition that are reported in Table 4 of the original study. Dickson, Gordon, and Huber (2015) correctly notes that both models are sensitive to model specification; our analysis suggests that these results may be attributable to post-treatment bias.

4.2. *Reanalysis 2: Broockman and Butler (2015)*

To further demonstrate the potential bias that can result from conditioning on post-treatment variables, we leverage replication data from Broockman and Butler (2015), which does *not* engage in post-treatment conditioning, to demonstrate how controlling for or selecting on post-treatment variables can distort experimental findings. The article reports the results of field experiments conducted in cooperation with sitting politicians who randomly varied the content of letters they sent to constituents. Below we use data from the original article to demonstrate the bias that can result from inappropriately conditioning on a manipulation check.[22]

Broockman and Butler's first study included a manipulation check measuring whether respondents reported having received a letter from the legislator, but correctly refrained from conditioning on this variable. We do so, however, to illustrate how it could affect the inferences that would be drawn from the study. We find that dropping cases that were assigned to treatment but failed the manipulation check (a common practice) makes the sample unbalanced — prior approval of the legislator, a key pre-treatment covariate, is significantly higher in the treatment group after these cases have been dropped. Specifically, a *t*-test comparing mean legislator approval between the control group (0.17) and the treatment

[22] The results below are based on a reanalysis of data from the first study in Broockman and Butler (2015), but we obtain substantively similar results when we condition on a manipulation check from the second study in the article as well — see the Online Appendix for further details.

Table 3: The effects of post-treatment bias: Legislator approval models

			Drop if fail manipulation check	
	Original	Covariate	Treatment/control	Treatment only
Sent policy letter (treatment)	0.135**	0.074	0.097	0.232**
	(0.058)	(0.064)	(0.118)	(0.071)
Prior legislator approval	0.220**	0.200**	0.237**	0.205**
	(0.025)	(0.026)	(0.042)	(0.029)
Recall receiving a letter		0.198**		
		(0.068)		
Constant	0.251**	0.214**	0.370**	0.253**
	(0.041)	(0.043)	(0.108)	(0.041)
R^2	0.31	0.34	0.34	0.34
N	193	183	66	146

* $p < .10$, ** $p < .05$. OLS regression results; standard errors in parentheses.

group (0.28) is not statistically significant for the full sample ($p = .21$). However, dropping respondents that were assigned to treatment and failed the manipulation check causes significant imbalance — we can reject the null hypothesis of no difference of means in legislator approval between the control (0.17) and treatment (0.56) groups ($p < .05$).[23]

To demonstrate the potential for post-treatment bias that conditioning on manipulation checks can create, we next reanalyze the experimental data from this study, which considers the effects of sending a policy letter to constituents who disagree with its content on legislator job approval.[24] Table 3 presents results from the following models: the original model estimated by Broockman and Butler (2015) that includes only a pre-treatment control for prior approval (first column), a model that includes the manipulation check as a covariate in the regression (second column), and models that instead drop respondents in both conditions or only those in the treatment condition who did not recall receiving a letter from the legislator (the third and fourth columns, respectively).[25]

[23] See Figure A7 in the Online Appendix for a visualization of the resulting imbalance.

[24] The authors present results with the outcome variable coded three different ways (Broockman and Butler 2015, 6). We only present results for the model where approval is coded as a binary outcome.

[25] Self-reported recall of receiving a letter from the legislator in this subsample (constituents who disagree with the issue position in question) was 36% ($n = 183$) overall — 55% in the treatment group ($n = 91$) and 17% in the control group ($n = 92$). This increase is statistically significant (Broockman and Butler 2015, 5).

The results indicate that the inferences we would draw from the Broockman and Butler (2015) data differ substantially depending on whether we control for or select on the post-treatment manipulation check variable. The first column verifies the authors' finding that sending a policy letter to voters who disagree with its content has a positive and reliable effect on legislator approval ($p < .05$). However, we cannot reject the null hypothesis of no effect when we control for the manipulation check (second column). Similarly, we cannot reject the null of no effect when we drop respondents who fail the manipulation check in both conditions even though the sample remains balanced on prior legislative approval (third column; balance test results available upon request). Finally, the average treatment effect estimate is instead biased upward if we drop respondents who fail the manipulation check in the treatment condition only, which as we show above leads to imbalance between the treatment and control groups in prior legislator approval. These results demonstrate that conditioning on manipulation checks can lead to substantively different conclusions using real-world data.[26]

5. RECOMMENDATIONS FOR PRACTICE

In this section, we provide recommendations to help researchers avoid the problems we describe above. We summarize several motivations for post-treatment conditioning — non-compliance, attrition, efficiency concerns, heterogeneous treatment effects, and mechanism questions — and briefly explain how to address these issues without inducing bias.

5.1. *Use instrumental variables to address non-compliance*

One frequent problem in experiments is noncompliance. Participants frequently fail to receive the assigned treatment due to logistical problems, failure to understand the rules of an experiment, or inattentiveness. These cases represent so-called one-sided noncompliance — the case in which some treatment group members fail to receive an assigned treatment but

[26] Of course, the difference between the original model estimate and the results obtained using post-treatment conditioning is not necessarily itself statistically significant. Our point instead is that scholars who condition on a post-treatment variable would reach mistaken conclusions in a null hypothesis significance test.

no control group members are treated. In other cases, scholars use an encouragement design or otherwise try to induce exogenous variation in a treatment of interest that cannot be manipulated directly. In these cases, scholars may face so-called "two-sided non-compliance" in which some control group members receive the treatment and some participants in the treatment group do not.

Dropping non-compliers or controlling for compliance will in general induce bias for the reasons described above. Its presence also changes the experimental estimand of a simple treatment-control comparison from the ATE to an intent-to-treat effect (ITT) that may not be the research question of interest. There are no easy solutions to this problem.

Besides estimating the ITT, the best approach to noncompliance is to estimate a two-stage least squares model using random assignment as an instrument for treatment status. The quantity estimated by this approach, which is known as the complier average causal effect (CACE), represents the average treatment effect among compliers — those respondents who would be treated if assigned to treatment but not otherwise. In the case of one-sided non-compliance, it is easily understood as the average treatment effect on the treated. However, interpretation is somewhat more difficult for two-sided non-compliance because it represents the ATE for an unobserved subset of compliers (Angrist, Imbens, and Rubin 1996; see Gerber and Green 2012, 131–209 for more on these points).

5.2. *Use double sampling, extreme value bounds, or instruments to account for attrition*

Experimental studies often suffer from attrition and non-response, leading many analysts to exclude observations from their final analysis. However, unless attrition and non-response are unrelated to potential outcomes and treatment, this practice is equivalent to conditioning on a post-treatment variable.

There are several approaches that aim to achieve a better estimation of treatment effects in the presence of non-random attrition. Gerber and Green (2012) recommend extreme value bounds (Manski 1989), where analysts estimate the largest and smallest ATEs that we could obtain if the missing information were filled in with extreme outcomes. An alternative approach to collect outcome data among some subjects with missing outcomes in a double sampling scheme (Gerber and Green 2012; Hansen and Hurwitz 1946). A recent approach that combines double sampling with extreme value bounds is presented in Aronow et al. (N.d.). Finally, Huber (2012) presents a method to ameliorate the bias arising from attrition through a combination of inverse probability weighting and instrumental variables for missingness.

5.3. *Use pre-treatment moderators, control variables, and attention checks*

Researchers often wish to control for other variables beside compliance measures in their experimental analyses. Though it is not necessary to do so (randomization eliminates omitted variable bias), regression adjustment for covariates has been shown to induce only minor bias and to potentially increase efficiency under realistic conditions (Green and Aronow N.d.; Lin 2013). Including control variables is therefore potentially appropriate as long as researchers avoid specification searches. However, this recommendation applies *only* to pre-treatment covariates that are independent of potential outcomes (Gerber and Green 2012, 97–105).

Similarly, some researchers may wish to test for heterogeneous treatment effects by interacting their treatment variable T with a potential moderator X. However, as we note above, controlling for and interacting a treatment effect indicator with a moderator that could be affected by the experimental manipulation risks post-treatment bias. As Huber and Lapinski (2006, 424) argue, moderators that are vulnerable to treatment spillovers like racial resentment should be collected prior to a manipulation due to the possibility of post-treatment bias. However, measuring a relevant moderator before an experimental manipulation does raise concerns about priming a relevant attitude or identity. We acknowledge the difficult design tradeoff that this possibility raises and discuss the need for further research on the topic in the conclusion.

A third related concern is that scholars often wish to use measures of respondent attention that are not manipulation checks to drop inattentive respondents (e.g., Oppenheimer, Meyvis, and Davidenko 2009; Berinsky, Margolis, and Sances 2014). All attention checks should be collected before the experimental randomization to avoid post-treatment bias. Researchers may neglect this issue when the content of the attention check is not directly related to the experimental randomization, but many treatments could potentially affect passage rates through other mechanisms (changing respondent engagement with the study, making certain considerations more or less accessible, affecting the contents of working memory, etc.). If passage rates on an attention check are affected by the treatment for any reason, then dropping respondents who fail it would be the equivalent of dropping cases on a post-treatment covariate and would again risk bias.[27]

5.4. *Use mediation models to study mechanisms*

Finally, some researchers include post-treatment covariates as control variables in an effort to test theories about potential mechanisms for the causal effect of interest or to try to estimate the direct effect of a treatment that does *not* pass through a given mediator. However, this approach, which is frequently attributed to Baron and Kenny (1986), does not identify the direct or indirect effects of interest absent additional assumptions. A better approach is a modern mediation method like that recommended by Imai, Keele, and Tingley (2010) or related alternatives such as marginal structural models (Robins, Hernan, and Brumback 2000) or structural nested mean models (Robins 1999).

6. CONCLUSION

This article provides the most systematic account to date of the problems with and solutions to a recurring problem in experimental political science: conditioning on post-treatment variables. We find that a significant fraction of the experimental studies published in the discipline's most prestigious general interest journals drop observations based on post-treatment variables or control for post-treatment variables in their statistical analysis. These practices are typically employed in an effort to address practical problems like non-compliance or to try to answer difficult inferential questions such as questions about causal

[27] This concern unfortunately applies even if the research cannot reject the null of no difference in passage rates between conditions (see Section 3.5).

mechanisms. Though these intentions are laudable, we demonstrate that post-treatment conditioning undermines the value of randomization and biases treatment effect estimates using analytical and simulation evidence as well as a reanalysis of data from two published studies. We conclude with a brief overview of recommendations for practice, including using only pre-treatment covariates as moderators, control variables, and attention checks; addressing noncompliance with instrumental variables models; and estimating mediation models to address questions about causal mechanisms.

As noted above, we recommend avoiding selecting on or controlling for post-treatment covariates. This issue does raise additional practical challenges. If a panel design cannot be used that includes a prior wave before the experimental randomization, scholars must ask respondents about relevant covariates *before* the experimental manipulation during a single survey. Such designs must be implemented carefully. In particular, asking questions about certain highly salient covariates like group identification immediately before an outcome variable can affect subsequent responses (e.g, Kosloff et al. 2010; Leach et al. 2010). Scholars may therefore be concerned about priming effects contaminating their study (e.g., Valentino, Hutchings, and White 2002, 78). Though such effects are not always observed, scholars should still seek to carefully separate pre-treatment questions from their experiment and outcome measures to avoid inadvertently affecting the treatment effects they seek to estimate. However, further research is needed on how to minimize potential priming effects and/or design experimental manipulations to estimate true (unprimed) causal effects without inducing post-treatment bias.

Before concluding, it is worth considering how the institutions and practices of academic research may encourage post-treatment bias. Many of the practices described above are included in experimental analyses in response to or in anticipation of reviewer demands. We hope this article helps convince reviewers and editors not to make such requests and provides evidence researchers can cite to justify avoiding such practices.

Finally, we wish to note that the practice of conditioning on post-treatment variables does not exclusively pertain to experimental studies. The prevalence of and bias from post-treatment conditioning is likely to be even greater in observational studies than in experiments (see, e.g., the estimates in Acharya, Blackwell, and Sen 2016). However, it is not unreasonable to expect that experimentalists should be particularly careful to avoid post-treatment bias. The internal validity that randomization provides in experiments makes the avoidance of post-treatment bias an easier – and arguably more important – task than in observational studies. In many cases, the usefulness of an experiment rests on its strong claim to internal validity, not the participants (often unrepresentative) or the manipulation (often somewhat artificial). And unlike in observational studies, the nature and timing of the treatment in experiments is typically unambiguous, making it easier for scholars to determine which variables are potentially post-treatment and to avoid conditioning on them.

In total, the evidence we provide demonstrates that post-treatment conditioning is a frequent and significant problem in political science. However, we also show that scholars can address the concerns that motivate the use of these practices using existing analytical approaches. Happily, then, the bias that post-treatment conditioning introduces into so much experimental research can easily be avoided.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* DOI: https://doi.org/10.1017/S0003055416000216.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91 (434): 444–455.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2014. *Mastering 'metrics: the path from cause to effect*. Princeton,NJ: Princeton University Press.

Antman, Francisca, and Brian Duncan. 2015. "Incentives to identify: racial identity in the age of affirmative action." *Review of Economics and Statistics* 97 (3): 710–713.

Arceneaux, Kevin. 2012. "Cognitive biases and the strength of political arguments." *American Journal of Political Science* 56 (2): 271–285.

Aronow, Peter, Jonathon Baron, and Lauren Pinson. N.d. "A note on dropping experimental subjects who fail a manipulation check." Unpublished manuscript. Downloaded November 3, 2016 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2683588.

Aronow, Peter M., Alexander Coppock, Alan Gerber, Donald P Green, and Holger L. Kern. N.d. "Combining Double Sampling and Bounds to Address Non-Ignorable Missing Outcomes in Randomized Experiments." Unpublished manuscript. Downloaded November 3, 2016 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2305788.

Banks, Antoine J, and Nicholas A Valentino. 2012. "Emotional substrates of white racial attitudes." *American Journal of Political Science* 56 (2): 286–297.

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.

Baum, Matthew A., Justin de Benedictis-Kessner, Adam Berinsky, Dean Knox, and Teppei Yamamoto. N.d. "Disentangling the causes and effects of partisan media choice in a polarized environment: Research to date and a way forward." Unpublished manuscript. Downloaded November 3, 2016 from http://www.democracy.uci.edu/newsevents/events/conference_files/baum_2016_effectsofpartisanmediachoice.pdf.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–753.

Blackwell, Matthew. 2013. "A framework for dynamic causal inference in political science." *American Journal of Political Science* 57 (2): 504–520.

Bolsen, Toby, Paul J Ferraro, and Juan Jose Miranda. 2014. “Are Voters More Likely to Contribute to Other Public Goods? Evidence from a Large-Scale Randomized Policy Experiment.” *American Journal of Political Science* 58 (1): 17–30.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. “Understanding interaction models: Improving empirical analyses.” *Political Analysis* 14 (1): 63–82.

Broockman, David E., and Daniel M. Butler. 2015. “The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication.” *American Journal of Political Science* DOI: `https://doi.org/10.1111/ajps.12243`.

Clifford, Scott. 2014. “Linking issue stances and trait inferences: A theory of moral exemplification.” *Journal of Politics* 76 (3): 698–710.

Corazzini, Luca, Sebastian Kube, Michel André Maréchal, and Antonio Nicolo. 2014. “Elections and deceptions: an experimental study on the behavioral effects of democracy.” *American Journal of Political Science* 58 (3): 579–592.

Dickson, Eric S., Sanford C. Gordon, and Gregory A. Huber. 2015. “Institutional Sources of Legitimate Authority: An Experimental Investigation.” *American Journal of Political Science* 59 (1): 109–127.

Druckman, James N, Jordan Fein, and Thomas J Leeper. 2012. “A source of bias in public opinion stability.” *American Political Science Review* 106 (02): 430–454.

Dunning, Thad, and Janhavi Nilekani. 2013. “Ethnic quotas and political mobilization: caste, parties, and distribution in Indian village councils.” *American Political Science Review* 107 (1): 35–56.

Elwert, Felix, and Christopher Winship. 2014. “Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable.” *Annual Review of Sociology* 40: 31–53.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gerber, Alan S., and Donald P. Green. 2012. *Field experiments: Design, analysis, and interpretation*. New York: WW Norton.

Gill, Jeff. 1999. “The insignificance of null hypothesis significance testing.” *Political Research Quarterly* 52 (3): 647–674.

Green, Donald P., and Peter M. Aronow. N.d. “Analyzing Experimental Data Using Regression: When is Bias a Practical Concern?” Unpublished manuscript. Downloaded November 3, 2016 from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1466886`.

Großer, Jens, Ernesto Reuben, and Agnieszka Tymula. 2013. “Political quid pro quo agreements: An experimental study.” *American Journal of Political Science* 57 (3): 582–597.

Hansen, Morris H, and William N Hurwitz. 1946. "The problem of non-response in sample surveys." *Journal of the American Statistical Association* 41 (236): 517–529.

Healy, Andrew, and Gabriel S Lenz. 2014. "Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy." *American Journal of Political Science* 58 (1): 31–47.

Huber, Gregory A., and John S. Lapinski. 2006. "The 'race card' revisited: Assessing racial priming in policy contests." *American Journal of Political Science* 50 (2): 421–440.

Huber, Martin. 2012. "Identification of average treatment effects in social experiments under alternative forms of attrition." *Journal of Educational and Behavioral Statistics* 37 (3): 443–474.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological Methods* 15 (4): 309–334.

Johns, Robert, and Graeme A.M. Davies. 2012. "Democratic peace or clash of civilizations? Target states and support for war in Britain and the United States." *The Journal of Politics* 74 (04): 1038–1052.

King, Gary, and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14 (2): 131–159.

Kosloff, Spee, Jeff Greenberg, Toni Schmader, Mark Dechesne, and David Weise. 2010. "Smearing the opposition: Implicit and explicit stigmatization of the 2008 US Presidential candidates and the current US President." *Journal of Experimental Psychology: General* 139 (3): 383–398.

Leach, Colin Wayne, Patricia M. Rodriguez Mosquera, Michael L.W. Vliek, and Emily Hirt. 2010. "Group devaluation and group identification." *Journal of Social Issues* 66 (3): 535–552.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7 (1): 295–318.

Malesky, Edmund, Paul Schuler, and Anh Tran. 2012. "The adverse effects of sunshine: a field experiment on legislative transparency in an authoritarian assembly." *American Political Science Review* 106 (04): 762–786.

Manski, Charles F. 1989. "Anatomy of the selection problem." *Journal of Human Resources* 24 (3): 343–360.

Oppenheimer, Daniel M, Tom Meyvis, and Nicolas Davidenko. 2009. "Instructional manipulation checks: Detecting satisficing to increase statistical power." *Journal of Experimental Social Psychology* 45 (4): 867–872.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference.* 2nd edition ed. Cambridge University Press.

Robins, James M. 1999. "Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models.".

Robins, James M, Miguel Angel Hernan, and Babette Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology* 11 (5): 550–560.

Rosenbaum, Paul R. 1984. "The consequences of adjustment for a concomitant variable that has been affected by the treatment.".

Transue, John E., Daniel J. Lee, and John H. Aldrich. 2009. "Treatment spillover effects across survey experiments." *Political Analysis* 17 (2): 143–161.

Utych, Stephen M, and Cindy D Kam. 2014. "Viability, Information Seeking, and Vote Choice." *Journal of Politics* 76 (1): 152–166.

Valentino, Nicholas A., Vincent L. Hutchings, and Ismail K. White. 2002. "Cues that matter: How political ads prime racial attitudes during campaigns." *American Political Science Review* 96 (1): 75–90.

Weiner, Marc D. 2015. "A Natural Experiment: Inadvertent Priming of Party Identification in a Split-Sample Survey." *Survey Practice* 8 (6).

Wooldridge, Jeffrey M. 2005. "Violating ignorability of treatment by controlling for too many factors." *Econometric Theory* 21 (5): 1026–1028.

Zhou, Haotian, and Ayelet Fishbach. 2016. "The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions." *Journal of Personality and Social Psychology* 111 (4): 493–504.

ONLINE APPENDIX

Coding past experiments for post-treatment bias

Classifying large and complex research projects is not always straightforward. We engaged in extensive discussions of many articles before making a determination about whether they included some form of post-treatment conditioning. We examined only statistical results either presented in the main text of the articles or supplemental analyses that were directly referenced in the main text.[1] When it was not possible to determine whether or not variables were measured before or after the treatment from the manuscript or appendices, we contacted the authors to learn more about the study design. In the end, we coded articles as having engaged in post-treatment conditioning if the article met *any* of the following criteria (although several articles more than one).

- Articles that control for a variable that the authors themselves show is post-treatment using a statistical model or graph;
- Articles that controlled for variables that were (a) measured after the treatment and (b) could have plausibly been affected by the treatment;[2]
- Articles that dropped cases due to a failed manipulation check or non-compliance with treatment assignment;
- Articles that drop subjects based on attention filters measured post-treatment or conduct subset analyses based on scores on post-treatment attention filters (Oppenheimer, Meyvis, and Davidenko 2009; Berinsky, Margolis, and Sances 2014);

[1] In some cases, the concerns we identify may therefore apply to robustness checks either presented or described in the main text rather than the primary experimental results.

[2] In most cases, these were unambiguous. For instance, in an experiment exposing subjects to information about named candidates' position on the death penalty, Clifford (2014, 705) controls for death penalty attitudes measured post-treatment. In a few cases, this decision is more ambiguous. Bolsen, Ferraro, and Miranda (2014), for instance, control for voting in post-treatment elections where the treatment was a persuasion message about water conservation. However, we ignored instances where researchers controlled for post-treatment variables that were clearly orthogonal to the treatments (e.g., gender or race when these measures were not directly relevant to the study).

- Articles implementing the Baron and Kenny (1986) approach to mediation analysis who present these results as the primary justification for their conclusions;
- Articles where subsamples of observations are analyzed that were selected based on one or more variables that the authors show are post-treatment (see first bullet above);
- Articles where subsamples of observations are analyzed that were selected based on one or more variables that were (a) measured after the treatment and (b) could have plausibly been affected by the treatment;[3]
- Articles that suffered from post-treatment attrition (n=6).

In a handful of unusual cases, we identified issues that, though technically wrong, are unlikely to change the reported results. For instance, the authors of one article dropped two cases because of perfect separation in the (post-treatment) outcome (Utych and Kam 2014).[4] In another case, seven subjects (out of 248) were dropped for failing to follow instructions (Banks and Valentino 2012). The experimental findings in these studies are unlikely to be strongly affected by post-treatment bias given the small number of cases affected. Still, scholars should employ analytical procedures that preserve the value of random assignment and avoid biasing their estimates in any way. The fact that scholars regularly engage in practices these practices despite the danger of biasing estimates indicates that the problem of post-treatment bias is still not widely recognized.

[3] For instance, Johns and Davies (2012) exclude respondents from a study that primed religious group identities based on a religious affiliation variable that was measured after the experimental manipulation. In a small number of cases, the authors tested these problematic variables to assess balance across treatment groups and found no treatment effect (e.g., Dunning and Nilekani 2013). We determined that these cases met our definition of inducing potential post-treatment bias, although the resulting bias is likely small.

[4] The authors technically control for a pre-treatment variable (race) in a model of Republican behavior, but in doing so create separation in their statistical model because both black Republicans in their sample gave the same response. Unfortunately, omitting these observations due to their outcome values is equivalent to post-treatment selection.

Simulation evidence of the consequences of post-treatment bias

In these simulations, we slightly alter the assumptions used in the main text by adding independent error terms to Equation (5) and generate data using the following model:

$$\begin{aligned} Y_i &= \kappa_Y u_i + \tilde{\epsilon}_{Y,i} \\ X_i &= \gamma T_i + \kappa_X u_i + \tilde{\epsilon}_{X,i} \end{aligned} \tag{1}$$

where $u_i \sim \mathcal{N}(0, \sigma_u^2)$, $\tilde{\epsilon}_{Y,i} \sim \mathcal{N}(0,1)$, $\tilde{\epsilon}_{X,i} \sim \mathcal{N}(0,1)$, and $u \perp\!\!\!\perp \tilde{\epsilon}_X \perp\!\!\!\perp \tilde{\epsilon}_Y$. In all of our examples below, we assume n=2,000 divided equally between the treatment and control conditions.

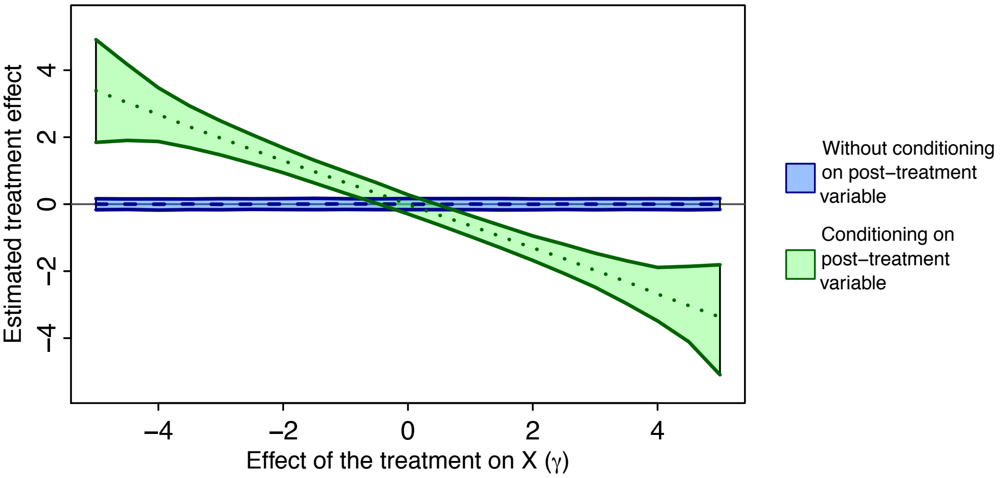
Simulations: Dropping cases based on post-treatment criteria

We first examine the consequences of dropping observations based on post-treatment criteria. Specifically, we simulate a scenario in which 15% of respondents are removed based based on the observed value of X_i (specifically, the 15% with the highest value of X_i). That is, we simulate data according to Equation (1) for different parameter values of γ and κ_Y. For each unique combination of parameter values, we simulate 2,000 samples and estimate a regression in which no observations are dropped and one in which 15% of observations are dropped. Our focus in these figures is on the 90% Monte Carlo intervals for the estimated treatment effect given each unique combination of parameters.

Our first set of simulations focuses on the effect of changing the γ parameter, which represents the effect of the treatment on the covariate X.[5] The blue shaded region in Figure A1 shows the 90% Monte Carlo interval for the estimate of treatment effect, which is centered at the true value of zero for all parameter settings. The green shaded region shows the same interval where 15% of observations have instead been dropped based on the values taken by X. In this case, the estimated treatment effect can be severely biased in either direction depending on the value of γ and only recovers the true treatment effect ($\tau = 0$) when $\gamma = 0$ as explained in previous sections (see Eq. 8).

[5] We fix $\kappa_X = 1$, κ_Y=1, and $\sigma_u = 2$.

Figure A1: Post-treatment bias when dropping cases as a function of treatment effect on X



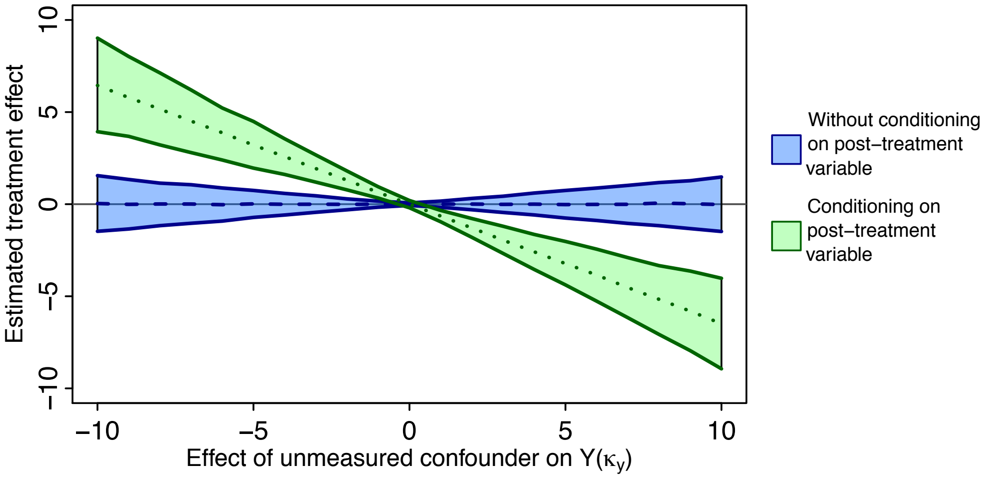
The plot shows the 90% Monte Carlo confidence intervals for point estimates of the treatment effect when 15% of the sample is dropped based on post-treatment criteria. Data were generated according to Model 1 for differing values of γ where $\kappa_X = 1$, κ_Y=1, and $\sigma_u = 2$. We fit 2,000 regressions for each parameter combination.

Our second set of simulations follows the same basic procedure, but now focuses on the effect of the unmeasured confounder on the outcome (κ_Y). The blue region in Figure A2 shows again that the point estimates are generally unbiased when the full sample is used. That is, for all values of κ_Y, the point estimates are centered at the true value of τ. However, when cases are dropped based on post-treatment criteria, the estimated treatment effects indicated by the green shaded region can be positive or negative depending on the specific value of κ_Y. This result is particularly disturbing because researchers cannot feasibly estimate κ_Y, which represents the effect of an unmeasured confounder u on Y.

Simulations: Directly controlling for post-treatment variables

We next turn to the related issue of directly controlling for post-treatment variables in a statistical model. Figure A3 shows the estimated treatment effect as a function of the effect of the treatment T on X (γ) with and without controls for the post-treatment variable X. This plot again demonstrates that point estimates of treatment effects are generally accurate without conditioning on a post-treatment variable (blue shaded region), but can be of almost

Figure A2: Post-treatment bias when dropping cases as a function of unobserved confounding



The plot shows the 90% Monte Carlo confidence intervals for point estimates of the treatment effect when 15% of the sample is dropped based on post-treatment criteria. Data were generated according to Model 1 for differing values of κ_Y where $\gamma = 1$, κ_Y=1, and $\sigma_u = 2$. We fit 2,000 regressions for each parameter combination.

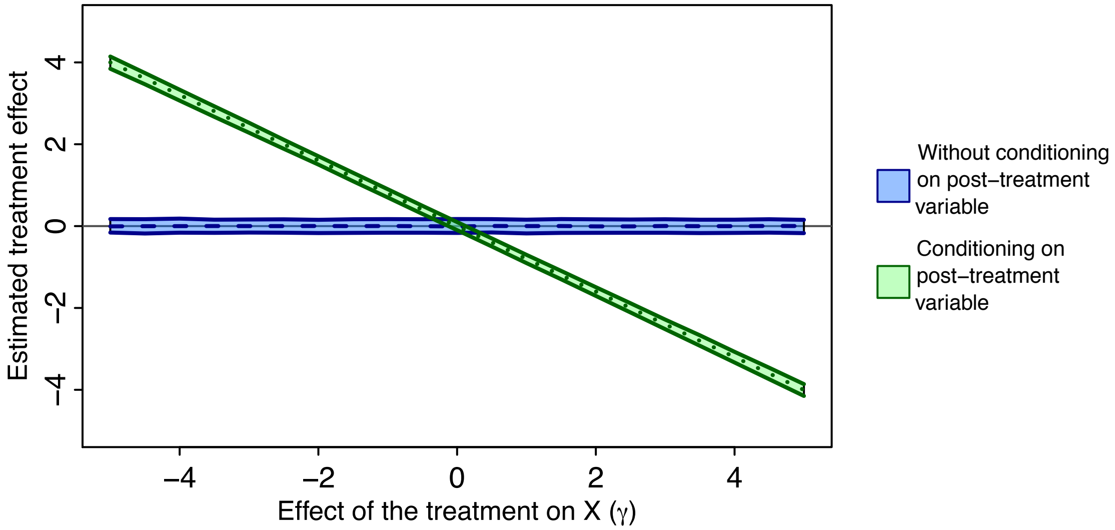
any size and in any direction when this conditioning occurs (green shaded region).

While it might be possible to estimate plausible values for the effect of the treatment on a measured covariate (γ), the same cannot be said for the effect of an unmeasured confounder on the outcome (κ_Y). As Figure A4 shows, a model that omits a post-treatment control generally yields unbiased estimates, but controlling for a post-treatment covariate can again induce severe bias of almost any size or direction depending on the values of this parameter.

Experimental balance in Dickson, Gordon, and Huber (2015)

The bold lines in each panel of Figure A5 correspond to the groups in Dickson, Gordon, and Huber (2015) that were assigned to the full information condition, while the dotted lines represent the limited information condition. These horizontal lines show the effect of being assigned to the appropriations treatment on the respective post-treatment variable for both the full and limited information groups (relative to the salary condition). Similarly, the vertical comparisons in the panels represent the effect of being in the full information condition, represented as an open circle, relative to the limited information condition (solid

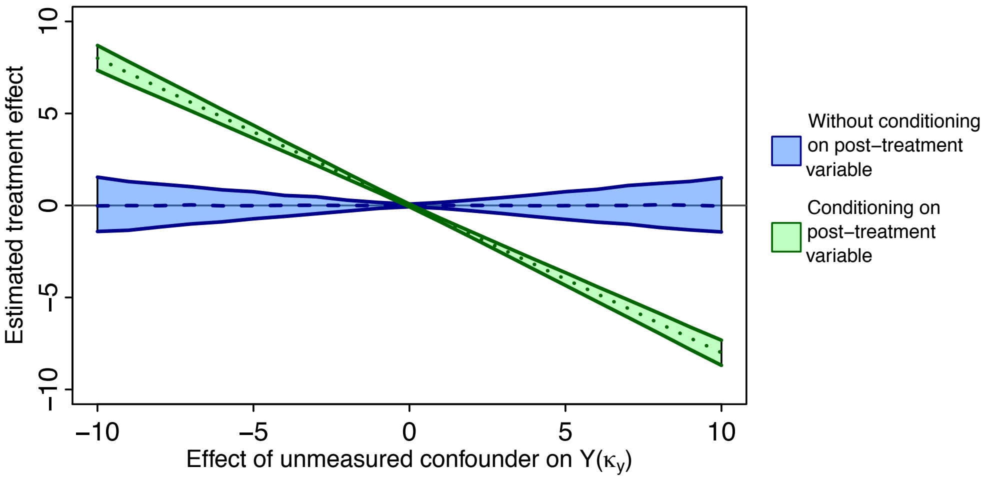
Figure A3: Post-treatment bias controlling for X as a function of the treatment's effect on X



The plot shows the 90% Monte Carlo confidence intervals for point estimates of the treatment effect when a post-treatment covariate is a included as a control variable in linear regression. Data were generated according to Model 1 for differing values of γ where $\kappa_X = 1$, κ_Y=1, and $\sigma_u = 2$. We fit 2,000 regressions for each parameter combination.

circle), for both the salary and appropriation groups. The top left panel of Figure A5 indicates that there is no significant difference in the prevalence of contributor targeting by treatment group (though again such a finding does not rule out post-treatment bias, as we show below). However, the other panels of the figure show that the appropriations manipulation has a causal effect on each of the lagged behavioral measures among those with low levels of information. As such, controlling for these variables could create post-treatment bias.

Figure A4: Post-treatment bias controlling for X as a function of unobserved confounding



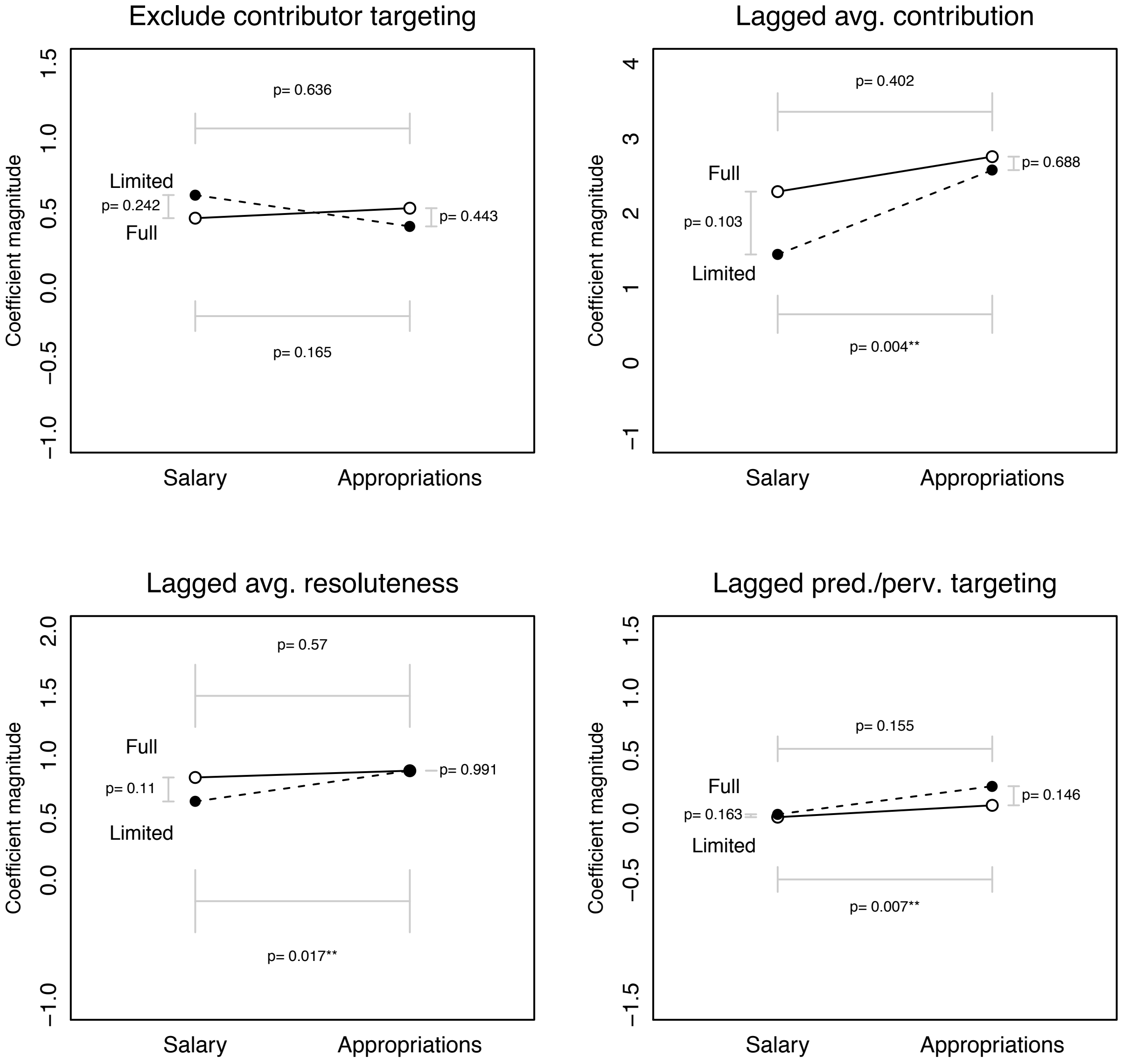
The plot shows the 90% Monte Carlo confidence intervals for point estimates of the treatment effect when post-treatment variables are included as "control" variables in linear regression. Data were generated according to Model 1 for differing values of κ_Y where $\gamma = 1$, κ_X=1, and $\sigma_u = 2$. We fit 2,000 regressions for each parameter combination.

Treatment effect differences in Dickson, Gordon, and Huber (2015)

The no-intercept models reported in Table A1 are the source of the treatment effect estimates in Table 2 in the main text. This approach differs from the subsample-based modeling strategy in the original study, which tests the effect of the appropriations treatment in the full information condition in Table 2 and the limited information condition in Table 4. Likewise, Table 4 of the original study estimates the effect of the limited information condition separately for the salary and appropriation conditions.

Figure A6 illustrates the treatment effect differences induced by post-treatment conditioning visually. The baseline results are shown in the top left panel (full sample, no post-treatment conditioning), the top right panel shows results when post-treatment covariates are included as controls, the bottom left panel shows results when we drop cases in which contributors were targeted, and the bottom right panel shows results using both practices. As in Figure A5, the graph allows for comparisons between conditions in the 2×2 design. In the figure, we highlight the differences in outcome means between treatment

Figure A5: Effect of treatments on four post-treatment variables



Data from Dickson, Gordon, and Huber (2015). Gray bars represent differences of means by experimental condition holding the other manipulation fixed. See text and the original study for further details.

Table A1: Predicting propensity to hinder or assist authorities (no intercept)

	Full sample	Lagged controls	Drop cases	Drop/controls
	(1)	(2)	(3)	(4)
Salary/full information	1.133^{***}	0.313	1.234^{***}	−0.009
	(0.349)	(0.536)	(0.327)	(0.457)
Appropriations/full information	0.078	−0.740	0.577	−0.799
	(0.376)	(0.527)	(0.356)	(0.522)
Appropriations/limited information	0.190	−0.400	−0.296	-1.643^{***}
	(0.420)	(0.790)	(0.586)	(0.542)
Salary/limited information	0.558	−0.217	0.493	−0.728
	(0.381)	(0.592)	(0.356)	(0.578)
Lagged avg. group contributions		0.276^{***}		0.405^{***}
		(0.106)		(0.101)
Lagged average resoluteness		0.373		0.439
		(0.368)		(0.463)
Lagged predatory/perverse targeting		-1.767^{**}		
		(0.851)		
Period indicators	Yes	Yes	Yes	Yes
R^2	0.141	0.222	0.246	0.339
N	457	432	309	286

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

combinations and include both the *p*-value for each difference and stars to indicate whether those differences are significant at conventional levels.

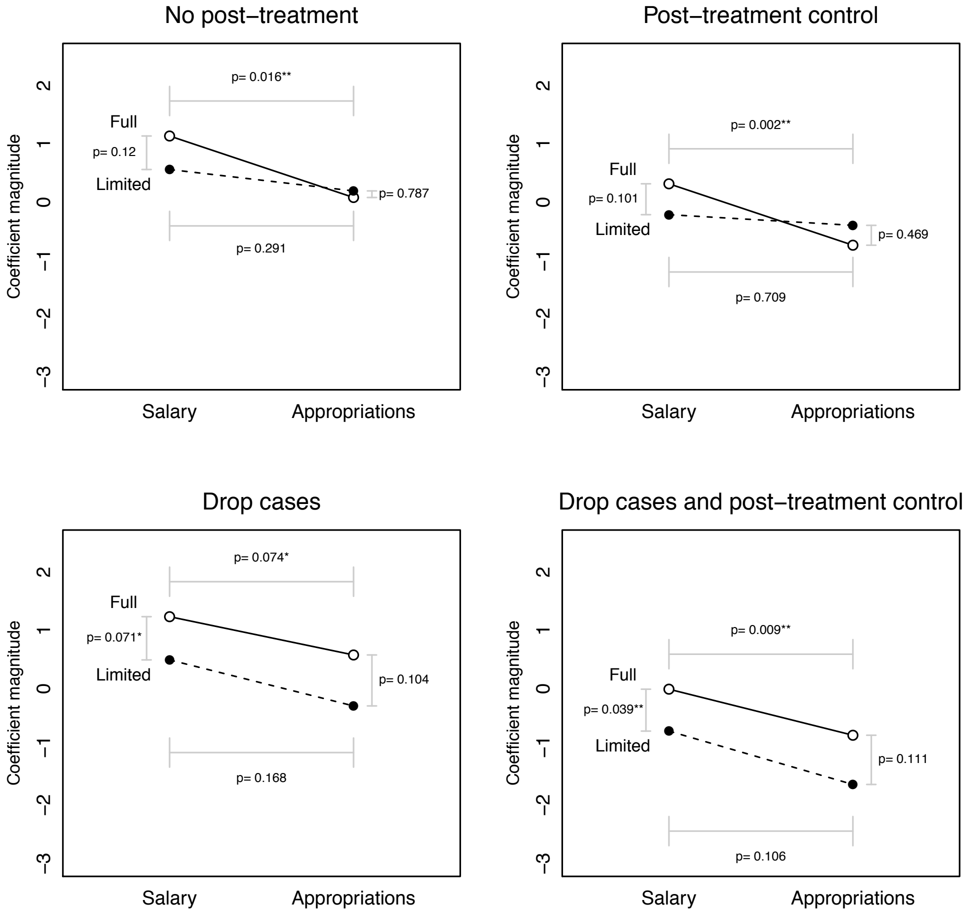
We also reported an equivalent pooled analysis below of the experimental results from Dickson, Gordon, and Huber (2015). Rather than omitting the intercept as in Table A1 above, we instead include indicators for the appropriations/limited information, salary/limited information, and appropriations/full information conditions. The omitted category is thus the salary/full information condition.[6]

Model 1 in Table A2 shows the results for a model omitting any potentially post-treatment controls or sample restrictions.[7] In this specification, we observe that the the appropriations manipulation decreases people's willingness to assist the authority regardless of their information status (relative to the salary/full information condition). However, when we add the lagged behavioral measures used in the original study as control variables

[6] As we note above, these estimates differ slightly from those in Dickson, Gordon, and Huber (2015) due to variation in subsample period effect estimates in the original study (details available upon request).

[7] For the purposes of this analysis, we follow Dickson, Gordon, and Huber (2015) in dropping sessions in which no attempted enforcement took place because the outcome measure is not defined.

Figure A6: Differences in treatment effect estimates between models



*p < .1; **p < .05. Data from Dickson, Gordon, and Huber (2015). See Table 2 for corresponding model results.

in model 2, the appropriations/limited information condition is no longer distinguishable from zero. In model 3, we instead drop cases where targeting of contributors is present following the exclusion in the original study. When we restrict the sample in this manner, the estimated treatment effects change substantially relative to model 1 – the magnitude of the appropriations/full information coefficient decreases considerably; the coefficient for the appropriations/limited information condition increases dramatically; and the estimated effect of being in the salary/limited information condition becomes statistically significant. These

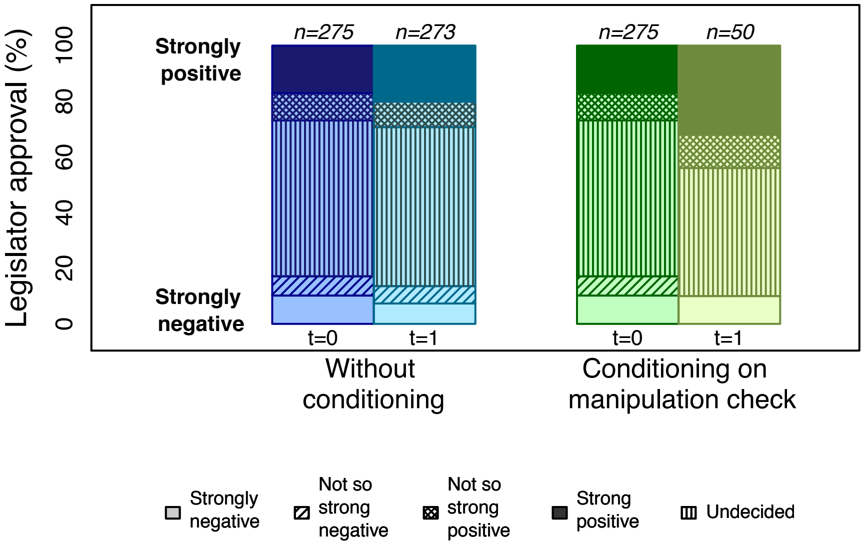
Table A2: Predicting the propensity to help or hinder authorities

	Full sample (1)	Lagged controls (2)	Drop cases (3)	Drop/controls (4)
Appropriations/full information	-1.055^{**}	-1.053^{***}	-0.657^{*}	-0.790^{***}
	(0.438)	(0.344)	(0.366)	(0.299)
Appropriations/limited information	-0.942^{**}	-0.713	-1.530^{**}	-1.634^{***}
	(0.380)	(0.502)	(0.593)	(0.561)
Salary/limited information	-0.575	-0.529	-0.742^{*}	-0.719^{**}
	(0.369)	(0.322)	(0.409)	(0.347)
Lagged avg. group contributions		0.276^{***}		0.405^{***}
		(0.106)		(0.101)
Lagged average resoluteness		0.373		0.439
		(0.368)		(0.463)
Lagged predatory/perverse targeting		-1.767^{**}		
		(0.851)		
(Intercept)	1.133^{***}	0.313	1.234^{***}	-0.009
	(0.349)	(0.536)	(0.327)	(0.457)
Period indicators	Yes	Yes	Yes	Yes
R^2	0.135	0.219	0.195	0.298
N	457	432	309	286

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$. Data from Dickson, Gordon, and Huber (2015). Models 3 and 4 ("Drop cases") exclude groups with any targeting of contributors as in Table 2 of the original study.

effects are striking given that the exclusion condition does not seem to be correlated with the treatment (see Figure A5). Finally, model 4 removes cases and controls for post-treatment variables similar to column 4 of Table 2 in Dickson, Gordon, and Huber (2015). In this case, all treatment combinations are highly significant.

Figure A7: How conditioning on a post-treatment variable can cause covariate imbalance



The bars on the left represent the distribution of legislator approval between the treatment and control conditions without dropping observations based on the manipulation check. The bars on the left represent the distribution if we drop observations in the treatment condition that failed the manipulation check.

Experimental imbalance from conditioning on manipulation checks

Figure A7 shows the distribution of prior legislator approval among the treatment and control groups in the Broockman and Butler (2015). The two bars on the left show the distribution of approval (ranging from strongly negative to strongly positive) among the control and treatment groups in the full sample, which is well-balanced. However, dropping observations in the treatment group that failed the manipulation check induces significant imbalance in prior approval, which can be seen in the two bars on the right.

Table A3: The effects of post-treatment bias: Legislator agreement models

			Drop if fail manipulation check	
	Original	Covariate	Treatment/control	Treatment only
Basic justification (treatment)	0.036	0.018	0.139*	0.263**
	(0.036)	(0.035)	(0.076)	(0.064)
Extensive justification (treatment)	0.044	0.020	0.001	0.117*
	(0.040)	(0.040)	(0.077)	(0.065)
Lagged opinion	-0.038	-0.038	0.007	-0.049
	(0.046)	(0.046)	(0.101)	(0.051)
Correctly identified position		0.171**		
		(0.033)		
Constant	0.355**	0.327**	0.621**	0.433**
	(0.049)	(0.047)	(0.098)	(0.064)
Basic − extensive justification	-0.008	-0.003	0.137	0.146*
	(0.047)	(0.046)	(0.088)	(0.087)
Dummy variables for strata	Yes	Yes	Yes	Yes
R^2	0.04	0.06	0.15	0.08
N	1076	1076	278	804

* $p < .10$, ** $p < .05$. OLS regression results with robust standard errors clustered by voter.

Conditioning on manipulation checks: Another reanalysis of Broockman and Butler (2015)

We also conduct a reanalysis of Broockman and Butler's second study, which compares agreement with a legislator's position between voters who received a content-free "control letter" and those sent one with either a basic or an extensive justification.[8] This study also included a manipulation check in which a random subset of respondents were asked to identify the position of the legislator. Once again, Broockman and Butler (2015) did *not* condition on correct answers to this question, which was considered in a separate analysis. However, we use it check to demonstrate the pernicious consequences of post-treatment conditioning.

The analysis in Table A3 is restricted to the subset of respondents who were asked about the legislator's position. As in Table 3, we present results from four models: the original model in Broockman and Butler (2015) (first column), a model that controls for whether respondents could correctly identify the legislator's position (second column), and

[8] Broockman and Butler (2015, 9) consider three measures of legislator agreement. As in Table 3, we again focus on the binary agreement measure here for ease of exposition.

models that drop respondents who could not correctly identify the legislator's position from both conditions or the treatment condition only (the third and fourth columns, respectively). Again, conditioning on a post-treatment variable creates substantively important differences in the conclusions we draw. The original model estimates null effects for both treatment variables in this subsample.[9] However, dropping respondents who fail the manipulation check in both conditions makes the basic justification treatment positive and statistically significant ($p < .10$; third column). Even worse, both treatments become statistically significant when we drop respondents who fail the manipulation check from the treatment condition only (basic $p < .05$, extensive $p < .10$; fourth column). In this model, we can also reject the null of no difference between treatments ($p < .10$), which was a relatively precise zero in the original model (see Table A3, which reports this auxiliary quantity in the sixth row).

[9] As reported in the original article (Broockman and Butler 2015, 9), the basic justification treatment effect is statistically significant at the $p < .05$ level in the full sample for two of the three models and at the $p < .10$ level for the binary agreement measure we use.