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
Campus Box 1063
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

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, eliminating observations based on post-treatment criteria, or subsetting the data based on post-treatment variables. 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. The data, code, and any additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science Dataverse* within the Harvard Dataverse Network at http://dx.doi.org/10.7910/DVN/EZSJ1S. 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, Molly Roberts, and three anonymous reviewers for helpful comments. All errors are our own.

1. INTRODUCTION

Political scientists increasingly rely on experimental studies because they allow researchers to obtain unbiased estimates of causal effects without identifying and measuring all confounders or engaging in complex statistical modeling. Under randomization, the difference between the average outcome of observations that received a treatment and the average outcome of those who did not is an unbiased estimate of the causal effect. Experiments are therefore a powerful tool for testing theories and evaluating causal claims while ameliorating concerns about omitted variable bias and endogeneity. For many, randomized controlled studies represent the gold standard of social science research.

Of course, this description of experiments is idealized. In the real world, things get messy. Some participants ignore stimuli or fail to receive their assigned treatment. Researchers may wish to understand the mechanism that produced an experimental effect or to rule out alternative explanations. Experimental practitioners are all too familiar with these and many other challenges in designing studies and analyzing results.

Unfortunately, researchers who wish to address these problems often resort to common practices including dropping participants who fail manipulation checks; controlling for variables measured after the treatment such as potential mediators; or subsetting samples based on post-treatment variables. Many applied scholars seem unaware that these common practices amount to conditioning on post-treatment variables and can bias estimates of causal effects. Further, this bias can be in any direction, it can be of any size, and there is often no way to provide finite bounds or eliminate it absent strong assumptions that are unlikely to hold in real-world settings. In short, conditioning on post-treatment variables can ruin experiments; we should not do it.

Though the dangers of post-treatment bias have long been recognized in the fields of statistics, econometrics, and political methodology (e.g., Rosenbaum 1984; Wooldridge 2005; King and Zeng 2006; Elwert and Winship 2014; Acharya, Blackwell, and Sen 2016), there is still significant confusion in the wider discipline about its sources and consequences. In this article, we therefore seek to 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 common practices that lead to post-treatment conditioning and document their prevalence in articles published in the field's top journals. We then provide analytical results that explain how post-treatment bias contaminates experimental analyses and demonstrate how it can distort treatment effect estimates using data from two real-world studies. We conclude by offering guidance on how to address practical challenges in experimental research without inducing post-treatment bias.

2. DON'T WE ALREADY KNOW THIS?

We first address the notion that the dangers of post-treatment bias are already well known. After all, published research in political science identified post-treatment bias (in passing) as problematic a decade ago (King and Zeng 2006, 147–148). 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 are either not understood or are being ignored — our review of the published literature suggests that it is widespread.

Of course, conditioning on post-treatment variables is not a practice that is exclusive to experimental research. Indeed, we believe the prevalence of and bias from post-treatment conditioning in observational research is likely greater (perhaps, much greater). Acharya, Blackwell, and Sen (2016), for instance, show that as many as four out of five observational studies in top journals may condition on post-treatment variables. We speculate that post-treatment bias may be even more common in less prestigious outlets or in books.

We focus on experiments because, first, it is reasonable to expect experimentalists to be *especially* careful to avoid post-treatment bias. 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 artificial). And unlike observational studies, the nature and timing of the treatment in experiments is typically unambiguous, making it easy for scholars to avoid conditioning on post-treatment variables. Second, for pedagogical purposes, explaining post-treatment bias in experiments allows for greater expositional clarity, reduces ambiguity about whether variables are measured post-treatment in the examples we discuss, and allows us to generate an unbiased estimate for purposes of comparison in our applications.

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 ($n = 75$). We coded each article for whether the authors subsetted the data 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 experimental outcomes.[1]

Table 1 presents a summary of our results. Overall, we find that 46.7% of the experimental studies published in APSR, AJPS, and JOP from 2012 to 2014 engaged in post-treatment conditioning (35 of 75 studies). Specifically, more than one in three studies engaged in at least one of two problematic practices — 21.3% (16 of 75) controlled for a post-treatment covariate in a statistical model and 14.7% of studies subsetted the data based on potential post-treatment criteria (11 of 75 studies reviewed) — and almost one in ten engaged in both (10.7%, 8 studies). Among those studies that controlled for a post-treatment variable, six used a mediation technique (8%). Further, while some studies lost cases due to post-treatment attrition (8.0%), the others chose to subset their samples or drop cases based on failed manipulation checks, noncompliance, attention screeners, or other post-treatment variables. Most strikingly, 12% of studies conditioned on a variable shown to be affected by the experimental treatment in analyses contained within the article itself (9 of 75).

[1] Additional details on these coding procedures as well as a listing of articles coded as having some form of post-treatment conditioning are provided in the Online Appendix.

Table 1: Post-treatment conditioning in experimental studies

Category	Prevalence
Engages in post-treatment conditioning	46.7%
Controls for/interacts with a post-treatment variable	*21.3%*
Drops cases based on post-treatment criteria	*14.7%*
Both types of post-treatment conditioning present	*10.7%*
No conditioning on post-treatment variables	52.0%
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 ($n = 75$).

In short, nearly half of the experimental studies published in our discipline's most prestigious journals during this period raise concerns about post-treatment bias. About one in four drop cases or subset the data based on post-treatment criteria and nearly a third include post-treatment variables as covariates. Further, few acknowledge potential concerns regarding the bias that post-treatment conditioning can introduce. Most tellingly, *nearly one in eight articles directly condition on one or more variables that the authors themselves treat as outcomes*[2]—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 is clear: it does not.

3. THE INFERENTIAL PROBLEMS CREATED BY POST-TREATMENT BIAS

The pervasiveness of post-treatment conditioning in experimental political science 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 (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,

[2] The analyses in question are not necessarily the main results of interest; in some cases, prior dependent variables are treated as covariates in auxiliary analyses. The concerns we describe still apply, however.

the bias that results from conditioning on post-treatment covariates is discussed only briefly (Gelman and Hill 2006, Section 9.7; Angrist and Pischke 2014, pp. 214-17). Even when the subject is treated fully (e.g., Gerber and Green 2012), it is 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 analyses in political science. We refer readers to, e.g., Imbens and Angrist (1994), Aronow, Baron, and Pinson (2015), Athey and Imbens (2016), and the works cited therein for more technical discussions.

3.1. *The intuition of post-treatment bias*

The intuition behind post-treatment bias may be best understood within the context of an example. Consider a hypothetical randomized trial testing whether a civic education program increases voter turnout in a mixed income school. In this example, we would estimate the effect of the intervention by comparing the turnout rate among those assigned to receive the civic education treatment with those who were not. These two *groups* serve as counterfactuals for each other because each group will in expectation be *similar* in terms of other variables such as socioeconomic status (SES) due to random assignment.

Conditioning on post-treatment variables eliminates the advantages of randomization because we are now comparing *dissimilar* groups. Imagine, for instance, that we wish to control for political interest of the subjects (as measured after the treatment) so that we can understand the effect the civic training class independent of subjects' political awareness. In this example, we assume that political interest is binary—it is measured as either high or low. Once we condition on the political interest variable by subsetting the data on political interest or including it as a covariate in a regression, we are now comparing the turnout rate of individuals who had low political interest *despite receiving the civic engagement training* (Group A) with those who have low political interest in the absence of the class (Group B).[3]

[3] Similarly, we are comparing people who had a high level of interest after taking the class to those with a high level of interest *despite not taking the class.*

If the training program worked, these groups are *not* similar. The training will surely lead to higher levels of political interest among students with a predisposition to become activated (e.g., higher SES students). The *point* of the experiment was to nudge individuals who *might* be interested in politics to become more politically active. Treated/low-interest students (Group A) will therefore consist disproportionately of individuals whose pre-treatment characteristics make them least likely to participate under any circumstances—those with the lowest SES. Meanwhile, Group B will have relatively more individuals with moderate levels of political interest and engagement (and correspondingly higher levels of SES) since no effort was made to help them become politically engaged.

In this example, comparing dissimilar groups could lead us to falsely conclude that the treatment had a *negative* effect on turnout. The untreated/low-interest subjects (Group B) might vote at a higher rate than the treated/low-interest subjects (Group A) because these groups differ by SES, not because the civic education program decreased participation.[4]

As this example illustrates, concerns about post-treatment bias are not really (or only) about the post-treatment variable itself. The problem is that by conditioning on a post-treatment variable we have unbalanced the treatment and control groups with respect to *every other possible confounder*. In this example, our attempt to control for one variable (political interest) introduced bias from imbalance in another variable (SES) that was not even included in the model and which the researchers may not have even measured.

3.2. *Why experiments generate unbiased estimates of treatment effects*

To understand more formally 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, $y_{[i,T=0]}$.

[4] We thank an anonymous reviewer for suggesting this explanatory approach.

The estimand of interest is the average treatment effect (ATE), which we denote:

$$\begin{aligned} \text{ATE} \;=\; \tau \;&=\; \mathbb{E}(\mathbf{y}_{[T=1]} - \mathbf{y}_{[T=0]}) \\ &=\; \mathbb{E}(\mathbf{y}_{[T=1]}) - \mathbb{E}(\mathbf{y}_{[T=0]}) \end{aligned} \tag{1}$$

Of course, we cannot observe *both* potential outcomes for each individual. Thus, we define a new estimand, the *difference in conditional expected values* (DCEV). This is,

$$\text{DCEV} = \Delta = \mathbb{E}(\mathbf{y}|T=1, \mathbf{X}=\mathbf{X}^*) - \mathbb{E}(\mathbf{y}|T=0, \mathbf{X}=\mathbf{X}^*), \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_p']$ is an $n \times p$ matrix of covariates and $\mathbf{X}^*$ represents their realized values. We focus on the DCEV because $\Delta = \tau$ given certain assumptions (these estimands are equivalent) and we can construct an unbiased estimate of Δ from observed data. A standard approach is to difference the conditional mean outcome among individuals we *observed* to have received a treatment, $\bar{\mathbf{y}}^{Obs}_{[1,\mathbf{X}^*]} = \text{mean}(\mathbf{y}|T=1, \mathbf{X}=\mathbf{X}^*)$, and the conditional mean outcome among those we *observed* who did not, $\bar{\mathbf{y}}^{Obs}_{[0,\mathbf{X}^*]} = \text{mean}(\mathbf{y}|T=0, \mathbf{X}=\mathbf{X}^*)$ (King and Zeng 2006). We denote this quantity, the *difference in conditional means* (DCM), as:

$$\text{DCM} = \hat{\Delta} = \bar{\mathbf{y}}^{Obs}_{[1,\mathbf{X}^*]} - \bar{\mathbf{y}}^{Obs}_{[0,\mathbf{X}^*]} \tag{3}$$

This estimate, $\hat{\Delta}$, is what is produced using standard regression analyses of experiments.

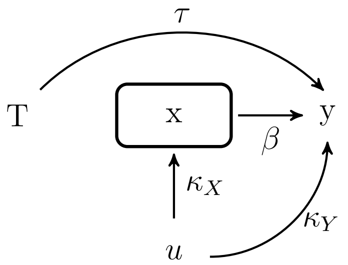
The reason that experiments work so well is that random assignment guarantees key assumptions[5] needed to ensure that $\Delta = \tau$, an equality which must hold to ensure that $\hat{\Delta}$ is an unbiased estimate of τ. Chief among these assumptions is

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

which states that treatment assignment is independent of potential outcomes conditional on

[5] Estimating a causal effect from an experiment requires several assumptions not discussed here. We focus on the assumption of interest for our purposes but see, e.g., Gerber and Green (2012).

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



covariates.

To see why this assumption is so critical, consider a graphical causal model where y is a linear function of a randomly assigned treatment T, a single covariate $x \in \{0, 1\}$, and unmeasured confounder u. Further, we assume that x is a pre-treatment co-variate, meaning that $T \perp\!\!\!\perp x$. Equation 4, which is shown visually in Figure 1[6], presents an example of a system of equations that meets these assumptions where c is a threshold constant and $\mathbb{1}(\cdot)$ is an indicator function.[7] Using our example above, y represents respondents' turnout decision, T represents the experimental civics education class, x represents respondents' *pre-treatment* political interest, and u represents the unmeasured confounder (SES).

$$\begin{aligned} y_i &= \alpha_Y + \tau T_i + \beta x_i + \kappa_Y u_i \\ x_i &= \mathbb{1}(\alpha_X + \kappa_X u_i > c), \end{aligned} \tag{4}$$

Substituting into Equation (2), we can show the following:

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

[6] Pearl (2009) shows 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.

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

Canceling terms, recalling that $\mathbb{E}(T|T=1,\mathbf{x}=\mathbf{x}^*)=1$ and $\mathbb{E}(T|T=0,\mathbf{x}=\mathbf{x}^*)=0$, and rearranging,[8] this can be expressed as:

$$\begin{aligned}\underbrace{\Delta}_{\text{DCEV}} = \underbrace{\tau}_{\text{ATE}} &+ \underbrace{\kappa_Y\Big(\mathbb{E}(\mathbf{u}|T=1,\mathbf{x}=\mathbf{x}^*)-\mathbb{E}(\mathbf{u}|T=0,\mathbf{x}=\mathbf{x}^*)\Big)}_{\text{Bias from imbalance in } u} \\ &+ \underbrace{\beta\Big(\mathbb{E}(\mathbf{x}|T=1,\mathbf{x}=\mathbf{x}^*)-\mathbb{E}(\mathbf{x}|T=0,\mathbf{x}=\mathbf{x}^*)\Big)}_{\text{Bias from imbalance in } x}\end{aligned} \tag{5}$$

Several aspects of Equation (5) are important. First, both of the terms on the right must be zero in expectation for Δ to be equivalent to τ—a necessary condition for $\hat{\Delta}$ to be an unbiased estimator of τ. In theory, that is precisely what experimental designs achieve. As long as we do not condition on a post-treatment variable, randomization guarantees that Assumption (1) is satisfied and both quantities go to zero. Assumption (1) implies that individuals in the treatment and control conditions will be similar in expectation with respect to unobserved confounders such as SES. In mathematical terms, $\mathbb{E}(\mathbf{u}|T=1,\mathbf{x}=\mathbf{x}^*)=\mathbb{E}(\mathbf{u}|T=0,\mathbf{x}=\mathbf{x}^*)$, which means that the expected bias from a lack of balance in SES is zero. Further, Assumption (1) requires that x is not causally related to T—i.e., that respondents' level of pre-treatment political interest is not a function of treatment assignment. Thus, $\mathbb{E}(\mathbf{x}|T=1,\mathbf{x}=\mathbf{x}^*)=\mathbb{E}(\mathbf{x}|T=0,\mathbf{x}=\mathbf{x}^*)$, which means that the second term is also exactly zero in expectation. More generally, data generated as shown in Figure 1 will satisfy Assumption (1). Any method that generates an unbiased estimate of the DCEV (Δ) will then also generate an unbiased estimate of the ATE (τ). For instance, a regression controlling for both the civic education treatment and prior political interest will, in expectation, provide the right estimate.

A second key feature of Equation (5) is that the bias resulting from imbalance in the observed or unobserved covariates can be *anything*. For any finite ATE, we can construct examples where the bias will be $-\infty$, ∞, or anything in between depending on the value of parameters like κ_Y (the effect of the unmeasured covariate on the outcome).

[8] Note that Model 4 also assumes that the main parameters in the model (τ, α_Y, β, τ, and κ_Y) do not vary as a function of T or $\mathbf{x}$, which is why we can move these parameters outside of the expectations. However, this simplifying assumption is not problematic for our argument. (Without it, the resulting bias will not evaporate or even necessarily decrease, but will instead simply be more difficult to characterize.)

Finally, while it might be plausible to estimate (and adjust for) the bias resulting from imbalance in x using the observed values in our data (e.g., political interest), Equation (5) shows if we violate Assumption (1) we would also need to somehow adjust for bias resulting from imbalances in the unobserved confounder u (e.g., SES). Adjusting for imbalance in unobservable variables is more challenging, requiring either the availability of exogenous instruments and/or stronger (and more limiting) assumptions such as no imbalance in unobservables conditional on observed covariates that are often implausible in practice.

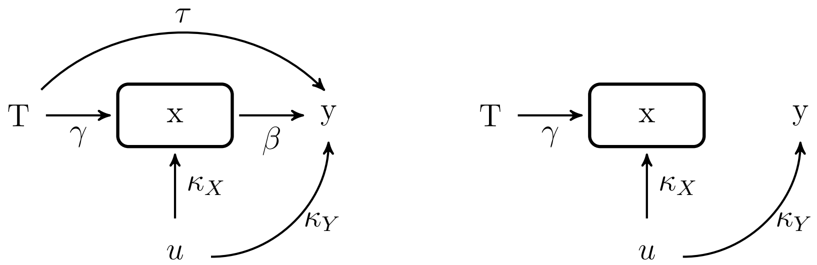
3.3. *The problem with conditioning on post-treatment variables*

We are now are ready to directly discuss post-treatment bias. In short, when we include a post-treatment variable in the set of conditioning variables either directly or indirectly, Assumption (1) is violated. As a result, $\tau \neq \Delta$ for the reasons discussed above. Standard estimates such as the difference in conditional means ($\hat{\Delta}$) will therefore be biased regardless of sample size, measurement precision, or estimation method.[9] Further, the bias of standard estimates such as $\hat{\Delta}$ can be in any direction and of any magnitude depending on the value of unknown (and unknowable) parameters (e.g., κ_Y, the effect of the unmeasured confounder on the outcome). Once we have conditioned on a post-treatment variable, we have eliminated the assurance of unconfoundedness provided by randomization.

To explain this point more clearly, we return to our example. We assume that the researcher estimates 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. This might occur, for instance, if we measured political interest after the civic education class was completed. As a result, the covariate (political interest) is now affected by the treatment and

[9] For expositional clarity, we omit edge cases that would allow us to condition on post-treatment confounders and generate an unbiased estimate of Δ. For instance, we assume that the influence of unmeasured confounders along the various causal paths will not somehow cancel out.

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



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

is thereby "post-treatment," meaning $\mathbb{E}(\mathbf{x}|T = 1) \neq \mathbb{E}(\mathbf{x}|T = 0)$. The assumed model is:

$$\begin{aligned} y_i &= \alpha_Y + \tau T + \beta x + \kappa_Y u_i \\ x_i &= \mathbb{1}(\alpha_X + \gamma T_i + \kappa_X u_i > c) \end{aligned} \tag{6}$$

Note that Equation 4 is identical to Equation 6 except that in the former we assumed that $\gamma = 0$ (no effect of the civic education class on political interest).

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 ($\beta = \tau = 0$). In our example, this assumption would mean that neither the civics class nor respondents' level of political interest affected turnout, but that the class did increase political interest ($\gamma \neq 0$). This situation, which is depicted in Figure 2b, can be written as:

$$\begin{aligned} y_i &= \alpha_Y + \kappa_Y u_i \\ x_i &= \mathbb{1}(\alpha_X + \gamma T_i + \kappa_X u_i > c) \end{aligned} \tag{7}$$

Note that Equation 7 is identical to Equation 6 except that in the former we assumed that $\beta = \tau = 0$ (no effect of either the intervention or the observed covariate on the outcome).

Under these circumstances it may seem harmless to condition on the post-treatment covariate x — after all, x has no effect on y.[10] This intuition is wrong. Even in such

[10] 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.

a favorable context, conditioning on x *still* 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 impossible without further assumptions (Elwert and Winship 2014).[11] In our example, conditioning on political interest unbalances the treatment and control groups on SES, which in turn causes our estimates of the causal effect of the civics class on turnout to be biased.

3.4. *Practices that lead to post-treatment bias*

Conceptually, there are two ways that researchers may condition on post-treatment variables: dropping (or subsetting) observations based on post-treatment criteria or controlling for post-treatment variables. We consider each below.

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. Sometimes conditioning on post-treatment variables is nearly unavoidable. The treatment itself may cause some respondents to be more likely to be omitted from the sample, a phenomenon which is usually termed non-random attrition. Zhou and Fishbach (2016) show that many online experiments experience significant differential attrition by experimental condition, which can also occur in field experiments (e.g., Horiuchi, Imai, and Taniguchi 2007). For instance, Malesky, Schuler, and Tran (2012) find that Vietnamese National Assembly delegates who were randomly selected to have websites built for them were less likely to be re-nominated (Table 7.1.1). As a result, analyses of the effect of this treatment on electoral outcomes inadvertently condition on a post-treatment variable (they are estimated only among legislators who were re-nominated). Similar problems can occur when analyzing the content of responses in audit experiments where some legislators do not reply (Coppock 2017).

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

In other instances, scholars intentionally condition on post-treatment variables. For instance, researchers frequently drop subjects who fail a post-treatment manipulation check or other measure of attention or compliance (including being suspicious of or guessing the purpose of a study). Healy and Lenz (2014, 37), for instance, exclude respondents who failed to correctly answer questions that were part of the treatment in a survey experiment. However, conditioning on these post-treatment measures can imbalance the sample with respect to observed or unobserved confounders. In particular, as Aronow, Baron, and Pinson (2015, 4) note, "the types of subjects who fail the manipulation check under one treatment may not be the same as those who fail under a different treatment" even if manipulation check passage rates are equal between conditions.

Finally, researchers may sometimes wish to estimate causal effects for different subsets of respondents but do not consider that the measure they use to define the subgroup was collected after the 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 affected by the treatment (see result 2 on page 589). Typically, this sort of intentional subsetting is driven by a desire to strengthen experimental findings. In our example, we might wish to estimate the effect of the civics education class only among low-interest students to show that the effect is not isolated to previously engaged students. Dropping respondents based on manipulation checks is often done to show that the estimated treatment effect is larger among compliers, which might appear to suggest that the treatment is working through the researchers' proposed mechanism. This reasoning is wrong. Selecting a portion of the data based on post-treatment criteria will not allow us to generate an unbiased estimate of the treatment effect within an interesting subset of respondents. Instead, we will obtain a *biased* estimate among an endogenously selected group.

Specifically, dropping cases or subsetting based on post-treatment criteria will unbalance the treatment and control conditions with respect to unmeasured confounders and bias our treatment effect estimates. For instance, consider data generated using Model (7) and assume we wish to analyze only low-interest observations ($x = 0$). Using Equation (2), we now have

$$\begin{aligned}
\Delta &= \mathbb{E}(\mathbf{y}|T = 1, \mathbf{x} = 0) - \mathbb{E}(\mathbf{y}|T = 0, \mathbf{x} = 0) \\
&= \mathbb{E}(\alpha_Y + \tau T + \beta\mathbf{x} + \kappa_Y\mathbf{u}|T = 1, \mathbf{x} = 0) \\
&\quad -\mathbb{E}(\alpha_Y + \tau T + \beta\mathbf{x} + \kappa_Y\mathbf{u}|T = 0, \mathbf{x} = 0) \\
&= \tau + \underbrace{\kappa_Y\big(\mathbb{E}(\mathbf{u}|T = 1, \mathbf{x} = 0) - \mathbb{E}(\mathbf{u}|T = 0, \mathbf{x} = 0)\big)}_{\text{Bias from imbalance in u when x=0}}.
\end{aligned} \tag{8}$$

Symmetrically, the bias when examining only high-interest subjects is.

$$\Delta = \tau + \underbrace{\kappa_Y\big(\mathbb{E}(\mathbf{u}|T = 1, \mathbf{x} = 1) - \mathbb{E}(\mathbf{u}|T = 0, \mathbf{x} = 1)\big)}_{\text{Bias from imbalance in u when x=1}}, \tag{9}$$

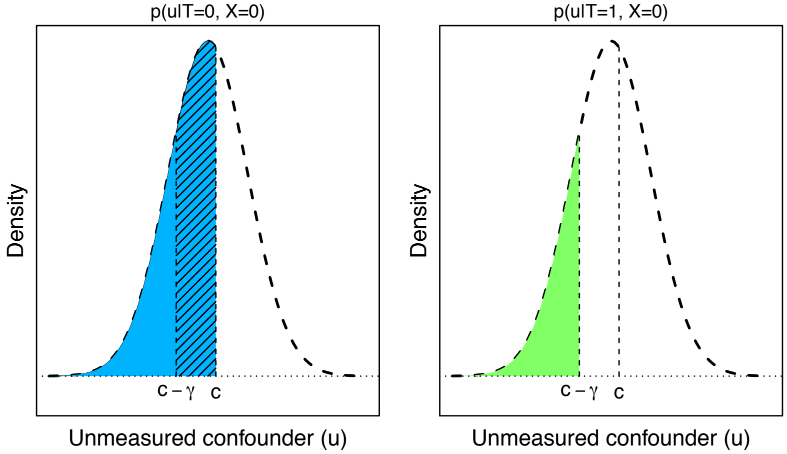
Although it is possible to construct examples where this bias is zero, it will not be zero in general. The reason is that the value of u must on average be lower for observations in the treatment group ($T = 1$) who also meet the selection criteria ($x = 0$) under the assumed data-generating process for x. In other words, units in the treatment group need lower values of u to stay below the threshold c. 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. In the context of our example, the low-interest subjects in the control group are being compared to respondents who maintained a low level of political interest *despite exposure to the civics education class*. In our simplified example, these are likely to be low SES students. This potential imbalance is illustrated in Figure 3, which shows an example of how the distribution of u will be imbalanced across treatment and control conditions when only selecting on low-interest ($x = 0$).[12]

Including post-treatment variables as covariates: A closely related practice is to control for one or more post-treatment covariates in a statistical model. In our example, this could occur if the post-treatment political interest variable were included as a covariate in a regression.

In some cases, well-intentioned scholars may engage in this practice in a mistaken

[12] In the Online Appendix, we provide exact calculations for the bias shown in this figure.

Figure 3: Example of 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=0) under the data-generating process in Equation 7. We assume $\alpha_x = 0$, $c > 0$, $\gamma > 0$, and that u is distributed normally.

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, analyze 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 and estimate a statistical model including an interaction term. For these models to be valid, the moderator x must *not* be affected by the experimental randomization. Spillover effects are possible even for strongly held attitudes like racial resentment after related interventions (e.g., Transue, Lee, and Aldrich 2009). Even variables that seem likely to remain fixed when measured after treatment such as measures of racial or partisan identification can be affected by treatments (e.g., Antman and Duncan 2015; Weiner 2015).

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 interacts it with 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 and 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 as a covariate — along with the treatment — in a model 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).

Regardless of the intention, including post-treatment variables as covariates for any of these reasons can bias estimates by creating imbalance with respect to the unmeasured confounder.[13] To see this more formally, we first need to define some quantities, which we will again illustrate in terms of our running example. Let $\Pr(\mathbf{x} = 1)$ be the marginal probability of being a high-interest student and $\Pr(\mathbf{x} = 0)$ be the marginal probability of being a low-interest student. Further, let $\mathbb{E}(\mathbf{u}|T = 0, \mathbf{x} = 0)$ and $\mathbb{E}(\mathbf{u}|T = 1, \mathbf{x} = 0)$ be the expected values of the unmeasured confounder (SES) for low-interest students in the control and treatment groups, respectively. These quantities would be, for instance, the expected value of the shaded areas in the left and right panels of Figure 3. Finally, $\mathbb{E}(\mathbf{u}|T = 0, \mathbf{x} = 1)$ and $\mathbb{E}(\mathbf{u}|T = 1, \mathbf{x} = 1)$ are the expected values of u for high-interest individuals.

We now want to calculate the DCEV when "controlling" for a post-treatment variable x, which is political interest in our example. Returning to Equation (2) and employing basic

[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 by assuming that $\beta = 0$. As shown in Equation 5, however, bias can also arise from imbalance in observed covariates when controlling for x $(\beta(\mathbb{E}(\mathbf{x}|T = 1) - \mathbb{E}(\mathbf{x}|T = 0)))$. While bias from imbalance in unobservables is even more problematic, it is also not possible to eliminate bias from imbalance in observables without additional assumptions (see, e.g., Baum et al. N.d.).

rules of probability, we get:

$$\begin{aligned}
\Delta &= \mathbb{E}(\mathbf{y}|T=1,\mathbf{x}=\mathbf{x}^*) - \mathbb{E}(Y|T=0,\mathbf{x}=\mathbf{x}^*) \\
&= \tau + \underbrace{\kappa_Y\big(\mathbb{E}(\mathbf{u}|T=1,\mathbf{x}=\mathbf{x}^*) - \mathbb{E}(\mathbf{u}|T=0,\mathbf{x}=\mathbf{x}^*)\big)}_{\text{Imbalance in u}} \\
&= \tau + \kappa_y\Big[\underbrace{\Pr(\mathbf{x}=0)}_{\text{Prob. low interest}}\ \underbrace{\big[\mathbb{E}(\mathbf{u}|T=1,\mathbf{x}=0) - \mathbb{E}(u|T=0,\mathbf{x}=0)\big]}_{\text{Imbalance when } x=0} \\
&\qquad + \underbrace{\Pr(\mathbf{x}=1)}_{\text{Prob. high interest}}\ \underbrace{\big[\mathbb{E}(\mathbf{u}|T=1,\mathbf{x}=1) - \mathbb{E}(\mathbf{u}|T=0,\mathbf{x}=1)\big]}_{\text{Imbalance when } x=1}\Big]
\end{aligned} \tag{10}$$

Note that this bias is simply a weighted combination of the exact same biases shown in Equations (8) and Equations (9) where the weights reflect the marginal probabilities of being either high or low interest students. Intuitively, this result shows that controlling for a post-treatment variable leads to a new bias that is simply a combination of the biased estimates we would get from selecting only cases where $x = 1$ 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 with standard methods.

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

4.1. *Analysis: An original study of judge perceptions*

We further demonstrate the pernicious effects of post-treatment bias with a simple experiment on cue-taking in judicial opinion conducted among 1,234 participants recruited from Amazon Mechanical Turk.[14] The study, which was conducted from April 24–25, 2017, builds on prior research investigating the effect of party and source cues on public opinion toward judges and courts (e.g., Burnett and Tiede 2015; Clark and Kastellec 2015). We specifically examine the effect of an implicit endorsement from President Trump on opinion toward a sitting state supreme court judge.

The study was conducted as follows.[15] After some initial demographic and attitudinal questions, each participant was shown a picture and a brief biography of Allison Eid, a justice on the Colorado Supreme Court. The treatment group was randomized to a version of the

[14] Like many Mechanical Turk samples, participants in the study skewed young (65% 18–34), male (58%), educated (53% hold a bachelor's degree or higher), and Democratic (59% including leaners).

[15] See the Online Appendix for the full instrument.

biography that included one additional fact: "Donald Trump named her as one of the 11 judges he might pick as a Supreme Court nominee." This information was not shown to the control group. After the experimental manipulation, respondents were asked how likely they were to retain Eid on the Colorado Supreme Court (for Colorado residents) or how likely they would be to do so if they lived in Colorado (for non-Colorado residents) on a four-point scale, which serves as our outcome variable. They were then also asked to evaluate her ideology on a seven-point scale from liberal (1) to conservative (7).

Model 1 in Table 2 reports the unconditional average treatment effect estimate of the endorsement on support for retaining Eid.[16] Given that participants disproportionately identify as Democrats, it is not surprising that Trump's endorsement reduced the likelihood of supporting Eid's retention by -0.214 ($p < .01$, 95% CI: -0.301, -0.127) on the four-point scale. This value is the treatment effect estimate of interest.

Table 2: Endorsement effect on retention vote conditioning on ideological distance

	Full sample		$\lvert\text{Distance}\rvert \leq 1$	$\lvert\text{Distance}\rvert > 1$
	(1)	(2)	(3)	(4)
Trump endorsement	-0.214*	-0.057	0.257*	-0.460*
	(0.044)	(0.041)	(0.063)	(0.054)
Ideological distance		-0.207*		
		(0.013)		
Constant	2.381*	2.724*	2.436*	2.319*
	(0.031)	(0.036)	(0.041)	(0.040)
N	1182	1178	504	674

*p < .01. Outcome variable is a four-point measure of the likelihood of voting to retain Eid. Ideological distance = |self-reported ideology - perception of Eid's ideology|.

Imagine, however, that a reviewer believes that the mechanism of the endorsement effect is Eid's perceived ideology rather than feelings about Trump. To try to account for this theory, the author could try to explore how the effect of the endorsement varies by perceived ideological distance to Eid. This distance is calculated as the absolute value of the difference between the respondents' self-placement on the 7-point ideology scale (measured pre-treatment) and the respondents' placement of Eid on the same scale (measured post-

[16] These results are estimated among the 1,182 respondents who answered the retention question. A total of 1,205 entered the manipulation. Attrition rates were 2.3% in control and 1.5% in treatment.

treatment). Unfortunately, because perceptions of Eid's ideology were measured after the manipulation, any analysis that conditions on ideological distance will be biased.

To illustrate this point, consider the other models in Table 2, which demonstrate just how severely post-treatment bias can distort treatment effect estimates. When we control for ideological distance to Eid in Model 2, for instance, the estimated treatment effect is no longer statistically significant (-0.057, 95% CI: -0.138, 0.024). While some might wish to interpret this coefficient as the direct effect of the Trump endorsement (controlling for perceived ideology), it is not. Instead, it is a biased estimate of the direct effect of the Trump endorsement, and the bias can be in any direction at all.

The bias becomes even worse if we condition on respondents who perceive themselves to be ideologically close to Eid ($\leq$ one point on the seven-point ideology scale) or not. The sign of the estimated treatment effect *reverses* in subsample of respondents who perceive themselves as being close to Eid, becoming positive (0.257, 95% CI: 0.132, 0.382), whereas the magnitude of the negative coefficient approximately doubles relative to the unconditional estimate among respondents who perceive themselves as further from Eid (-0.460; 95% CI: -0.567, -0.353). These effects are opposite in sign and in both cases highly significant ($p < .01$ in both directions).

However, all of these subsample estimates are also biased. As described in Section 3.4, conditioning on ideological distance actually unbalances the sample by respondents' self-reported ideology *even though self-reported ideology is measured pre-treatment.* For instance, among respondents who perceive themselves as ideologically close to Eid, treatment group respondents are significantly more conservative on our seven-point ideology scale than are control group respondents (4.633 versus 3.652, $p < .01$ in a t-test). The reason is that the treatment *increases* perceptions of Eid's conservatism (from 3.821 in the control group to 4.781 in the treatment group, $p < .01$ in a t-test). As a result, control group participants who think Eid is centrist on average and perceive themselves to be relatively close to her are being compared to treatment group participants who think she is close to them after finding out she was endorsed by President Trump. By conditioning on a post-treatment variable, we have unbalanced the treatment and control groups in terms of ideology and unmeasured confounders and thus biased the treatment effect estimate.

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

To further 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 manipulates rules and information to assess their effect on citizens' propensity to support or hinder authorities.

Participants were assigned to groups in which they were randomly assigned to be the authority or citizens. Each group played multiple sessions in which citizens first decide whether they wanted to contribute to a common pot, of which each citizen and the authority receive a share later. After observing contributions, the authority decides whether to target a citizen for enforcement for failing to contribute to the pot. If a member was penalized, 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 manipulated how authorities were compensated: fixed wage (*salary*) versus compensation based on penalties collected (*appropriations*). The other dimension, transparency, varied the amount of information citizens received about the actions of other players: knowing only that someone had been targeted but not knowing contributions (*limited information*) versus fully observing contributions and target selection (*full information*).

The study follows two common approaches in the literature on experimental economics and behavioral games that raise concerns about post-treatment bias.[17] First, DGH exclude cases of so-called "perverse" targeting of a contributor when at least one citizen did not contribute (119). Intuitively, dropping these cases might seem to allow them to focus on

[17] DGH is described as "experimental" in its title and invokes causal inference as a key rationale for its design: "because participants are randomly assigned to institutional environments, we are able to avoid selection problems and other obstacles to causal inference that complicate observational studies" (110).

treatment effects among individuals who correctly understood the incentives. However, perverse targeting is a post-treatment behavior given the expected effect of the manipulations. Second, DGH controls for lagged average contributions, average resoluteness, and perverse or predatory targeting to try to ensure that the effects of the treatments at time t are not fully mediated by behavior and outcomes in previous periods (122). Unfortunately, the lagged measures are themselves affected by the manipulations. As a result, both approaches provide biased treatment effect estimates that do not correspond to meaningful causal estimands.[18]

Table 3 demonstrates that post-treatment conditioning induces substantial differences in the estimated effects of DGH's treatments.[19] 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 on 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. In addition, we find that the magnitude of the effect estimates varies substantially when we condition on post-treatment variables. Most notably, the appropriations treatment effect estimate in the limited information condition more than doubles in magnitude and becomes nearly statistically significant in the fourth column ($p < .11$).

These findings offer new insight into the results in Dickson, Gordon, and Huber (2015). We replicate the appropriations treatment effect for full information groups, but our analysis raises concerns about post-treatment bias for both the limited information effect in the salary

[18] We show that these variables were affected by the treatment assignment in the Online Appendix.

[19] 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 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.

Table 3: 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.

condition and the appropriation effect in the limited information condition. Dickson, Gordon, and Huber (2015) notes that both models are sensitive to model specification; our analysis suggests that these results are attributable to post-treatment bias.[20]

5. RECOMMENDATIONS FOR PRACTICE

In this section, we provide recommendations to help researchers avoid the problems we describe above. The most important advice we have to offer is simple: do not condition on post-treatment variables. Do not control for them in regressions. Do not subset your data based on them. However, we recognize that following this guidance can be difficult. We therefore briefly summarize several motivations for post-treatment conditioning below — non-compliance, attrition, efficiency concerns, heterogeneous treatment effects, and mechanism questions — and explain how to address these issues without inducing bias using the most common and practical methods available.[21]

[20] See the Online Appendix for further analysis of Dickson, Gordon, and Huber (2015) and an additional demonstration of post-treatment bias using data from Broockman and Butler (2015).

[21] A full review of these literatures is beyond the scope of this article; see the cited works for more.

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

Researchers often wish to control for other variables in their analyses. Though it is not necessary to do so (randomization eliminates omitted variable bias in expectation), regression adjustment for covariates has been shown to induce only minor bias and to potentially increase efficiency under realistic conditions (e.g., Lin 2013). Including control variables is therefore potentially appropriate, but *only* covariates that are unrelated to the treatment and preferably measured in advance (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, this design risks post-treatment bias if the moderator could be affected by the experimental manipulation. Moderators that are vulnerable to treatment spillovers like racial resentment should be measured pre-treatment (see, e.g., Huber and Lapinski 2006, 424).[22]

Third, scholars often wish to use measures of respondent attention (separate from 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 differentially affect the types of participants who pass these measures via other mechanisms (e.g., changing respondent engagement or affecting the contents of working memory), thereby imbalancing the sample. In this scenario, dropping respondents based on post-treatment attention checks is the equivalent of selecting on a post-treatment covariate and would again risk bias.

5.2. *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 experimental

[22] Measuring moderators before a manipulation does raise concerns about priming. We acknowledge this possibility and discuss the need for further research on the topic in the conclusion.

rules, or inattentiveness. In other cases, scholars use an encouragement design or otherwise try to induce exogenous variation in a treatment 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 treatment group members do not.

There are no easy solutions to this problem. For the reasons stated above, simply dropping cases or controlling for compliance status in a regression model can lead to biased estimates of the ATE. Two possible solutions are fairly easy to implement but both require researchers to focus on different causal estimands. The simplest is to calculate the difference in outcomes between respondents *assigned* to receive treatment and those *assigned* to receive the control, which is an unbiased estimate of the intention to treat (ITT) effect. Although simple to execute (just ignore compliance status), this estimand may not correspond well with the underlying research question.

Another approach to noncompliance is to estimate a two-stage least squares model using random assignment as an instrument for treatment status. Here again, however, we are estimating a different estimand known as the complier average causal effect (CACE). While perfectly valid, interpretation can be difficult since the estimand represents the treatment effect for a subset of compliers. Interpretation is especially thorny in the presence of two-sided non-compliance where compliance status cannot be directly observed and an additional monotonicity assumption (no defiers) must be invoked (Angrist, Imbens, and Rubin 1996; see Gerber and Green 2012, 131–209 for more on these points).

5.3. *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 better estimate treatment effects in the presence of non-random attrition. If we are willing to assume that missingness is not a function of unmeasured confounders, we can use familiar methods such as imputation or marginal structural models. Under more realistic assumptions, however, the choices are more limited: Gerber and Green (2012) recommend extreme value bounds (Manski 1989), where analysts estimate the largest and smallest ATEs possible if missing information were filled in with extreme outcomes. An alternative approach is to collect outcome data among some subjects with missing outcomes (Coppock et al. 2017), which combines double sampling with extreme value bounds. Finally, Huber (2012) seeks to reduce bias from attrition using inverse probability weighting and instrumental variables for missingness.

5.4. *Understand the costs of mediation analysis*

Some researchers include post-treatment covariates as control variables in an effort to test theories about causal mechanisms or to try to estimate the direct effect of a treatment that does not pass through a potential 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 including sequential ignorability, which essentially assumes away the possibility of unmeasured confounders. Many mediation methods like 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) are founded on the *exact same assumption*. The most common mediation models all rely in some way on the assumption that researchers have access to every relevant covariate.

The lesson here is not to that studying mechanisms is impossible or that researchers should give up on trying to understand causal paths. However, there is no free lunch when analyzing mediators in an experiment. For example, Bullock, Green, and Ha (2010) outline experimental designs that facilitate the study of causal mediation by directly manipulating post-treatment mediators as well as treatment assignments. This approach is not only very difficult to execute (it requires a treatment that affects the mediator but not the outcome) but also subject to criticism for implausible assumptions. Imai, Tingley, and Yamamoto (2013) outline several designs that allow researchers to estimate mediation effects, but these too come with additional assumptions (e.g., a consistency assumption) or require use of less intuitive estimands (e.g., average complier indirect effects).

Scholars, reviewers, and editors should recognize that any attempt to include post-treatment variables in a mediation analysis comes at an inferential cost. Unpacking the "black box" of experimental treatments must be paid for in the form of assumptions, biased estimates, or both. Absent any additional assumptions, the best we may be able to do may resemble the "implicit mediation analysis" outlined by Gerber and Green (2012, Section 10.6). Alternatively, one may estimate mediation effects under stronger assumptions while providing a sensitivity analysis to violations of those assumptions per Imai, Keele, and Tingley (2010).

5.5. *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 a hypothesis test about balance in x between the treatment and control conditions. Scholars might, for instance, conduct a bivariate regression testing if x (political interest) differs based on T (the civics class).

However, failing to reject the null hypothesis $H_0 : E(\mathbf{x}|T = 0) = E(\mathbf{x}|T = 1)$ does not rule out post-treatment bias in analyses that condition on x. First, even in the simplified examples presented above, post-treatment bias will not be eliminated unless the effect of the treatment on the covariate (γ) is precisely zero—something that cannot be established using traditional hypothesis testing. Failing to reject the null hypothesis is not direct evidence for the null. Second, we made the simplifying assumption in our examples above that critical parameters including the treatment effect (τ), the effect of the treatment on the covariate (γ), and the effects of the confounders on outcomes and covariates (κ_Y, κ_X) were constant for each individual. There is no reason to believe that these assumptions are correct in real world data. Without them, we cannot be sure a variable is not post-treatment unless we accept the sharp null of no effect *for any unit.* Indeed, Aronow, Baron, and Pinson (2015) show that in a more general setting, it will often not be possible to provide bounds that exclude $-\infty$ and ∞ for the potential bias from conditioning on a post-treatment variable. In the end, the best solution is not to test for post-treatment bias but rather to carefully design experimental protocols that prevent it in the first place.

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 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 identifying causal mechanisms. Though these intentions are laudable, we demonstrate that post-treatment conditioning undermines the value of randomization and biases treatment effect estimates using analytical results as well as a reanalysis of real-world data from two 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 being realistic about the assumptions required for mediation analysis.

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 before an outcome variable can affect subsequent responses (e.g, Kosloff et al. 2010; Leach et al. 2010). For instance, scholars may 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 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.

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 appear to be driven by authors' efforts to show that their proposed mechanism is responsible for the treatment effect. Reviewers often ask authors to try and rule out alternative explanations in this way. However, once an experiment has been conducted, it is not possible to rule out alternative mechanisms without the possibility of post-treatment bias. As shown above, standard approaches such as controlling for intervening variables or subsetting data are incorrect. Similarly, mediation analyses require strong assumptions that may be inconsistent with the goals of experimental research. We hope this article helps convince reviewers and editors not to request such post-hoc statistical analyses and provides evidence researchers can cite to justify avoiding such practices.

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 M., Jonathon Baron, and Lauren Pinson. 2015. "A Note on Dropping Experimental Subjects Who Fail a Manipulation Check." Available at SSRN: `https://ssrn.com/abstract=2683588`.

Athey, Susan, and Guido Imbens. 2016. "The econometrics of randomized experiments." *arXiv preprint arXiv:1607.00698*.

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.

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`.

Bullock, John G, Donald P Green, and Shang E Ha. 2010. "Yes, but what's the mechanism?(don't expect an easy answer)." *Journal of personality and social psychology* 98 (4): 550.

Burnett, Craig M., and Lydia Tiede. 2015. "Party Labels and Vote Choice in Judicial Elections." *American Politics Research* 43 (2): 232–254.

Clark, Tom S., and Jonathan P. Kastellec. 2015. "Source Cues and Public Support for the Supreme Court." *American Politics Research* 43 (3): 504–535.

Coppock, Alexander. 2017. "Comment on White, Nathan, and Faller (2015)." June 19, 2017. Downloaded June 22, 2017 from `https://acoppock.github.io/papers/coppock_comment_WNF.pdf`.

Coppock, Alexander, Alan S. Gerber, Donald P. Green, and Holger L. Kern. 2017. "Combining Double Sampling and Bounds to Address Nonignorable Missing Outcomes in Randomized Experiments." *Political Analysis* 25. Unpublished manuscript. Downloaded November 3, 2016 from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2683588`.

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.

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.

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.

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.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Designing and analyzing randomized experiments: Application to a Japanese election survey experiment." *American Journal of Political Science* 51 (3): 669–687.

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, Dustin Tingley, and Teppei Yamamoto. 2013. “Experimental designs for identifying causal mechanisms.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (1): 5–51.

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

Imbens, Guido W., and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62 (2): 467–75.

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.” In *Computation, causation, and discovery*, ed. C. Glymour and G. Cooper. AAAI Press/The MIT Press.

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.” *Journal of the Royal Statistical Society Series A (General)* 147 (5): 656–666.

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

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

This Online Appendix provides additional analyses, examples, extended notation, coding specifications, and figures in six main sections:

- **Coding of articles** (A-2–A-8): This section contains the specific criteria and rules used to identify the articles with post-treatment conditioning issues and provides details on the articles in which the authors identified these practices.

- **Mathematical expressions and illustrations of post-treatment bias** (A-9–A-12): This section provides a more detailed illustration of the inferential problems that post-treatment bias creates through figures and mathematical notation.

- **Simulation evidence of post-treatment bias** (A-12–A-16): In this section, we provide evidence of the pernicious effects and inferential problems that post-treatment bias implies through several simulation exercises.

- **Additional reanalysis of Dickson, Gordon, and Huber (2015)** (A-15–A-20): This section presents additional analyses of the study conducted by Dickson, Gordon, and Huber (2015) that is described in the main text.

- **Reanalysis of Broockman and Butler (2015)** (A-18–A-24): This section contains an additional reanalysis of published data. In contrast to the original study, we control for post-treatment variables and drop cases based on post-treatment criteria (e.g., manipulation checks) to illustrate the effects that this practice has on the inferences that researchers may reach in the real world.

- **Judge experiment questionnaire** (A-21–A-29): This section provides the experimental instrument used in the study reported in section 4.1.

A1. CODING OF ARTICLES

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. 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. 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 met more than one):

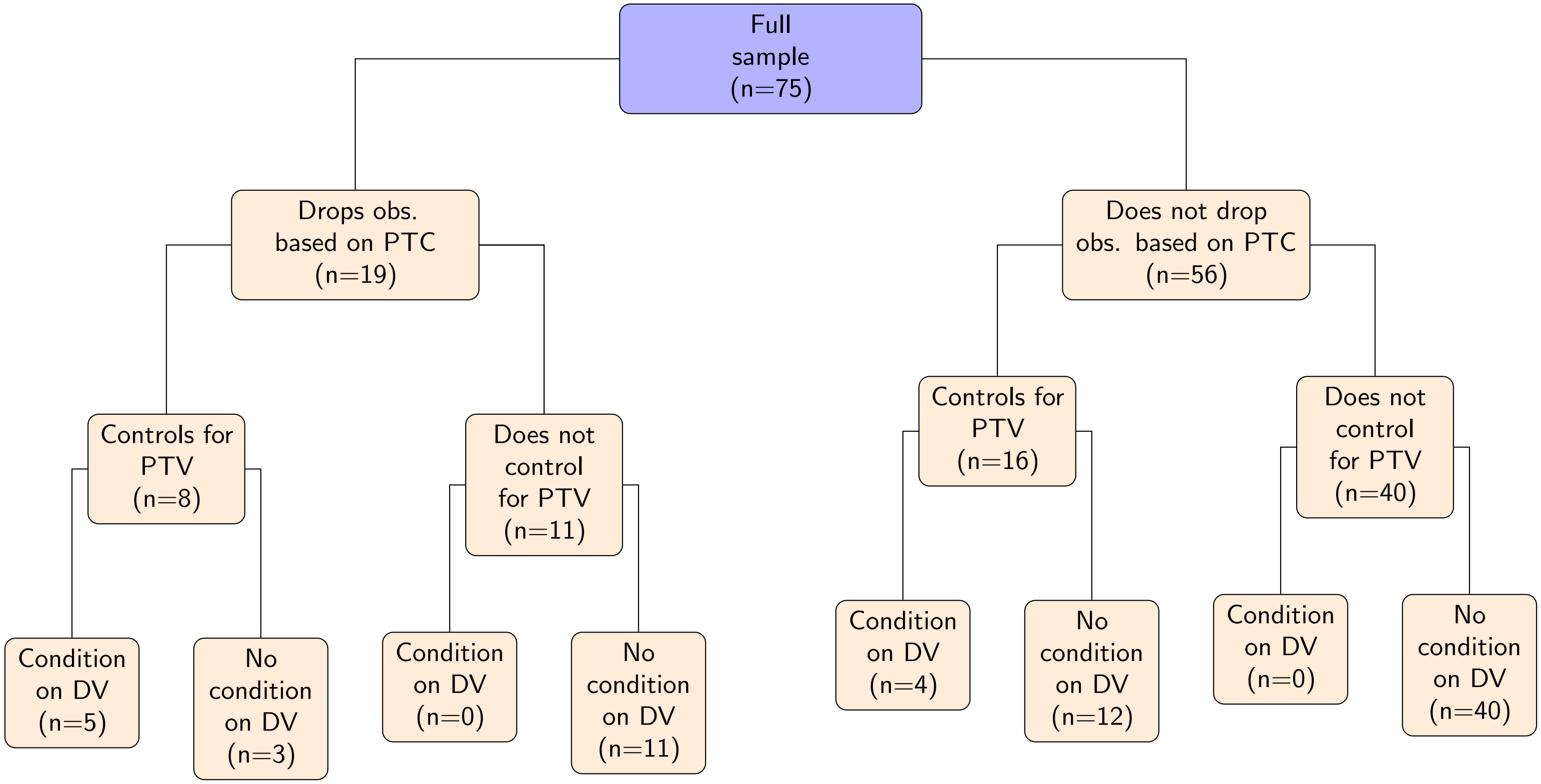
- Articles that control for a variable that the authors themselves show is post-treatment using a statistical model or graph;
- Articles that controlled for variables that were (a) measured after the treatment and (b) could have plausibly been affected by the treatment;[1]
- Articles that dropped cases 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);
- Articles implementing mediation analysis;
- 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;[2]
- Articles that suffered from post-treatment attrition.

[1] 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).

[2] 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.

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).[3] 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. For example, if only seven subjects failed to follow instructions, why not leave them in? The fact that scholars regularly engage in these practices despite the danger of biasing estimates indicates that the problem of post-treatment bias is still not widely recognized.

[3] 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.


Full sample (n=75)
Drops obs. based on PTC (n=19)
Does not drop obs. based on PTC (n=56)
Controls for PTV (n=8)
Does not control for PTV (n=11)
Controls for PTV (n=16)
Does not control for PTV (n=40)
Condition on DV (n=5)
No condition on DV (n=3)
Condition on DV (n=0)
No condition on DV (n=11)
Condition on DV (n=4)
No condition on DV (n=12)
Condition on DV (n=0)
No condition on DV (n=40)

Note: PTC=Post-treatment criteria; PTV= Post-treatment variable; DV=Dependent variable.

Table A1: Coding of articles that include a post-treatment conditioning practice

Author	Title	Journal details	PT issues	PT variable	Location in paper
Toby Bolsen, Paul J. Ferraro, Juan Jose Miranda	Are Voters More Likely to Contribute to Other Public Goods? Evidence from a Large-Scale Randomized Policy Experiment	AJPS 58(1):17–30.	Control for/interact with PTV	Vote frequency	Footnote 6 on pg. 21 Table 1
David Samuels and Cesar Zucco Jr.	The Power of Partisanship in Brazil: Evidence from Survey Experiments	AJPS 58(1):212–25	Drop based on post-treatment criteria	Non-response to PID and experimental questions	Pg. 219
Andrew Healy and Gabriel S. Lenz	Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy	AJPS 58(1):31–47	Drop based on post-treatment criteria	Failed to evaluate a term. Failed an attention test	Pg. 37
Tali Mendelberg, Christopher F. Karpowitz, Nicholas Goedert	Does Descriptive Representation Facilitate Womens Distinctive Voice? How Gender Composition and Decision Rules Affect Deliberation	AJPS 58(2):291–306	Control for/interact with PTV*	Ratio of care to financial issues Care frequency and verbosity	Figure 3 (DV), Table 3 (Control) Pg. 298 Table A6 in Appendix
Luca Corazzini, Sebastian Kube, Michel Andre Marechal, Antonio Nicolo	Elections and Deceptions: An Experimental Study on the Behavioral Effects of Democracy	AJPS 58(3): 579–592	Control for/interact with PTV*	Campaign promises Approval	Pg. 585 and Figure 1 (DV), Table 2 (Control) Pg. 587–8 and Tables 4 and 5
Adam Berinsky, Michele Margolis, and Michael Sances	Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys	AJPS 58(3):739–753	Drop based on post-treatment criteria	Attention filters	Pg. 750–1
Stephen M. Utych, Cindy D. Kam	Viability, Information Seeking, and Vote Choice	JOP 76(1):152–166	Drop based on post-treatment criteria	Non-compliance, vote for unexpected candidate	Pg. 156
Markus Prior	Visual Political Knowledge: A Different Road to Competence?	JOP 76(1): 41–57	Control for/interact with PTV*	Political knowledge	Table 1 (DV), Table 2 (Control)
Scott Clifford	Linking Issue Stances and Trait Inferences: A Theory of Moral Exemplification	JOP 76(3): 698–710	Drop based on post-treatment criteria	Death penalty attitudes	Appendix and Figure 1 Personal communication
Cindy D. Kam and Elizabeth J. Zechmeister	Name Recognition and Candidate Support	AJPS 57(4): 971–986	Control for/interact with PTV$^+$	Perceptions of viability	Table 2
Jens Großer, Ernesto Reuben, and Agnieszka Tymula	Political Quid Pro Quo Agreements: An Experimental Study	AJPS 57(3): 582–597	Drop based on post-treatment criteria and Control for/interact with PTV*	Subset by tax level in time 2 Changes in taxes and transfers	Pg. 588–589 and Figure 1 (DV) Pg. 592 and Table 2 (Control) Table 3 (Subset)
Matthew S. Levendusky	Why Do Partisan Media Polarize Viewers?	AJPS 57(3): 611–623	Drop based on post-treatment criteria	Attrition	Pg. 619 and Table 3

Author	Title	Journal details	PT issues	PT variable	Location in paper
T. K. Ahn, Robert Huckfeldt, Alexander K. Mayer, John Barry Ryan	Expertise and Bias in Political Communication Networks	AJPS 57(2): 357–373	Control for/interact with PTV	Information purchased	Pg. 362 and Table 2
Michael R. Tomz and Jessica L. P. Weeks	Public Opinion and the Democratic Peace	APSR 107(4):849–865	Control for/interact with PTV and Drop based on post-treatment criteria*+	Perceptions of threat Attrition	Table 3 (DV), Table 5 (control) Pg. 854
Andrew Beath, Fotini Christia, Ruben Enikolopov	Empowering Women through Development Aid: Evidence from a Field Experiment in Afghanistan	APSR 107(3): 540–557	Control for/interact with PTV	Project completion	Pg. 554 and Table A8
Thad Dunning and Janhavi Nikelani	Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils	APSR 107(1): 35–56	Drop based on post-treatment criteria	Party ID	Pg. 52 and Table 6A Footnote 66
Daryl G. Press, Scott D. Sagan, and Benjamin A. Valentino	Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-Use of Nuclear Weapons	APSR 107(1): 188–206	Drop based on post-treatment criteria	Manipulation check	Pg. 196
Samara Klar	The Influence of Competing Identity Primes on Political Preferences	JOP 75(4): 1108–24	Drop based on post-treatment criteria	Party ID	Pg. 1114 Online appendix
Dennis Chong and James N. Druckman	Counterframing Effects	JOP 75(1): 1–16	Drop based on post-treatment criteria	Attrition	Pg. 7
Michael Tesler	The Return of Old-Fashioned Racism to White Americans' Partisan Preferences in the Early Obama Era	JOP 75(1): 110–123	Control for/interact with PTV	Opposition to interracial dating Racial resentment	Pg. 120 and Table 3 Personal communication
James N. Druckman and Thomas J. Leeper	Learning More from Political Communication Experiments: Pretreatment and Its Effects	AJPS 56(4): 875–96	Control for/interact with PTV and Drop based on post-treatment criteria	Media consumption and Need to evaluate Attrition	Pg. 886 and Figure 9 Footnotes 18 and 21 `https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/17218&version=3.2`
Guy Grossman and Delia Baldassarri	The Impact of Elections on Cooperation: Evidence from a Lab-in-the-Field Experiment in Uganda	AJPS 56(4): 964–85	Control for/interact with PTV	Monitor profiles	Pg. 972 and Table 1
Xiaobo Lu, Kenneth Scheve, Matthew J. Slaughter	Inequity Aversion and the International Distribution of Trade Protection	AJPS 52(3): 638–54	Control for/interact with PTV	Personal income	Pg. 650 Personal communication

Author	Title	Journal details	PT issues	PT variable	Location in paper
Michael Tesler	The Spillover of Racialization into Health Care: How President Obama Polarized Public Opinion by Racial Attitudes and Race	AJPS 56(3): 690–704	Control for/interact with PTV	Racial resentment	Table A4 in the Appendix Personal communication
Kevin Arceneaux	Cognitive Biases and the Strength of Political Arguments	AJPS 56(2): 271–85	Control for/interact with PTV^{+}	Anxiety and anger	Pg. 275–6 and Table 1
Antoine J. Banks and Nicholas A. Valentino	Emotional Substrates of White Racial Attitudes	AJPS 56(2): 286–97	Drop based on post-treatment criteria	Manipulation check	Pg. 289
Dona-Gene Mitchell	It's About Time: The Lifespan of Information Effects in a Multiweek Campaign	AJPS 56(2): 298–311	Control for/interact with PTV and Drop based on post-treatment criteria	Correct memory Attrition	Pg. 305 and Table 1 Pg. 308
Marco Battaglini, Salvatore Nunnari and Thomas R. Palfrey	Legislative Bargaining and the Dynamics of Public Investment	APSR 106(2): 407–29	Control for/interact with PTV	Investment level Herfindahl index Inequality of proposal	Pg. 421 and Table 7
James N. Druckman, Jordan Fein and Thomas J. Leeper	A Source of Bias in Public Opinion Stability	APSR 16(2): 430–54	Drop based on post-treatment criteria and Control for/interact with PTV^{*+}	Health opinion and search behavior Attitude certainty Attrition	Table 4 Footnote 19 on Pg. 441 Pg. 433 and 441
Christopher F. Karpowitz, Tali Mendelberg and Lee Shaker	Gender Inequality in Deliberative Participation	APSR 106(3): 533–47	Control for/interact with PTV^{*+}	Speaking behavior	Table 3 (DV), Table 5 (control)
Edmund Malesky, Paul Schuler and Anh Tran	The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in an Authoritarian Assembly	APSR 106(4): 762–86	Control for/interact with PTV and Drop based on post-treatment criteria*	Debate speech count Nominated candidates	Table 6 (DV), Table 7 (control) Table 7
James L. Gibson and Gregory A. Caldeira	Campaign Support, Conflicts of Interest, and Judicial Impartiality: Can Rescusals Rescue the Legitimacy of Courts?	JOP 74(1): 18–34	Control for/interact with PTV	Knowledge of the case	Coding in Appendix and pg. 28
Matthew S. Levendusky and Michael C. Horowitz	When Backing Down is the Right Decision: Partisanship, New Information, and Audience Costs	JOP 74(2): 323–38	Control for/interact with PTV^{+}	Judgement of competence	Footnote 21 on pg. 334
Kyle Mattes	What Happens when a Candidate Doesn't Bark? "Cursed" Voters and Their Impact on Campaign Discourse	JOP 74(2): 369–82	Drop based on post-treatment criteria and Control for/interact with PTV*	Campaign choice Opposition behavior	Table 3 Pg. 377

Author	Title	Journal details	PT issues	PT variable	Location in paper
Robert Johns and Graeme A. M. Davies	Democratic Peace of Clash of Civilizations? Target States and Support for War in Britain and the United States	JOP 74(4): 1038–52	Control for/interact with PTV and Drop based on post-treatment criteria	Religious affiliation and attendance Authoritarianism, nationalism, and social dominance	Pg. 1046 Pg. 1044 Table 3 `http://www.tessexperiments.org/data/johns798.html`

Note: PTV = Post-treatment variable; * = condition on a dependent variable; + = the study implements mediation and/or causal path analysis; DV = Dependent variable.[4]

[4] Note: The outcomes we describe as DVs are not necessarily the main outcome variable of the study; we use the term to refer to any variable that is modeled as a DV in analyses presented in the main text or supplementary material of the study. In cases where an article controls for or interacts with a post-treatment variable that is a DV, we indicate first where a variable is treated as an outcome and then where the authors condition on it.

A2. MATHEMATICAL EXPRESSIONS AND ILLUSTRATIONS OF POST-TREATMENT BIAS

A2.1. *Imbalance in treatment with respect to confounders*

As explained in the main text, one of the central implications of practices that involve conditioning on post-treatment variables in an experimental setting is the loss of the benefits of randomization. In Section 3, we show that post-treatment conditioning creates imbalance between the treatment and control groups with respect to an unmeasured confounder u. Recall that conditioning on a post-treatment variable implies comparing outcomes between groups with different treatment assignments but the same post-treatment values. For example, if we subset or drop cases based on a a particular value of the post-treatment variable, say $x = 1$, then we will compare the outcome y between treatment and control groups among a subsample defined by x. However, the balance in treatment with respect to confounders that randomization is designed to achieve does not hold among this group.

To illustrate this problem, recall the data generating process specified in Equation (7) in the main text. Further assume that $\alpha_x = 0$, u is distributed normally, $c > 0$, and $\gamma > 0$. First consider the case where we look only at cases where $x = 1$. 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 A1. 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 A1. 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. Further, note that u will on average be higher in the control condition, which means that the bias that is induced in this example 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.

Specifically, note that the distribution of u in these figures is simply a truncated normal distribution. Thus, we know:

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

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 (shown in Figure 3 in the main text), but the problem is the same. The expected values of the unmeasured confounder u in this case would be as follows:

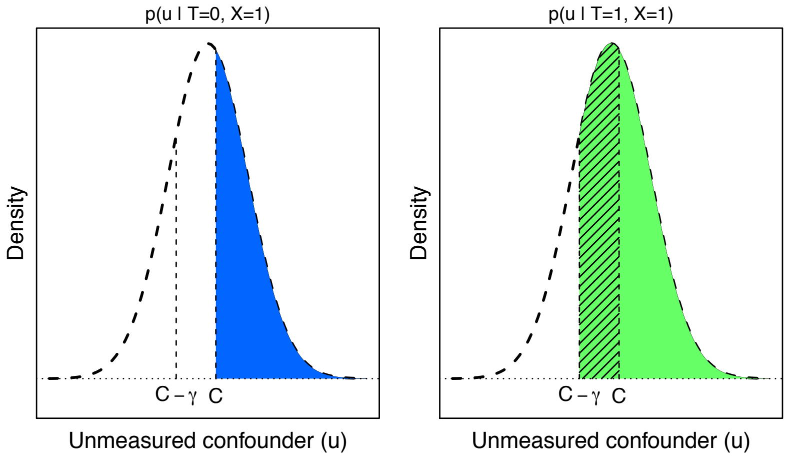
$$\begin{aligned} \mathbb{E}(\mathbf{u}|T = 1, \mathbf{x} = 0) &= -\sigma_u(\phi(\tfrac{C-\gamma}{\sigma_u}))/(\Phi(\tfrac{c-\gamma}{\sigma_u})), \text{ and} \\ \mathbb{E}(\mathbf{u}|T = 0, \mathbf{x} = 0) &= -\sigma_u(\phi(\tfrac{c}{\sigma_u}))/(\Phi(\tfrac{c}{\sigma_u})). \end{aligned} \tag{2}$$

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

A2.2. *The consequences of imbalance*

We next present two visualizations to help illustrate how inappropriately conditioning on a post-treatment variable can bias our estimated treatment effect. First, Figure A2 shows how imbalance in u induced by post-treatment conditioning can lead to mistaken inferences.

Figure A1: 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 7. We assume $\alpha_x = 0$, $C > 0$, $\gamma > 0$, and that u is distributed normally.

Figure A2: How post-treatment conditioning can create spurious treatment effect estimates

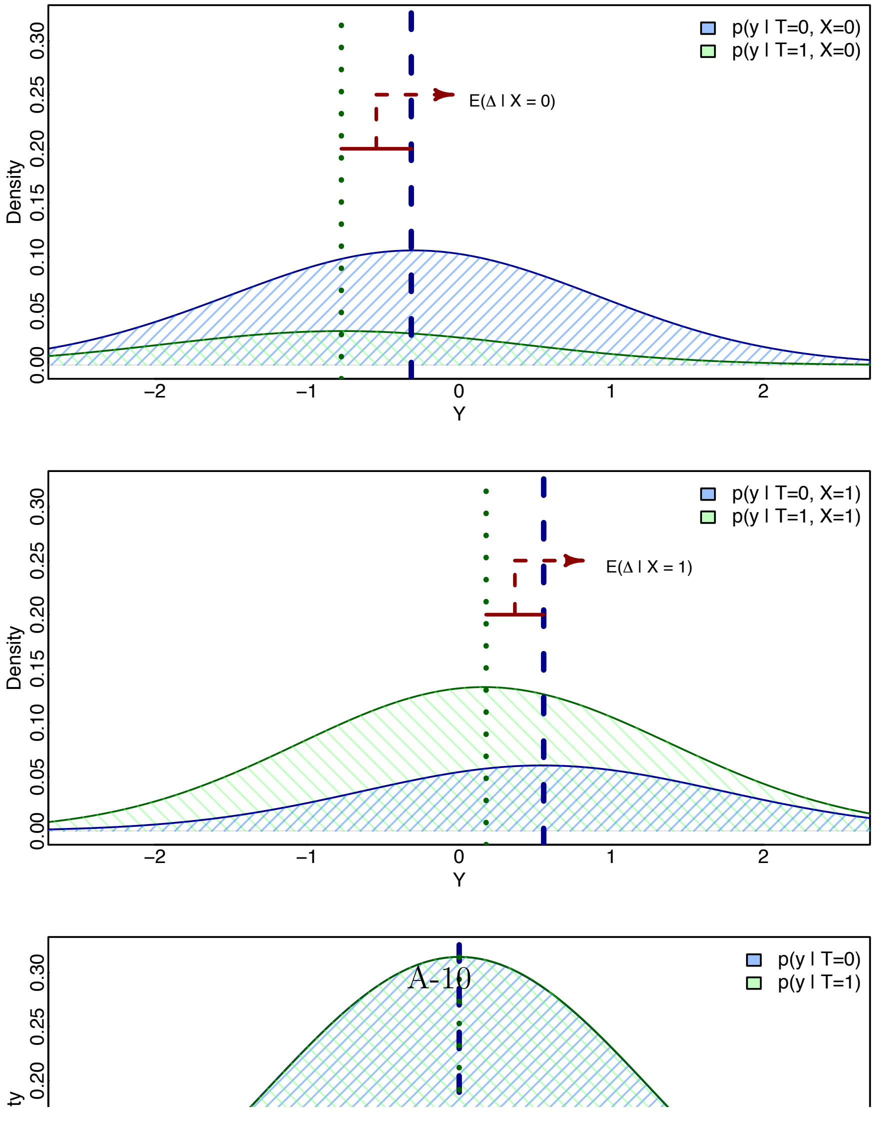
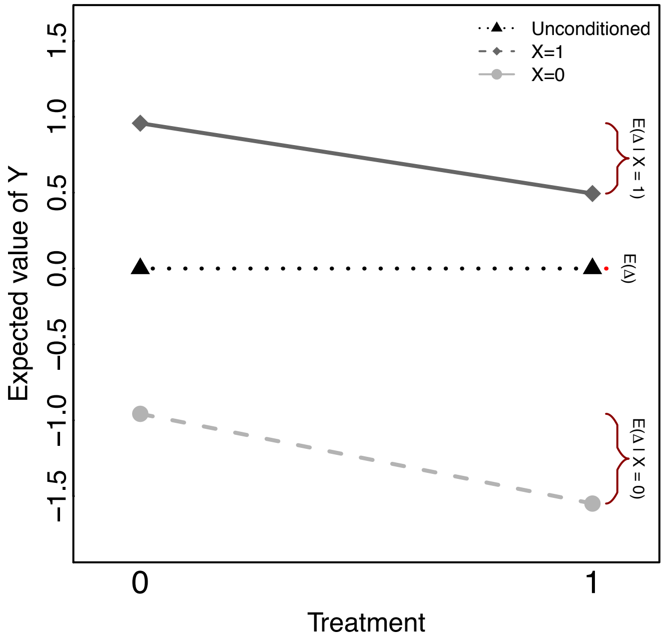


Figure A3: How conditioning on a post-treatment variable can bias treatment effect estimates



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 7 and assume that $\alpha_x = 0$, $C > 0$, $\gamma > 0$, and u distributed normally.

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).

To help visualize the bias, Figure A3 shows the expected value of the outcome for each possible combination of y and x in our example. The figure shows that the treatment appears to have a *negative* effect on the outcome when holding constant the value of x either by subsetting or through statistical control. However, the effect of the treatment in this simulated data is actually zero, as shown by the unconditional relationship between T and y. Conditioning on the post-treatment variable will lead us to the wrong conclusion.

A2.3. *Simulation evidence of post-treatment bias*

To further demonstrate the pernicious effects of conditioning on post-treatment variables, we simulate data by slightly altering the assumptions used above by adding independent error terms to Equation (7) and generate data using the following model:

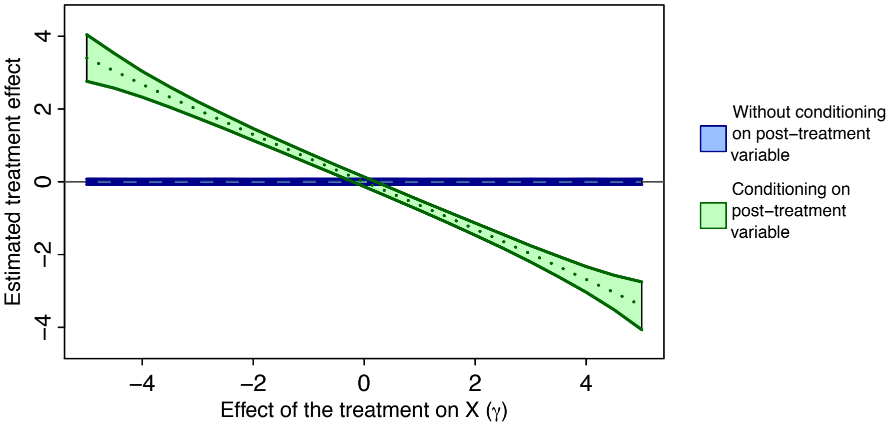
$$\begin{aligned} y_i &= \kappa_Y u_i + \tilde{\epsilon}_{y,i} \\ x_i &= \gamma T_i + \kappa_X u_i + \tilde{\epsilon}_{x,i} \end{aligned} \qquad (3)$$

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.

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 on the observed value of our covariate x (specifically, the 15% with the highest value of x). That is, we simulate data according to Equation (3) for different parameter values of γ and κ_Y. For each unique combination of parameter values, we simulate 10,000 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 considers 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 A4 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 $\gamma = 0$.

Figure A4: Post-treatment bias when dropping cases as a function of treatment effect on X



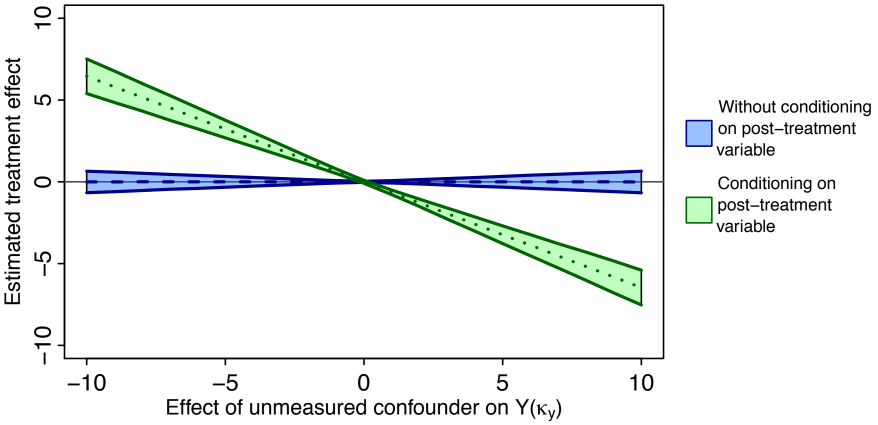
The plot shows the 90% Monte Carlo confidence intervals for point estimates of the treatment effect when 15% of the sample is dropped based on post-treatment criteria. Data were generated according to Model 3 for differing values of γ where $\kappa_X = 1$, κ_Y=1, and $\sigma_u = 2$. We fit 10,000 regressions 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 A5 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

[5] We fix $\kappa_x = 1$, κ_y=1, and $\sigma_u = 2$.

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.

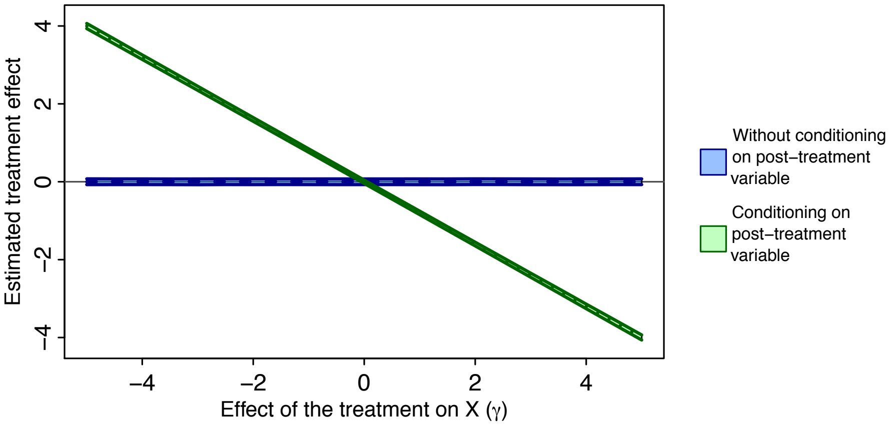
Figure A5: Post-treatment bias when dropping cases as a function of unobserved confounding



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 3 for differing values of κ_Y where $\gamma = 1$, κ_Y=1, and $\sigma_u = 2$. We fit 10,000 regressions for each parameter combination.

Finally, we show that these same problems persist if instead of dropping cases, we simply include the post-treatment variable in the regression equation. Figure A6 shows that a model that omits a post-treatment control generally yields unbiased estimates—the blue shaded area is centered at the true value of $\tau = 0$. However, controlling for a post-treatment covariate can again induce severe bias of almost any size or direction depending on the values of γ. Note again that the magnitude and even the direction of the potential bias depends on the relationship between the unmeasured confounder u and the outcome y, which cannot be directly estimated and is difficult to diagnose.

Figure A6: Post-treatment bias controlling for x as a function of the γ (the treatment's effect on x)

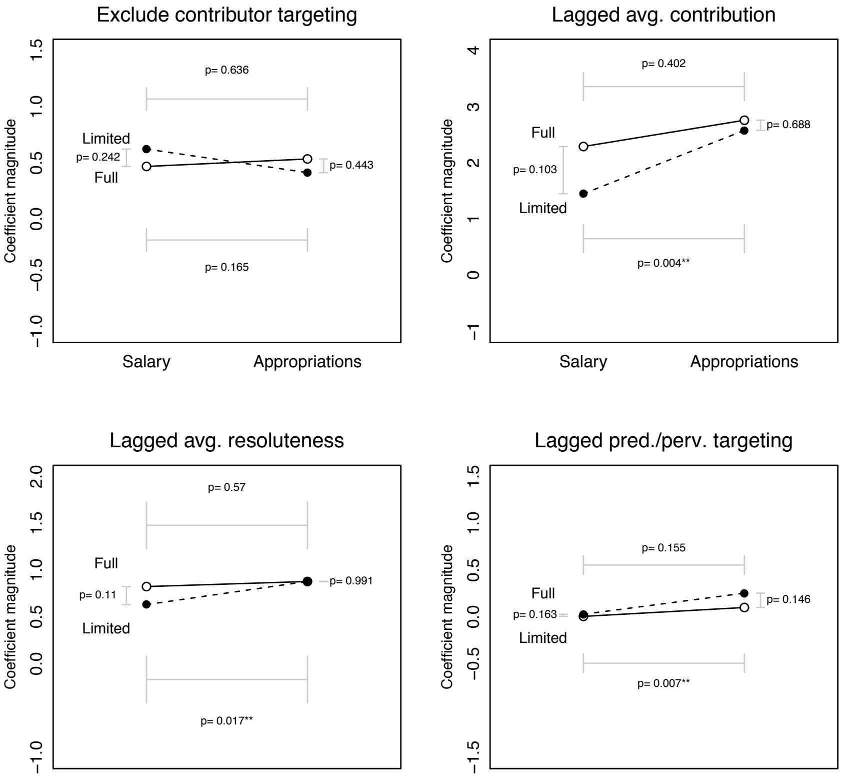


The plot shows the 90% Monte Carlo confidence intervals for point estimates of the treatment effect when a post-treatment covariate is a included as a control variable in linear regression. Data were generated according to Model 3 for differing values of γ where $\kappa_X = 1$, κ_Y=1, and $\sigma_u = 2$. We fit 10,000 regressions for each parameter combination.

A3. ADDITIONAL REANALYSIS OF DICKSON, GORDON, AND HUBER (2015)

Experimental balance: In this section, we provide further reanalysis of Dickson, Gordon, and Huber (2015). For example, Figure A7 tests and illustrates the effect of the treatments on the different post-treatment variables used in the analysis. The bold lines in each panel of Figure A7 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 circle), for both the salary and appropriation groups. The top left panel of Figure A7 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 A7: 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.

Treatment effect differences in Dickson, Gordon, and Huber (2015): The pooled no-intercept models reported in Table A3 are the source of the treatment effect estimates in Table 3 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 separately among respondents assigned to the full information condition in Table 2 and among those assigned to the limited information condition in Table 4. Rather than estimating

Table A3: 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

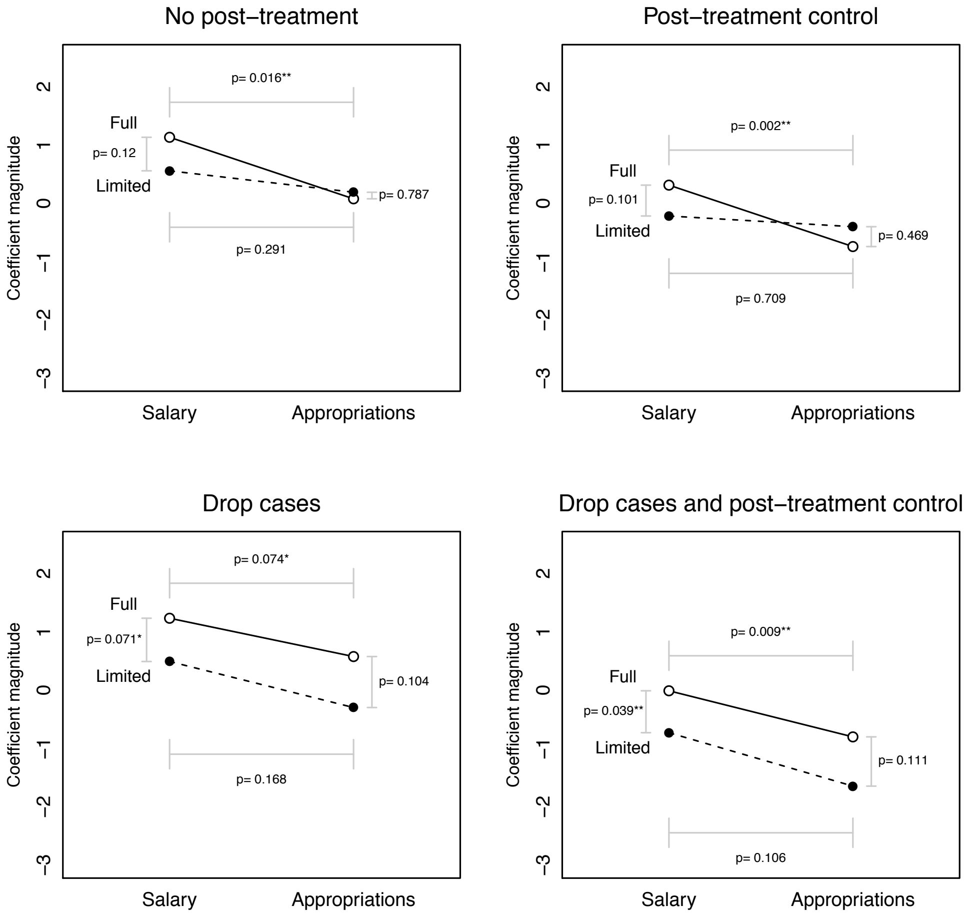
$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

these effects separately on subsets of the data, we estimate an equivalent interaction model on the full data in Table A3. For example, the coefficient of -0.382 of the Appropriations treatment in Table 4 (original study) indicates that under this condition, there is lower net assistance to the authority than under the Salary condition but only among subjects in a limited information context. In our study, this effect is obtained in Table A3 by taking the difference of the coefficients of "Appropriations/limited information"(0.190) and "Salary/limited information" (0.558), which is equal to -0.368.[6]

Figure A8 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 A7, the graph allows for comparisons between conditions in the 2×2 design. In the figure, we highlight the differences in outcome means between treatment combinations and include both the *p*-value for each difference and stars to indicate whether those differences are statistically significant at conventional levels.

[6] Note that because of the inclusion of period indicators, the estimated treatment effects differ slightly between the subset and the interaction approaches.

Figure A8: Differences in treatment effect estimates between models



*p < .1; **p < .05. Data from Dickson, Gordon, and Huber (2015). See Table 3 for corresponding model results.

Reanalysis of Broockman and Butler (2015)

To further demonstrate the bias that can result from conditioning on post-treatment variables, we use 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.[7]

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.[8] 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 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 ($p < .05$).[9]

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.[10] Table A4 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).[11]

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 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.[12]

[7] We obtain substantively similar results when we condition on a manipulation check from the second study in the article as well — see below for further details.

[8] Careful attention to research design can allow researchers to condition on manipulation checks. For example, placebo-controlled experiments may allow researchers to condition on compliance directly so long as compliance is measured before the treatment (e.g., Broockman, Kalla, and Sekhon 2017).

[9] See Figure A9 for a visualization of the resulting imbalance.

[10] 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.

[11] 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).

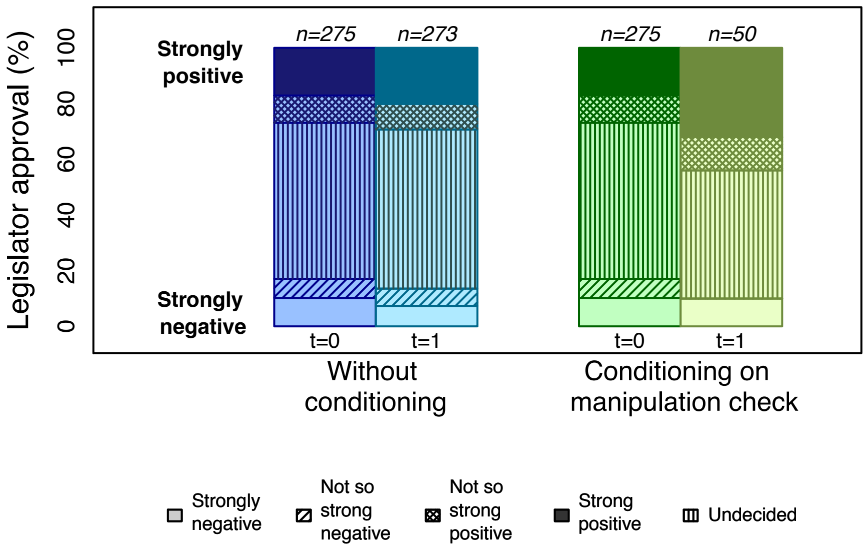
[12] Of course, the difference between the original model estimate and the results obtained using post-treatment conditioning is not necessarily itself statistically significant. Our point instead is that scholars who condition on a post-treatment variable would reach mistaken conclusions in a null hypothesis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

Table A4: 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.

Figure A9: 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 A9 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 A5: 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: 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.[13] 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. We do so in order to demonstrate the pernicious consequences of post-treatment conditioning.

The analysis in Table A5 is restricted to the subset of respondents who were asked about the legislator's position. As in Table A4, 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 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.[14] 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 A5, which reports this auxiliary quantity in the sixth row).

[13] Broockman and Butler (2015, 9) consider three measures of legislator agreement. As in Table A4, we again focus on the binary agreement measure here for ease of exposition.

[14] 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.

Judge experiment questionnaire

How old are you?
-Under 18
-18-24
-25-34
-35-44
-45-54
-55-64
-65-74
-75-84
-85 or older

In what state do you currently reside? [pulldown menu]

What is your gender?
-Male
-Female
-Other

Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else?
-Republican
-Democrat
-Independent
-Something else

Display This Question:
If Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else? Democrat Is Selected
Would you call yourself a strong Democrat or not a very strong Democrat?
-Strong Democrat
-Not very strong Democrat

Display This Question:
If Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else? Republican Is Selected
Would you call yourself a strong Republican or not a very strong Republican?
-Strong Republican
-Not very strong Republican

Display This Question:
If Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else? Independent Is Selected
Or Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else? Something else Is Selected
Do you think of yourself as closer to the Republican Party or to the Democratic Party?
-Closer to the Republican Party
-Closer to the Democratic Party
-Neither

Do you approve or disapprove of the way Donald Trump is handling his job as president?
-Strongly approve
-Somewhat approve

-Somewhat disapprove
-Strongly disapprove

When it comes to politics, would you describe yourself as liberal, conservative, or neither liberal nor conservative?
-Very conservative
-Somewhat conservative
-Slightly conservative
-Moderate; middle of the road
-Slightly liberal
-Somewhat liberal
-Very liberal

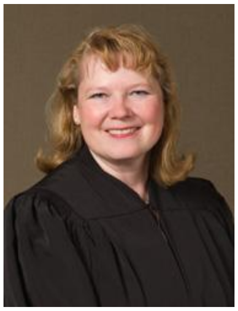
What is the highest degree or level of school you have completed?
-Did not graduate from high school
-High school diploma or the equivalent (GED)
-Some college
-Associate degree
-Bachelor's degree
-Master's degree
-Professional or doctorate degree

Please check one or more categories below to indicate what race(s) you consider yourself to be.
-White
-Black or African American
-American Indian or Alaska Native
-Asian/Pacific Islander
-Multi-racial
-Other

Are you of Spanish or Hispanic origin or descent?
-Yes
-No
-Don't know

We are interested in what people think about judges. Please read the following information about a judge carefully. We will then ask you some questions about that judge afterward.

[treatment - randomized]



Allison Eid has been a Colorado Supreme Court justice since 2006. Here are some facts about Justice Eid.
-She earned a law degree at the University of Chicago and served as a law clerk for the U.S. Fifth Circuit Court of Appeals and the U.S. Supreme Court.
-After working as a litigator at a private law firm, she became a law professor at the University of Colorado.

-She is the former solicitor general for the state of Colorado.
-Donald Trump named her as one of the 11 judges he might pick as a Supreme Court nominee.

[control - randomized]

Allison Eid has been a Colorado Supreme Court justice since 2006. Here are some facts about Justice Eid.
-She earned a law degree at the University of Chicago and served as a law clerk for the U.S. Fifth Circuit Court of Appeals and the U.S. Supreme Court.
-After working as a litigator at a private law firm, she became a law professor at the University of Colorado.
-She is the former solicitor general for the state of Colorado.

Display This Question:
If In what state do you currently reside? -Colorado Is Not Selected
If you were a resident of Colorado, how likely would you be to vote to retain Eid on the Colorado Supreme Court in the future?
-Extremely likely
-Very likely
-Somewhat likely
-Not at all likely

Display This Question:
If In what state do you currently reside? -Colorado Is Selected
How likely are you to vote to retain Eid on the Colorado Supreme Court in the future?
-Extremely likely
-Very likely
-Somewhat likely
-Not at all likely

Based on what you know, would you describe Allison Eid as liberal, conservative, or neither liberal nor conservative?
-Very conservative
-Somewhat conservative
-Slightly conservative
-Moderate; middle of the road
-Slightly liberal
-Somewhat liberal
-Very liberal

We would like to know how well you think the phrases below describe Allison Eid. For each, please indicate if it describes her extremely well, very well, moderately well, slightly well, or not well at all.

She is moral
She is knowledgeable
She is intelligent

She is honest
She really cares about people like me

Response options:
-Extremely well
-Very well
-Moderately well
-Slightly well
-Not well at all

References

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

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.

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.

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`.

Broockman, David E., Joshua L. Kalla, and Jasjeet S. Sekhon. 2017. "The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs." Forthcoming, *Political Analysis*. Downloaded June 30, 2017 from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2742869`.

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

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.

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.

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." *Journal of Politics* 74 (04): 1038–1052.

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.

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