AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It

Jacob M. Montgomery Washington University in St. Louis
Brendan Nyhan Dartmouth College
Michelle Torres Washington University in St. Louis

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 posttreatment variables in statistical models, eliminating observations based on posttreatment criteria, or subsetting the data based on posttreatment 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 posttreatment 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 posttreatment 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.*

Replication Materials: 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: https://doi.org/10.7910/DVN/EZSJ1S.

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 of those who 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 posttreatment variables. Many applied scholars seem unaware that these common practices amount to conditioning on posttreatment 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

Jacob M. Montgomery is Associate Professor, Department of Political Science, Washington University in St. Louis, Campus Box 1063, One Brookings Drive, St. Louis, MO 63130 (jacob.montgomery@wustl.edu). Brendan Nyhan is Professor, Department of Government, Dartmouth College, HB 6108, Hanover, NH 03755 (nyhan@dartmouth.edu). Michelle Torres is PhD Candidate, Department of Political Science, Washington University in St. Louis, One Brookings Drive, Campus Box 1063, Saint Louis, MO 63130 (smtorres@wustl.edu).

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

American Journal of Political Science, Vol. 62, No. 3, July 2018, Pp. 760–775

 DOI: 10.1111/ajps.12357

unlikely to hold in real-world settings. In short, conditioning on posttreatment variables can ruin experiments; we should not do it.

Though the dangers of posttreatment bias have long been recognized in the fields of statistics, econometrics, and political methodology (e.g., Acharya, Blackwell, and Sen 2016; Elwert and Winship 2014; King and Zeng 2006; Rosenbaum 1984; Wooldridge 2005), 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 posttreatment bias in experimental political science research. We first identify common practices that lead to posttreatment conditioning and document their prevalence in articles published in the field's top journals. We then provide analytical results that explain how posttreatment 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 posttreatment bias.

Don't We Already Know This?

We first address the notion that the dangers of posttreatment bias are already well known. After all, published research in political science identified posttreatment bias (in passing) as problematic over a decade ago (King and Zeng 2006, 147–48). More recent work has amplified these points in the context of observational research (Acharya, Blackwell, and Sen 2016; Blackwell 2013). Some readers may wonder whether this exercise is needed given the increasingly widespread understanding of causal analysis in the discipline. In this section, we show that the dangers of posttreatment conditioning are either not understood or are being ignored—our review of the published literature suggests that it is widespread.

Of course, conditioning on posttreatment variables is not a practice that is exclusive to experimental research. Indeed, we believe the prevalence of and bias from posttreatment 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 posttreatment variables. We speculate that posttreatment 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 posttreatment 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 are typically unambiguous, making it easy for scholars to avoid conditioning on posttreatment variables. Second, for pedagogical purposes, explaining posttreatment bias in experiments allows for greater expositional clarity, reduces ambiguity about whether variables are measured posttreatment in the examples we discuss, and allows us to generate an unbiased estimate for purposes of comparison in our applications.

Table 1 Posttreatment Conditioning in Experimental Studies

Category	Prevalence
Engages in posttreatment conditioning	46.7%
Controls for/interacts with a posttreatment variable	21.3%
Drops cases based on posttreatment criteria	14.7%
Both types of posttreatment conditioning present	10.7%
No conditioning on posttreatment variables	52.0%
Insufficient information to code	1.3%

Note: The sample consists of 2012–14 articles in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* including a survey, field, laboratory, or lab-in-the-field experiment (n = 75).

To demonstrate the prevalence of posttreatment 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 the *Journal of Politics* (*JOP*) that included one or more survey, field, laboratory, or lab-in-the-field experiments from 2012 to 2014 (n = 75). We coded each article for whether the authors subsetted the data based on potentially posttreatment 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

[1] Additional details on these coding procedures as well as a listing of articles coded as having some form of posttreatment conditioning are provided in the supporting information.

APSR, *AJPS*, and *JOP* from 2012 to 2014 engaged in posttreatment conditioning (35 of 75 studies). Specifically, more than 1 in 3 studies engaged in at least one of two problematic practices—21.3% (16 of 75) controlled for a posttreatment covariate in a statistical model, and 14.7% of studies subsetted the data based on potential posttreatment criteria (11 of 75 studies reviewed)—and almost 1 in 10 engaged in both (10.7%, eight studies). Among the studies that controlled for a posttreatment variable, six used a mediation technique (8%). Further, while some studies lost cases due to posttreatment attrition (8.0%), others chose to subset their samples or drop cases based on failed manipulation checks, noncompliance, attention screeners, or other posttreatment 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).

In short, nearly half of the experimental studies published in our discipline's most prestigious journals during this period raise concerns about posttreatment bias. About 1 in 4 drop cases or subset the data based on posttreatment criteria, and nearly a third include posttreatment variables as covariates. Further, few acknowledge potential concerns regarding the bias that posttreatment conditioning can introduce. Most tellingly, *nearly 1 in 8 articles directly conditions on variables that the authors themselves show as being an outcome of the experiment*–an unambiguous indicator of a fundamental lack of understanding among researchers, reviewers, and editors that conditioning on posttreatment variables can invalidate results from randomized experiments. Empirically, then, the answer to the question of whether the discipline already understands posttreatment bias is clear: It does not.

The Inferential Problems Created by Posttreatment Bias

The pervasiveness of posttreatment 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 posttreatment bias. To be sure, the subject has been covered extensively in technical work in statistics and econometrics dating back at least to Rosenbaum (1984). What the literature lacks, however, is a treatment of this subject that is both rigorous and accessible to nontechnical readers. Indeed, in many popular textbooks, the bias that results from conditioning on posttreatment covariates is discussed only briefly (Angrist and Pischke 2014, pp. 214–17; Gelman and Hill 2006, Section 9.7). 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 posttreatment bias will help improve experimental designs and analyses in political science. We refer readers to, for example, Imbens and Angrist (1994), Aronow, Baron, and Pinson (2015), Athey and Imbens (2016), and the works cited therein for more technical discussions.

The Intuition of Posttreatment Bias

The intuition behind posttreatment 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 posttreatment 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).[2]

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 pretreatment characteristics make them least likely to participate

[2] 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.*

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.[3]

As this example illustrates, concerns about posttreatment bias are not really (or only) about the posttreatment variable itself. The problem is that by conditioning on a posttreatment 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.

Why Experiments Generate Unbiased Estimates of Treatment Effects

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

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

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

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

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

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

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

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

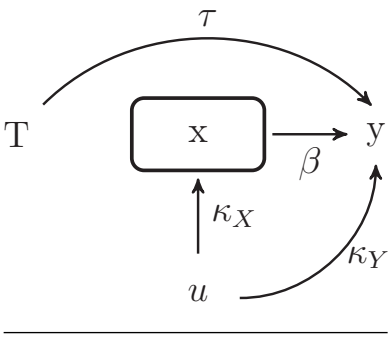
The reason experiments work so well is that random assignment guarantees key assumptions[4] needed to ensure that $\Delta = \tau$, an equality that must hold to ensure $\hat{\Delta}$ is an unbiased estimate of τ. Chief among these assumptions is

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

which states that treatment assignment is independent of potential outcomes conditional on 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 pretreatment covariate, meaning that $T \perp\!\!\!\perp x$. Equation (4), which is shown visually in Figure 1,[5]

FIGURE 1 Causal Graph When the Covariate Is Unaffected by the Treatment



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

[4]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).

[5]Pearl (2009) shows that the graphical causal model approach is equivalent to the potential outcomes framework we use above. It

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.[6] Using our example above, y represents respondents' turnout decision, T represents the experimental civics education class, x represents respondents' *pretreatment* 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:

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

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,[7] this can be expressed as

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

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 posttreatment 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—that is, that respondents' level of pretreatment 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).

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.

The Problem with Conditioning on Posttreatment Variables

We are now ready to directly discuss posttreatment bias. In short, when we include a posttreatment 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.[8] 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)

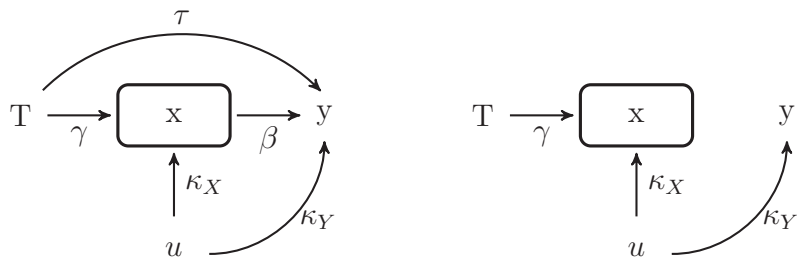
is often especially helpful in clarifying which research designs can accurately recover causal estimates, which is why we employ it here.

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

[7] Note that Model 4 also assumes that the main parameters in the model ($\tau, \alpha_Y, \beta, \tau$, 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.)

[8] For expositional clarity, we omit edge cases that would allow us to condition on posttreatment 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 Posttreatment Variable



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

parameters (e.g., κ_Y, the effect of the unmeasured confounder on the outcome). Once we have conditioned on a posttreatment 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 is thereby "posttreatment," 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 $\gamma = 0$ (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 $\beta = \tau = 0$ (no effect of either the intervention or the observed covariate on the outcome).

Under these circumstances, it may seem harmless to condition on the posttreatment covariate x—after all, x has no effect on y.[9] This intuition is wrong. Even in such a favorable context, conditioning on x *still* leads to inconsistent estimates because the posttreatment 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).[10] 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.

Practices That Lead to Posttreatment Bias

Conceptually, there are two ways that researchers may condition on posttreatment variables: dropping (or subsetting) observations based on posttreatment criteria or controlling for posttreatment 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 posttreatment variables is nearly unavoidable. The treatment itself may cause some respondents to be more likely to be omitted from the sample, a phenomenon usually termed *nonrandom attrition.*

[9] 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 posttreatment bias that arises from unblocking the path from u to y.

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

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 renominated (table 7.1.1). As a result, treatment effect estimates among legislators who were renominated inadvertently condition on a posttreatment variable. Similar problems can occur when analyzing the content of responses in audit experiments where some legislators do not reply (Coppock 2017).

In other instances, scholars intentionally condition on posttreatment variables. For instance, researchers frequently drop subjects who fail a posttreatment 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 posttreatment 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 in their paper), 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 posttreatment 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 posttreatment 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

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

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

$$\begin{aligned}\Delta \quad &= \tau\\ &\quad + \underbrace{\kappa_Y(\mathbb{E}(\mathbf{u}|T=1,\mathbf{x}=1) - \mathbb{E}(\mathbf{u}|T=0,\mathbf{x}=1))}_{\text{Bias from imbalance in u when x=1}}. \qquad (9)\end{aligned}$$

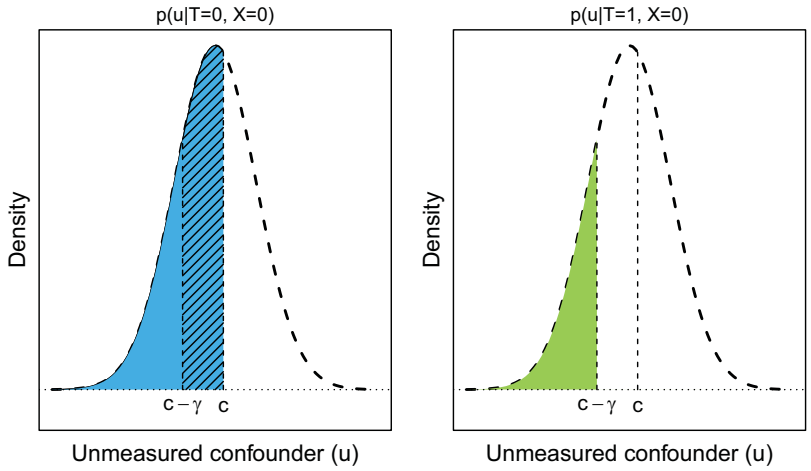
Although it is possible to construct examples where this bias is zero, it will not be zero in general. The reason is that the value of u must on average be lower for observations in the treatment group ($T = 1$) who also meet the selection criteria ($x = 0$) under the assumed data-generating process 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 ($x = 0$).[11]

Including Posttreatment Variables as Covariates. A closely related practice is to control for one or more posttreatment covariates in a statistical model. In our example, this could occur if the posttreatment 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 effort to prevent omitted variable bias (which is not a concern in experiments). In

[11] In the supporting information, we provide exact calculations for the bias shown in this figure.

FIGURE 3 Example of How Conditioning on a Posttreatment Variable Unbalances Randomization



Note: Expected distributions of an unmeasured confounder u for control (left panel) and treatment groups (right panel) when the population is selected based on posttreatment criteria ($x = 0$) under the data-generating process in Equation (7)are shown. We assume $\alpha_x = 0$, $c > 0$, $\gamma > 0$, and that u is distributed normally.

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 posttreatment (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 posttreatment 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 posttreatment variables are included in models is to try to address complex questions about causal mechanisms (e.g., mediation). For example, Corazzini et al. (2014) study 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 posttreatment variables as covariates for any of these reasons can bias estimates by creating imbalance with respect to the unmeasured confounder.[12] 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

[12]To simplify exposition, we focus here only on the bias resulting from the imbalance in u induced by controlling for the posttreatment 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.).

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 posttreatment variable x, which is political interest in our example. Returning to Equation (2) and employing basic rules of probability, we get

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

Note that this bias is simply a weighted combination of the exact same biases shown in Equation (8) and Equation (9), where the weights reflect the marginal probabilities of being either high- or low-interest students. Intuitively, this result shows that controlling for a posttreatment 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.

How Posttreatment Bias Can Contaminate Real-World Data

Analysis: An Original Study of Judge Perceptions

We further demonstrate the pernicious effects of posttreatment bias with a simple experiment on cue taking in judicial opinion conducted among 1,234 participants recruited from Amazon Mechanical Turk.[13] The study, which was conducted 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.[14] 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 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 4-point scale, which serves as our outcome variable. They were then also asked to evaluate her ideology on a 7-point scale ranging 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.[15] 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.

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 pretreatment) and the respondents' placement of Eid on the same scale (measured posttreatment). 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 posttreatment bias can distort treatment effect estimates. When

[13] 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).

[14] See the supporting information for the full instrument.

[15] 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.

TABLE 2 Endorsement Effect on Retention Vote Conditioning on Ideological Distance

	Full Sample		|Distance| ≤ 1	|Distance| > 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	1,182	1,178	504	674

Note: Outcome variable is a 4-point measure of the likelihood of voting to retain Eid. Ideological distance = |self-reported ideology – perception of Eid's ideology|.
*p < .01.

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 (≤ 1 point on the 7-point ideology scale) or not. The sign of the estimated treatment effect *reverses* in the subsample of respondents who perceive themselves as being close to Eid, becoming positive (.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 (−.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 the subsection "Practices That Lead to Posttreatment Bias," conditioning on ideological distance actually unbalances the sample by respondents' self-reported ideology *even though self-reported ideology is measured pretreatment.* For instance, among respondents who perceive themselves as ideologically close to Eid, treatment group respondents are significantly more conservative on our 7-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 posttreatment variable, we have unbalanced the treatment and control groups in terms of ideology and unmeasured confounders and thus biased the treatment effect estimate.

Reanalysis: Dickson, Gordon, and Huber (2015)

To further illustrate the consequences that posttreatment practices may have on real-world inferences, we replicate and reanalyze Dickson, Gordon, and Huber (2015; henceforth 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 want 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 observed 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 posttreatment bias.[16] First, DGH exclude cases of so-called "perverse" targeting of a contributor when at least one citizen did not contribute (2015, 119). Intuitively, dropping these cases might seem to allow them to focus on treatment effects among individuals who correctly understood the incentives. However, perverse targeting is a posttreatment 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 (2015, 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.[17]

Table 3 demonstrates that posttreatment conditioning induces substantial differences in the estimated effects of DGH's treatments.[18] The first column, which omits any posttreatment 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 posttreatment 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 posttreatment bias for both the limited information effect in the salary condition and the appropriation effect in the limited information condition. Dickson, Gordon, and Huber (2015) note that both models are sensitive to model specification; our analysis suggests that these results may be attributable to posttreatment bias.[19]

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 posttreatment 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 posttreatment conditioning below—noncompliance, 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.[20]

Use Pretreatment 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

[16] 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" (2015, 110).

[17] We show that these variables were affected by the treatment assignment in the supporting information.

[18] 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 supporting information for full model results.

[19] See the supporting information for further analysis of Dickson, Gordon, and Huber (2015) and an additional demonstration of posttreatment bias using data from Broockman and Butler (2017).

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

TABLE 3 Treatment Effect Differences by Posttreatment Conditioning

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

Note: Data are 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.
*p < .1; **p < .05; ***p < .01.

treatment variable T with a potential moderator x. However, as we note above, this design risks posttreatment bias if the moderator could be affected by the experimental manipulation. Moderators that are vulnerable to treatment spillovers like racial resentment should be measured pretreatment (see, e.g., Huber and Lapinski 2006, 424).[21]

Finally, scholars often wish to use measures of respondent attention (separate from manipulation checks) to drop inattentive respondents (e.g., Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009). All attention checks should be collected before the experimental randomization to avoid posttreatment 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 posttreatment attention checks is the equivalent of selecting on a posttreatment covariate and would again risk bias.

Use Instrumental Variables to Address Noncompliance

One frequent problem in experiments is noncompliance. Participants frequently fail to receive the assigned treatment due to logistical problems, failure to understand experimental 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 noncompliance" 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 noncompliance 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).

[21] 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.

Use Double Sampling, Extreme Value Bounds, or Instruments to Account for Attrition

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

There are several approaches that aim to better estimate treatment effects in the presence of nonrandom 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.

Understand the Costs of Mediation Analysis

Some researchers include posttreatment 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*. Thus, 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 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 posttreatment 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 posttreatment 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).

The Inadequacy of Empirical Tests for Posttreatment Bias

Finally, it is important to note that posttreatment bias cannot be easily diagnosed or remedied empirically. A common belief apparent in the literature is that researchers can rule out posttreatment 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 posttreatment bias in analyses that condition on x. First, even in the simplified examples presented above, posttreatment 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 posttreatment 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 posttreatment variable. In the end, the best solution is not to test for posttreatment bias but rather to carefully design experimental protocols that prevent it in the first place.

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 posttreatment variables. We find that a significant fraction of the experimental studies published in the discipline's most prestigious journals drop observations based on posttreatment variables or control for posttreatment variables in their statistical analysis. These practices are typically employed in an effort to address practical problems like noncompliance or to try to answer difficult inferential questions such as identifying causal mechanisms. Though these intentions are laudable, we demonstrate that posttreatment 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 pretreatment 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 posttreatment 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 pretreatment 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 posttreatment 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 to 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 posttreatment 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 posttreatment 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 posttreatment 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* 110(3): 512–29.

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): 444–55.

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): 710–13.

Arceneaux, Kevin. 2012. "Cognitive Biases and the Strength of Political Arguments." *American Journal of Political Science* 56(2): 271–85.

Athey, Susan, and Guido Imbens. 2017. "The Econometrics of Randomized Experiments." *Handbook of Economic Field Experiments* 1: 73–140.

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): 1173–82.

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): 739–53.

Blackwell, Matthew. 2013. "A Framework for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57(2): 504–20.

Broockman, David E., and Daniel M. Butler. 2017. "The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication." *American Journal of Political Science* 61(1): 208–21.

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): 550–58.

Burnett, Craig M., and Lydia Tiede. 2015. "Party Labels and Vote Choice in Judicial Elections." *American Politics Research* 43(2): 232–54.

Clark, Tom S., and Jonathan P. Kastellec. 2015. "Source Cues and Public Support for the Supreme Court." *American Politics Research* 43(3): 504–35.

Coppock, Alexander. 2017. "Comment on White, Nathan, and Faller (2015)." 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(2): 188–206.

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): 579–92.

Dean Knox, Teppei Yamamoto, Matthew A. Baum, and Adam Berinsky. Forthcoming. "Design, Identification, and Inference in Patient Preference Trials." Working paper http://web.mit.edu/teppei/www/research/MediaChoiceMethod.pdf.

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): 109–27.

Druckman, James N., Jordan Fein, and Thomas J. Leeper. 2012. "A Source of Bias in Public Opinion Stability." *American Political Science Review* 106(2): 430–54.

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 Multi-Level/Hierarchical Models.* New York: Cambridge University Press.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* New York: Norton.

Großer, Jens, Ernesto Reuben, and Agnieszka Tymula. 2013. "Political Quid Pro Quo Agreements: An Experimental Study." *American Journal of Political Science* 57(3): 582–97.

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): 669–87.

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): 421–40.

Huber, Martin. 2012. "Identification of Average Treatment Effects in Social Experiments under Alternative Forms of Attrition." *Journal of Educational and Behavioral Statistics* 37(3): 443–74.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15(4): 309–34.

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.

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): 131–59.

Kosloff, Spee, Jeff Greenberg, Toni Schmader, Mark Dechesne, and David Weise. 2010. "Smearing the Opposition: Implicit and Explicit Stigmatization of the 2008 U.S. Presidential Candidates and the Current U.S. President" *Journal of Experimental Psychology: General* 139(3): 383–98.

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): 535–52.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *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(4): 762–86.

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources* 24(3): 343–60.

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): 867–72.

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

Peter, M. Aronow, Jonathon, Baron and Lauren, Pin-son. Forthcoming. "A Note on Dropping Experimental Subjects who Fail a Manipulation Check." *Political Analysis.*

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. Clark Glymour and Gregory F. Cooper. Menlo Park, CA, and Cambridge, MA: AAAI Press/MIT Press, 349–405.

Robins, James M., Miguel Angel Hernan, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5): 550–60.

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): 656–66.

Transue, John E., Daniel J. Lee, and John H. Aldrich. 2009. "Treatment Spillover Effects across Survey Experiments." *Political Analysis* 17(2): 143–61.

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): 1026–28.

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.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

- Coding of articles
- Mathematical expressions and illustrations of post-treatment bias
- Simulation evidence of post-treatment bias
- Additional reanalysis of Dickson, Gordon, and Huber (2015)
- Reanalysis of Broockman and Butler (2015)
- Judge experiment questionnaire