Political Analysis (2025), **33**, 361–377
doi:10.1017/pan.2025.3

ARTICLE

Priming Bias Versus Post-Treatment Bias in Experimental Designs

Matthew Blackwell[1], Jacob R. Brown[2], Sophie Hill[3], Kosuke Imai[4] and Teppei Yamamoto[5]

[1]Associate Professor, Department of Government, Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA; [2]Assistant Professor, Department of Political Science, Boston University, Boston, MA, USA; [3]PhD Student, Department of Government, Harvard University, Cambridge, MA, USA; [4]Professor, Department of Government and Department of Statistics, Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA; [5]Professor, Faculty of Political Science and Economics, Waseda University, Tokyo, Japan

Corresponding author: Matthew Blackwell; Email: mblackwell@gov.harvard.edu

(Received 30 July 2024; revised 15 October 2024; accepted 22 November 2024; published online 21 March 2025)

Abstract

Conditioning on variables affected by treatment can induce post-treatment bias when estimating causal effects. Although this suggests that researchers should measure potential moderators before administering the treatment in an experiment, doing so may also bias causal effect estimation if the covariate measurement primes respondents to react differently to the treatment. This paper formally analyzes this trade-off between post-treatment and priming biases in three experimental designs that vary when moderators are measured: pre-treatment, post-treatment, or a randomized choice between the two. We derive nonparametric bounds for interactions between the treatment and the moderator under each design and show how to use substantive assumptions to narrow these bounds. These bounds allow researchers to assess the sensitivity of their empirical findings to priming and post-treatment bias. We then apply the proposed methodology to a survey experiment on electoral messaging.

Keywords: bounds; interactions; heterogeneous effects; measurement; moderation; sensitivity analysis

Edited by: Daniel J. Hopkins and Brandon M. Stewart

1. Introduction

Ascertaining heterogeneous treatment effects is an integral part of many survey experiments. Researchers are often interested in how treatment effects vary across respondents with different characteristics. For example, we may be interested both in how implicit versus explicit racial cues affect support for a particular policy but also in how those effects differ by levels of racial resentment (Valentino, Hutchings, and White 2002). Or we may want to know whether the effect of land-based electoral appeals might depend on the voters' sense of land security (Horowitz and Klaus 2020). These questions of effect heterogeneity allow researchers to explore potential causal mechanisms and design more targeted and effective future treatments.

To examine such treatment effect heterogeneity, we must measure the relevant covariates, such as racial resentment or land security in the aforementioned examples, at some point during the survey experiment. The question of *when* we measure these moderators, however, is a source of methodological debate. On the one hand, a long tradition in political science has recognized the potential *priming bias* of a *pre-test design*, where covariates are measured prior to treatment (e.g., Klar 2013; Klar, Leeper, and Robison 2020; Morris, Carranza, and Fox 2008; Schiff, Montagnes, and Peskowitz 2022; Transue 2007). For example, asking a respondent about their party identification might lead them to evaluate

the treatment in a more partisan or political light, resulting in biased causal effect estimates. Several studies have documented priming effects from a range of different covariates (see Klar *et al.* 2020, for a review). Some find that certain priming effects can last for weeks (Chong and Druckman 2010).

On the other hand, the practice of measuring moderators after treatment, what we call a *post-test design*, has come under scrutiny due to the possibility for *post-treatment bias* (Acharya, Blackwell, and Sen 2016; Montgomery, Nyhan, and Torres 2018; Rosenbaum 1984). In particular, if covariates are affected by the treatment, then conditioning on those covariates—as required when assessing effect heterogeneity—can bias the estimation of conditional average treatment effect and any interactions that compare such effects. In our empirical application, the key moderating variable of land insecurity is a subjective, perceived measure and thus potentially affected by the framing of a political appeal around land rights. Though treatment is unlikely to affect measurements of many moderators like basic demographics, researchers investigating manipulable moderators face a dilemma about when to measure these covariates when designing an experiment.

In this paper, we apply the nonparametric identification and bounding approach of Manski (1995) and Balke and Pearl (1997) to formally analyze the trade-off between priming and post-treatment biases under different experimental designs, and we propose principled ways to analyze data (see Imai and Yamamoto 2010, for a similar analysis of measurement error).

We begin by deriving nonparametric bounds to show that neither the pre-test nor post-test design provides much information about conditional average treatment effects or interactions without additional assumptions. Next, we show how three potentially plausible assumptions can narrow the bounds. The first is *priming monotoncity*, which assumes that whether the moderator is measured pre- or post-treatment only affects the outcome in one direction. The second is *moderator monotonicity*, which assumes that measuring the covariates after treatment can move the value of that moderator only in one direction. The third assumption is *stable moderator under control*, whereby the covariate under the control condition cannot be affected by the timing of treatment. None of these assumptions can point identify the interaction between the treatment and a moderator, but they can substantially narrow the bounds and sometimes be informative about the sign of such an interaction.

To further sharpen our inference, we generalize the two standard experimental designs and consider a *randomized placement design*, where the experimenter randomly assigns respondents to either the pre-test or post-test design. We demonstrate how to estimate nonparametric bounds in this context and how to incorporate assumptions that connect the pre-test and post-test arms to further narrow the bounds. In Supplementary Material, we also develop a parametric Bayesian approach to incorporate pre-treatment covariates in the analysis to sharpen our inferences and quantify estimation uncertainty.

We also derive sensitivity analysis procedures for all three designs. In these analyses, we vary the proportion of respondents whose outcomes or moderators are affected by when the moderator is measured and assess how the bounds change as a function of this sensitivity parameter. This procedure allows researchers and readers to gauge the credibility of an estimated interaction in light of a more nuanced set of assumptions, rather than the blunt instruments of the monotonicity and stability.

Several recent studies have empirically explored the trade-off between priming and post-treatment bias. Albertson and Jessee (2023) find that the effect of moderator placement has little effect on the estimated interaction in a question-wording experiment. Furthermore, they find no evidence for an average effect of treatment on the moderator, potentially reducing concerns about post-treatment bias. Sheagley and Clifford (2025) compared several experiments when the moderators were measured just prior to treatment or in a prior survey wave. The authors found that estimated effects and interactions were similar across these conditions and concluded that priming bias may not be a widespread concern for experimental studies in political science.

Both of these papers present compelling evidence for the specific experiments conducted in their empirical assessments, but as Sheagley and Clifford (2025) warn us, "we should be cautious in generalizing [these] findings to the wide variety of studies run by political scientists." Our approach, on the other hand, provides a general methodological toolkit that can be applied to any experimental design and allows researchers to include substantive assumptions to tailor the framework to their applications.

The rest of the paper proceeds as follows. We first introduce a motivating empirical example of how land insecurity moderates the effectiveness of land-based appeals by politicians from Horowitz and Klaus (2020). We next describe the notation and basic assumptions of the pre-test and post-test designs. We then derive the sharp nonparametric bounds and sensitivity analyses for the pre-test, post-test, and randomized placement designs. Next, we apply the proposed methods to the empirical example. Finally, we conclude by suggesting directions for future research.

2. Motivating Example

We illustrate the trade-off between the pre- and post-treatment measurement of a moderator using a survey experiment conducted by Horowitz and Klaus (2020). This study investigates the effectiveness of land-based appeals to increase a politician's electoral support in Kenya's Rift Valley. It focuses on a region where contemporary ethnic divisions are, in part, a function of long-standing conflicts over land. The authors conduct a survey experiment that tests the effect of candidate appeals about land ownership and related ethnic grievances on electoral support. Their finding shows little overall effect of such appeals on electoral support. The study also examines heterogeneity by land insecurity, polling respondents on how precarious they feel their ownership may be over said land. The authors argue that respondents who express greater land insecurity would be more likely to see material gain from electing a candidate who makes land ownership appeals. They find that land insecurity is positively associated with the treatment effect.

In our analysis, we focus on two randomly assigned conditions: a control condition in which participants heard a generic campaign speech with no direct reference to the land issue, and a treatment condition that additionally referenced the land issue.[1] The outcome is the participant's reported likelihood of supporting the candidate, which we dichotomize (likely to support the candidate versus not) to illustrate our proposed methods.

One of the main hypotheses tested by Horowitz and Klaus (2020) is whether individuals who are personally experiencing land insecurity are more responsive to land-based appeals by politicians. While the evidence from the full sample is inconclusive, among respondents belonging to the "insider" ethnic group (Kalenjins), there is a positive and statistically significant interaction between the treatment condition and a dummy variable for land insecurity, in line with the authors' expectations.

Both types of bias could be of concern in this context, as land insecurity is measured by asking respondents to rate the security of their land rights. Asking this question *before* treatment may raise the salience of the land issue and thus attenuate any treatment effect of hearing a land-based appeal by a politician, a form of priming bias. Conversely, if the experimenter asks respondents about land security *after* treatment, then their responses may be affected by the content of the speech, yielding post-treatment bias.

To address these concerns, Horowitz and Klaus (2020) use (what we call) the *randomized placement design*, in which questions relating to the respondent's land rights were randomly assigned to occur before or after the treatment. Figure 1 shows the "naïve" estimates of the treatment-moderator interaction, which are all positive and substantively large, implying that the effect of a land-based appeal on electoral support is 25–50 percentage points larger for land-insecure respondents compared land-secure respondents. However, these estimates only reach conventional levels of statistical significance with the increased statistical power of the combined pre/post sample.

What can we learn about the true interaction from these estimates? We show in the following sections that the observed data often tells us very little about interactions without strong assumption. In Section 6, we return to this empirical example to illustrate how researchers can assess the possible impact of priming bias and post-treatment bias on their substantive findings using our proposed methods.

[1] To preserve statistical power, we focus on a treatment group that combines two conditions: one stating the land issues is their top priority and one that additionally reference an ethnic grievance.

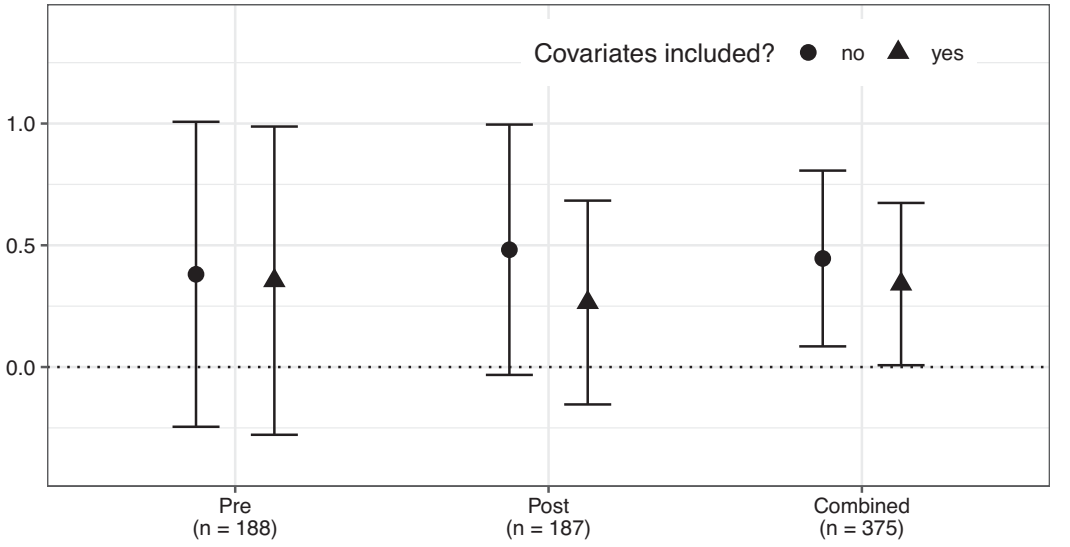

Figure 1. Estimates of treatment-moderator interaction using pre-test, post-test, and combined data from Horowitz and Klaus (2020), with and without covariates (age, gender, education, and closeness to own ethnic group). Error bars indicate 95% confidence intervals.

3. Experimental Designs and Causal Quantities of Interest

We now lay out the formal notation of our setting and describe the causal quantities of interest. Let T_i represent the binary treatment variable for unit $i \in \{1,\ldots,n\}$, indicating a certain experimental manipulation received by the unit. We use Y_i to denote a binary outcome of interest, and M_i to represent an observed binary moderator of interest.[2] Our goal is to develop a methodology that can be used to understand how the average treatment effect of T_i on Y_i varies as a function of M_i.

In this paper, we consider three experimental designs for this goal: the pre-test design, the post-test design, and the randomized placement design. These three designs differ in the timing of measurement for the potential moderators. In the pre-test design, the experimenter measures the moderator, M_i, before treatment assignment, ensuring that these measurements are unaffected by treatment. In the post-test design, the experimenter measures moderators after treatment assignment. Let Z_i be an indicator for measuring M_i before ($Z_i = 0$) or after treatment ($Z_i = 1$).

We will analyze these experimental designs using the potential outcomes framework (e.g., Holland 1986). Let $Y_i(t,z)$ represent the potential outcome with respect to the treatment status t and the timing of covariate measurement, z. In our empirical application, for example, $Y_i(1,1)$ is whether respondent i would support the hypothetical candidate if they were given the speech with referencing the land issue (that is, treatment, $t = 1$) and we measured the land security moderator *after* treatment ($z = 1$). We make the consistency assumption for the potential outcomes, such that $Y_i = Y_i(T_i,Z_i)$.

Our goal is to estimate treatment effects on the outcome free of priming bias, which requires us to measure moderators *after* treatment. To this end, we define the "true" potential outcomes of interest as $Y_i^*(t) \equiv Y_i(t,1)$ and let $Y_i(t) \equiv Y_i(t,0)$ to be a possibly mismeasured proxy observed under the pre-test design. Priming bias can occur when the pre-test and post-test measurements of the outcome differ, $Y_i(t) \neq Y_i^*(t)$.[3]

Similarly, the measures of the moderator, self-reported land security, can be affected by the when it is measured and the land grievance framing treatment. Let $M_i(t,z)$ be the potential value of the

[2] We focus on the binary setting to ease exposition, though many of the methods could be extended to more general cases.

[3] While we follow the literature and refer to the difference between the two potential outcomes, i.e., $Y_i(t)$ and $Y_i^*(t)$, as priming "bias," some researchers may be interested in $Y_i(t)$ even though this alternative outcome may be affected by the measurement of moderator.

moderator that would be observed for unit i when the treatment is set to t and the variable is measured in design z. Under the pre-test design, the treatment cannot affect the moderator, so we refer to $M_i(0,0) = M_i(1,0) \equiv M_i^*$ as the "true" moderator for unit i. On the other hand, under the post-test design, we let $M_i(t) \equiv M_i(t,1)$ to be potentially mismeasured proxy for that true moderator. Under those designs, the treatment can affect respondents' moderator, so that $M_i(0) \neq M_i(1)$, which can lead to post-treatment bias.

We can explore heterogeneous treatment effects by estimating the following quantities of interest,

$$\tau(m) \equiv \mathbb{E}(Y_i^*(1) - Y_i^*(0) \mid M_i^* = m), \tag{1}$$

$$\delta \equiv \tau(1) - \tau(0), \tag{2}$$

where $m \in \{0,1\}$. The first quantity characterizes the post-test conditional average treatment effect (CATE) as a function of the pre-test value of the moderator. This CATE is the effect of land-based appeals on support for a politician (when unprimed) conditional on a level of true land insecurity. The second quantity of interest compares the CATE between two subpopulations with different levels of the pre-test moderator, which is often called an *interaction effect*. These two quantities of interest formalize the dilemma about the choice of pre-test versus post-test experimental designs. Identifying $\tau(m)$ requires observing the true potential outcomes of interest ($Y_i^*(t)$) and true moderator (M_i^*) for the same unit, but this can never occur because the former requires $Z_i = 1$ and the latter requires $Z_i = 0$.

While the pre-test design allows us to observe the true moderator, it may suffer from priming bias because asking questions about the moderator might change the causal effect of the treatment by cueing respondents. Thus, under the pre-test design, neither $\tau(m)$ nor δ nor even the ATE, $\mathbb{E}(Y_i^*(1) - Y_i^*(0))$, is identified. In contrast, under the post-test design, the ATE can be estimated without bias, and yet this design may result in post-test bias for $\tau(m)$ and δ when the treatment affects the moderator.

4. Identification Analysis under the Three Experimental Designs

We now show what we can learn from each of the three possible experimental designs. Unfortunately, our analysis demonstrates that none of the three designs are informative about the sign of the interaction effect without further assumptions. For each design, however, we derive sharp bounds for the interaction effect and show how to narrow these bounds with additional substantive assumptions. We also develop a sensitivity analysis procedure that allows researchers to vary the strengths of such assumptions and assess their implications.

4.1. Pre-test Design

We first investigate the identifying power of the pre-test design, where we observe (Y_i, M_i, T_i) among the units for whom $Z_i = 0$. We begin by formalizing the randomization of treatment in this setting.

Assumption 1 (Pre-test Randomization).

$$\{Y_i(t,z), M_i(t,1)\} \perp\!\!\!\perp T_i \mid M_i^* = m, Z_i = 0, \tag{3}$$

for $t = 0,1$, $m \in \{0,1\}$.

The assumption allows for the use of stratified randomization based on the pre-treatment moderator. It is straightforward to show that this assumption also holds even when randomization is done without conditioning on the pre-treatment moderator.

By computing the difference in the average observed outcomes between different treatment groups, researchers can identify the following quantity,

$$\begin{aligned}\tau_{pre}(m) &\equiv \mathbb{E}(Y_i \mid T_i = 1, M_i = m, Z_i = 0) - \mathbb{E}(Y_i \mid T_i = 0, M_i = m, Z_i = 0) \\ &= \mathbb{E}(Y_i(1) - Y_i(0) \mid M_i^* = m),\end{aligned} \tag{4}$$

where the equality follows from Assumption 1 and the consistency assumption. Unfortunately, equation (4) does not generally equal the true (i.e., unprimed) CATE, $\tau(m)$. The key problem is that the pre-test design gives us information about the correct values of the moderator M_i^*, but provides no information about the outcomes we want to investigate, $Y_i^*(t)$. In particular, the conditional distribution of $Y_i(t)$ given $M_i^* = m$ may differ from that of $Y_i^*(t)$ given $M_i^* = m$, since asking questions about the moderator before measuring outcome may influence subsequent responses. Thus, $\tau_{pre}(m)$ represents the CATE of the land grievance treatment on the *primed* support for the candidate. Similarly, δ is not generally equal to the unprimed interaction effect, $\delta_{pre} \equiv \tau_{pre}(1) - \tau_{pre}(0)$.

With no restrictions on the priming bias, we cannot rule out extreme possibilities, such as all respondents changing their values of Y_i in response to the priming. This could occur if asking about land security caused all supporters to oppose the hypothetical candidate and vice versa. While perhaps unlikely, this scenario is possible under the randomization assumption alone. Thus, the nonparametric sharp (i.e., shortest possible) bounds under the pre-test design remain identical to the logical bounds without any data,

$$\tau(m) \in [-1,\ 1], \tag{5}$$

$$\delta \in [-2,\ 2]. \tag{6}$$

In other words, the pre-test design is completely uninformative about the causal effects of interest unless one is willing to make additional assumptions about the joint distribution of $Y_i^*(t)$ and $Y_i(t)$.

We can derive an expression for the amount of priming bias as

$$\begin{aligned}\tau_{pre}(m) - \tau(m) = &\left\{\mathbb{P}(Y_i^*(1) = 0, Y_i(1) = 1 \mid M_i^* = m) - \mathbb{P}(Y_i^*(0) = 0, Y_i(0) = 1 \mid M_i^* = m)\right\} \\ &- \left\{\mathbb{P}(Y_i^*(1) = 1, Y_i(1) = 0 \mid M_i^* = m) - \mathbb{P}(Y_i^*(0) = 1, Y_i(0) = 0 \mid M_i^* = m)\right\},\end{aligned} \tag{7}$$

where we have suppressed the conditioning on $Z_i = 0$ to reduce notational burden. The first term of this bias for the CATE is the effect of treatment on probability of being positively primed (moving from $Y_i^* = 0$ to $Y_i = 1$) and the second term is the effect of treatment on being negatively primed (moving from $Y_i^* = 1$ to $Y_i = 0$). Positively primed individuals in our application are those who would become supportive of the candidate only when asked about land security prior to treatment. This bias is unidentifiable because it depends on the joint distribution of the unprimed $Y_i^*(t)$ and primed $Y_i(t)$ outcomes, but we only observe $Y_i(t)$ under the pre-test design.

Some scholars, such as Sheagley and Clifford (2025), have shown empirically that δ and δ_{pre} appear close to each other in several experiments where the moderator was measured in a prior survey wave (and so was less prone to priming bias). Should we conclude from this that future studies will not suffer from priming bias? Perhaps, but we should properly view this as an additional assumption rather than a result implied by the design of the experiment. In particular, we would have to assume that either there is no priming at all, treatment does not affect priming, or that the treatment effects on positively and negatively primed individuals cancel out. Below, we will see how to incorporate assumptions about limited priming that would allow a sensitivity analysis for pretest designs.

4.1.1. *Narrowing the Pre-test Bounds under Additional Assumptions*

What assumptions can we place on the pre-test design to narrow the bounds? Any such assumption would have to place restrictions on the joint distribution of the primed and unprimed outcomes. A common set of assumptions for this type of setting would be a *monotone treatment response* assumption for the priming effect (Manski 1997). In particular, we consider a priming monotonicity assumption, which states that the effect of asking the moderator before treatment can only move the outcome in a single direction.

Assumption 2 (Priming Monotonicity). $Y_i(t) \geq Y_i^*(t)$ *for all* $t = 0, 1$.

The assumption implies that the effect of moving from post-test ($Z_i = 1$) to pre-test ($Z_i = 0$), which we call the priming effect, can only increase the outcome. In our application, this assumption implies that asking about land rights before reading the campaign speech treatment increases support for the hypothetical candidate. As stated, this assumption holds across levels of treatment, though it is possible to assume the reverse direction ($Y_i(t) \leq Y_i^*(t)$) or even to have a different effect direction for each level of treatment. The assumption narrows the bounds on our quantities of interest, as stated in the following proposition.

Proposition 1 (Pre-test Sharp Bounds). *Let* $P_{tzm} = \mathbb{P}(Y_i = 1 \mid T_i = t, Z_i = z, M_i = m)$. *Under Assumptions 1 and 2, we have the following sharp bounds:*

$$\tau(m) \in [-P_{00m},\, P_{10m}], \tag{8}$$

and

$$\delta \in [-P_{100} - P_{001},\, P_{101} + P_{000}]. \tag{9}$$

This proposition implies that, although priming monotonicity narrows the bounds compared to the uninformative randomization bounds, the bounds still will always contain zero. Thus, without further assumptions, the pre-test design cannot rule out no CATE for any level of the moderator nor can it rule out no interaction.

4.1.2. Sensitivity Analysis for the Pre-test Design

Given the empirical results such as Sheagley and Clifford (2025) that point to a limited role of priming in some experimental settings, we now develop a sensitivity analysis for pre-test designs to determine how sensitive results are to deviations from the "no priming bias" assumption. In particular, we constrain the proportion of respondents primed, or more precisely, the proportion of respondents whose value of $Y_i^*(t)$ is different from $Y_i(t)$. Formally, we state this restriction as

$$\mathbb{P}(Y_i^*(t) = 1, Y_i(t) = 0 \mid M_i^* = m) + \mathbb{P}(Y_i^*(t) = 0, Y_i(t) = 1 \mid M_i^* = m) \;\leq\; \theta \tag{10}$$

for all $t, m \in \{0, 1\}$. Note that if $\theta = 0$, then we have $Y_i^*(t) = Y_i(t)$ for all i, and the data under the pre-test design alone identify the quantities of interest. By increasing the value of θ, a sensitivity analysis gradually restricts the severity of the priming effects in the empirical setting.

Proposition 2 (Pre-test Sharp Bounds under Restricted Priming Effects). *Suppose that Assumptions 1 and 2, and restriction (10) hold. Then, we have the following sharp bounds:*

$$\tau(m) \in [\tau_{pre}(m) - \theta, \tau_{pre}(m) + \theta], \qquad \delta \in [\delta_{pre} - 2\theta, \delta_{pre} + 2\theta].$$

This proposition states that if the proportion of primed respondents in each treatment arm is no larger than θ, then we can bound the true interaction with an interval of length 4θ around the naïve pre-test interaction (for the CATE, the interval width is exactly half, i.e., 2θ). These bounds can be informative (that is, they exclude 0) if $|\delta_{pre}| > 2\theta$. Thus, researchers can conduct a sensitivity analysis by setting θ to different values and see how the bounds change.

4.2. Post-test Design

Next, we consider the post-test design, where we observe (Y_i, M_i, T_i) among the units for whom $Z_i = 1$. The randomization of the treatment implies the following ignorability assumption.

Assumption 3 (Post-Test Randomization).

$$\{Y_i(t,z), M_i(t), M_i^*\} \perp\!\!\!\perp T_i \mid Z_i = 1,$$

for $t = 0, 1$.

How informative is Assumption 3 alone about our quantities of interest (i.e., $\tau(m)$ and δ) without making any other assumption? The standard CATE estimator would be unbiased for

$$\tau_{post}(m) \equiv P_{11m} - P_{01m} = \mathbb{E}(Y_i^*(1) \mid M_i(1) = m) - \mathbb{E}(Y_i^*(0) \mid M_i(0) = m), \tag{11}$$

where the equality follows from Assumption 3 and the fact that $\mathbb{P}(Z_i = 1) = 1$ for all i.

Under the post-test design, we observe the true (i.e., unprimed) potential outcome of interest $Y_i^*(t)$, and yet the moderator may be observed with error, i.e., $M_i(t) \neq M_i^*$. Similar to the case of the pre-test design, therefore, $\tau_{post}(m)$ does not generally equal $\tau(m)$ because the conditional distribution of $Y_i^*(t)$ given $M_i(t) = m$ may differ from that of $Y_i^*(t)$ given $M_i^* = m$. In fact, $\tau_{post}(m)$ may not equal $\tau(m)$ even if the effects of treatment and moderator measurement timing do not change the marginal distribution of the moderator, i.e., $\mathbb{P}(M_i(0) = m) = \mathbb{P}(M_i(1) = m) = \mathbb{P}(M_i^* = m)$ for all $m \in \mathcal{M}$. Restricting the *marginal* distribution of the moderator does not help because effect heterogeneity is a function of the *joint* distribution of the counterfactual outcomes and moderators. Thus, under the post-test design, neither $\tau(m)$ nor δ is nonparametrically identified.

Despite the lack of point identification, the design can contain some information about our quantities of interest. The following proposition derives the sharp randomization-only bounds under the post-test design.

Proposition 3 (No-assumption Post-test Sharp Bounds). *Let* $P_t = \mathbb{P}(Y_i = 1 \mid T_i = t, Z_i = 1)$. *Under Assumptions 3, we have* $\delta \in [\delta_L, \delta_U]$, *where*

$$\begin{aligned}
\delta_L &= -1 - \max\left\{\min\left(\frac{1-P_0}{P_1}, \frac{1-P_1}{P_0}\right), \min\left(\frac{P_1}{1-P_0}, \frac{P_0}{1-P_1}\right)\right\}, \\
\delta_U &= 1 + \min\left\{\max\left(\frac{1-P_0}{P_1}, \frac{1-P_1}{P_0}\right), \max\left(\frac{P_1}{1-P_0}, \frac{P_0}{1-P_1}\right)\right\}.
\end{aligned} \tag{12}$$

Furthermore, these bounds are sharp.

The proof of this result is given in Supplementary Material A.2 and relies on a standard linear programming approach often used in bounding causal quantities (e.g., Balke and Pearl 1997). Unfortunately, these bounds are often quite wide in practice. In particular, the bounds can never be narrower than $[-1, 1]$ since the post-test data are completely uninformative about the true moderator. As a result, the sharp bounds under the post-test design can be quite wide and sometimes cover the entire logical range, $[-2, 2]$, for δ. In sum, without additional assumptions, the post-test design can only provide limited information about the causal quantities of interest.

4.2.1. Narrowing the Post-test Bounds under Additional Assumptions

While the bounds only using the randomization can be too wide to be useful in practice, we may be willing to entertain other assumptions on the causal structure that will narrow the bounds. We rely on a principal stratification approach in which we stratify the units according to how their moderator values react to treatment and moderator measurement timing (Frangakis and Rubin 2002). Let $\mu_s(t) = \mathbb{P}(Y_i^*(t) = 1 \mid S_i = s)$, where S_i represents the principal strata defined by the moderator, $\{M_i(1), M_i(0), M_i^*\}$. Without making any assumption, S_i can take any of the 2^3 values in

$$\mathcal{S} = \{111, 011, 101, 001, 110, 010, 100, 000\}.$$

Let $\rho_s = \mathbb{P}(S_i = s)$ be the probability of a unit falling into one of the strata, such that $\sum_{s \in \mathcal{S}} \rho_s = 1$. Finally, we denote the marginal probability of the true pre-test moderator as $Q_* = \mathbb{P}(M_i^* = 1)$.

The first assumption we consider is that the effect of post-treatment measurement of the moderator has a *monotonic* effect on the moderator for every unit.

Assumption 4 (Moderator Monotonicity). $M_i(t) \geq M_i^*$ *for all* $t = 0,1$.

In the context of the motivating example, this assumption requires no unit to have, say, lower land security values if we ask about it after the respondent reads the politician's speech rather than before. While weaker than assuming there is no measurement error in the post-test moderator, the plausibility of this assumption will depend on the study context. In the post-test design, we cannot verify this assumption because we never observe M_i^*. The assumption rules out several possible principal strata, ensuring that S_i can only take one of the following values: $111, 110, 010, 100,$ or 000.[4] We now present the sharp bounds under this assumption.

Proposition 4 (Post-test Sharp Bounds under Monotonicity). *Let* $Q_{tz} = \mathbb{P}(M_i = 1 \mid T_i = t, Z_i = z)$. *Under Assumptions 3 and 4, we have sharp bounds* $\delta \in [\delta_{L2}, \delta_{U2}]$, *where*

$$\delta_{L2} = \frac{P_{111}Q_{11} - P_{011}Q_{01}}{Q_*} - \frac{P_{110}(1-Q_{11}) - P_{010}(1-Q_{01})}{1-Q_*} + \frac{\max\{P_{011}Q_{01} - Q_*, 0\} - \max\{P_{111}Q_{11}, Q_{11} - Q_*\}}{Q_*(1-Q_*)},$$

$$\delta_{U2} = \frac{P_{111}Q_{11} - P_{011}Q_{01}}{Q_*} - \frac{P_{110}(1-Q_{11}) - P_{010}(1-Q_{01})}{1-Q_*} + \frac{\min\{0, Q_* - P_{111}Q_{11}\} + \min\{P_{011}Q_{01}, Q_{01} - Q_*\}}{Q_*(1-Q_*)}.$$

We provide the derivation of these bounds (and those in the next proposition) in Supplementary Material A.3. These bounds will be narrower than the randomization bounds given in Proposition 3 for two reasons. First, with the randomization assumption alone, we could only leverage the observed strata within levels of treatment—further stratification in terms of the moderator provided no information because it placed no restriction on the relationship between the pre-test and post-test versions of the moderator. Under moderator monotonicity (Assumption 4), we can leverage P_{tzd} and Q_{tz} to narrow the bounds. Second, moderator monotonicity places bounds on the true value of Q_* since it must be less than $\min(Q_{11}, Q_{01})$.

While the moderator monotonicity assumption does narrow the bounds, they are often still quite wide and usually contain 0. To further narrow the bounds, we consider another assumption that the moderator is stable in the control arm of the study.

Assumption 5 (Stable Moderator under Control). $M_i^* = M_i(0)$.

This assumption implies that the moderator under control in the post-test design is the same as the moderator as if it was measured pre-test. This assumption may be plausible in experimental designs where the control condition is neutral or similar to the pre-test environment. In our empirical example, this would mean that hearing the generic campaign speech, which does not mention the land issue, does not affect perceived land insecurity. Under both Assumptions 4 and 5, the only values that principal strata that S_i can take are $\{111, 100, 000\}$, further narrowing the bounds as follows.

Proposition 5 (Post-test Sharp Bounds under Moderator Monotonicity and Stability). *Under Assumptions 3, 4, and 5, we have* $Q_* = Q_{01}$ *and sharp bounds* $\delta \in [\delta_{L3}, \delta_{U3}]$, *where*

$$\delta_{L3} = \frac{P_{111}Q_{11}}{Q_{01}} - P_{011} - \frac{P_{110}(1-Q_{11})}{1-Q_{01}} + P_{010} - \min\left\{1, \frac{P_{111}Q_{11}}{Q_{11}-Q_{01}}\right\} \cdot \frac{Q_{11}-Q_{01}}{Q_{01}(1-Q_{01})},$$

$$\delta_{U3} = \frac{P_{111}Q_{11}}{Q_{01}} - P_{011} - \frac{P_{110}(1-Q_{11})}{1-Q_{01}} + P_{010} - \max\left\{0, \frac{P_{111}Q_{11}-Q_{01}}{Q_{11}-Q_{01}}\right\} \cdot \frac{Q_{11}-Q_{01}}{Q_{01}(1-Q_{01})}.$$

[4] While we present a positive version of monotonicity for both treatment levels, it is possible to derive bounds under a negative version of the assumption or with differing directions for each treatment condition.

These bounds demonstrate how the magnitude of the treatment effect on the moderator affects identification in the post-test design. A unit's moderator is affected by treatment whenever $M_i(1) \neq M_i(0)$, which corresponds to the $S_i = 100$ principal stratum under monotonicity and stable moderator under control. Note that under these assumptions, ρ_{100} represents the magnitude of the treatment-moderator effect. Since $Q_{11} = \rho_{111} + \rho_{100}$ and $Q_{01} = \rho_{111}$, we can identify this effect with the usual difference in (population) means, $Q_{11} - Q_{01} = \rho_{100}$. The maximum possible width of the sharp bounds depends on this effect, with

$$\max\{\delta_{U3} - \delta_{L3}\} = \frac{Q_{11} - Q_{01}}{Q_{01}(1 - Q_{01})} = \frac{\rho_{100}}{\rho_{000}(\rho_{111} + \rho_{100})},$$

so that the bounds can be relatively informative if the post-treatment average effect on the moderator is small.

4.2.2. Sensitivity Analysis under Limited Effects on the Moderator

While the monotonicity and stable moderator assumptions can considerably narrow the nonparametric bounds on our causal quantities, they rule out entire principal strata, which may be stronger than is justified for a particular empirical setting. We now consider an alternative approach to bounds that does not rule out any particular principal strata but rather places restrictions on the proportion of units whose moderators are affected by treatment.

In particular, we propose a sensitivity analysis that limits the proportion of respondents whose moderator value changes between the pre-test and post-test, regardless of the treatment condition (contrast this with Assumption 5 which applies to the control condition only). We operationalize this via the following constraint,

$$\mathbb{P}(S_i \notin \{111, 000\}) \leq \gamma.$$

Note that γ must be greater than $|Q_{11} - Q_{01}|$ for the bounds to be feasible since $|Q_{11} - Q_{01}| = |\rho_{101} + \rho_{100} - \rho_{011} - \rho_{010}|$. We vary the value of γ from $|Q_{11} - Q_{01}|$ to 1 and see how the nonparametric bounds on the value of δ change as we gradually allow a larger treatment effect on the moderator. We obtain these new bounds by adding this additional constraint to linear program, which we can then solve numerically. This approach is a more flexible way to allow for limited heterogeneous treatment effects on the moderator in any direction. Researchers can also combine this sensitivity analysis with the monotonicity and stable moderator assumptions.

4.3. Randomized Placement Design

Finally, we examine a combined pre/post design called the randomized placement design, where in addition to treatment, the timing of moderator measurement, Z_i, is also randomized. Bounds from this design will improve upon the post-test bounds because we can identify the marginal probability of the true moderator as

$$Q_* \;=\; \mathbb{P}(M_i^* = 1) \;=\; \mathbb{P}(M_i = 1 \mid Z_i = 0).$$

Unfortunately, for the randomization-only bounds in Equation (12), the sign of δ cannot be identified for any value of $Q_* \in (0,1)$.

With the randomized placement design, we have a slightly more complicated set of principal strata since now we must handle both the pre-test and post-test potential outcomes. In particular, we define the following quantities that characterize the joint distribution of the pre-test and post-test potential outcomes for a given treatment level, t, and principal strata, s:

$$\psi_{y_1 y_0 s}(t) \;=\; \mathbb{P}(Y_i^*(t) = y_1, Y_i(t) = y_0 \mid S_i = s)$$

where $\psi_{y_1 y_0 s}(t) \geq 0$ and $\sum_{y_1} \sum_{y_0} \psi_{y_1 y_0 s}(t) = 1$ for all t. Furthermore, let $\mathcal{S}_m^* = \{s : s \in \mathcal{S}, M_i^* = m\}$ be the set of principal strata with the true value of the moderator equal to m. Then, we can write the interaction between the treatment and the moderator as

$$\delta = \sum_{y_0=0}^{1} \left\{ \sum_{s_1 \in \mathcal{S}_1^*} \frac{\rho_{s_1}}{Q_*} (\psi_{1 y_0 s_1}(1) - \psi_{1 y_0 s_1}(0)) - \sum_{s_0 \in \mathcal{S}_0^*} \frac{\rho_{s_0}}{1 - Q_*} (\psi_{1 y_0 s_0}(1) - \psi_{1 y_0 s_0}(0)) \right\},$$

and the observed strata in the pre-test and post-test arms as

$$P_{t1m} Q_{t1} = \mathbb{P}(Y_i = 1, D_i = m \mid T_i = t, Z_i = 1) = \sum_{y_0=0}^{1} \sum_{s \in \mathcal{S}(t,1,m)} \psi_{1 y_0 s}(t) \rho_s,$$

$$P_{t0 m_*} = \mathbb{P}(Y_i = 1 \mid T_i = t, Z_i = 0, M_i = m_*) = \frac{\sum_{y_1=0}^{1} \sum_{s \in \mathcal{S}_{m_*}^*} \psi_{y_1 1 s}(t) \rho_s}{\sum_{s \in \mathcal{S}_{m_*}^*} \rho_s},$$

for all values of y, m, t, and m_*.

Without further assumptions, the pre-test data are helpful only insofar as they identify the marginal distribution of the moderator, Q_*. We can narrow the bounds under the randomized placement design using all of the substantive assumptions described above: priming monotoncity, moderator monotonicity, and stability under control. Each of these implies restrictions on the above principal strata that can be incorporated into a linear programming problem, which we solve numerically to derive the bounds.

4.3.1. Sensitivity Analysis in the Randomized Placement Design

The randomized placement design allows us to combine the sensitivity analysis procedure of the pre- and post-test designs. In particular, we can impose both of the restrictions from above simultaneously:

$$\gamma \geq \mathbb{P}(S_i \notin \{111, 000\}),$$
$$\theta \geq \mathbb{P}(Y_i^*(t) = 1, Y_i(t) = 0 \mid M_i^* = m_*) + \Pr(Y_i^*(t) = 0, Y_i(t) = 1 \mid M_i^* = m_*).$$

Again, these conditions restrict (a) how much the treatment and moderator measurement timing affects the moderator, and (b) how much the moderator measurement timing affects the outcome. To incoprate these restrictions into the bounds, we rewrite them in the principal strata described above:

$$(1 - \rho_{111} - \rho_{000}) \leq \gamma,$$
$$\frac{\sum_{s \in \mathcal{S}_{m_*}^*} (\psi_{10s}(t) + \psi_{01s}(t)) \rho_s}{\sum_{s \in \mathcal{S}_{m_*}^*} \rho_s} \leq \theta.$$

Then, we can easily add these restrictions to the optimization problem that produces the bounds on the interaction effect.

We can conduct sensitivity analyses on the randomized placement design by varying the values of γ and θ and seeing how the values of the bounds change. There are several ways to conduct and present such a two-dimensional sensitivity analysis. One would be to plot the parameters on each axis and demarcate the regions where the bounds are informative (e.g., do not include zero) and where they are not. A second approach would be to choose a small value for one of the two parameters consistent with a researcher's beliefs. For instance, if a researcher believes that the moderator is unlikely to be affected by treatment, then they could choose a small value for γ and investigate the sensitivity of the bounds to different amounts of priming, as measured by θ.

5. Statistical Inference

The above bounds are stated in terms of population quantities when, in fact, we only ever have sample data. We can easily obtain estimates for the bounds by plugging in sample versions of these probabilities or, for the randomized placement design, solving the linear programming problem using the sample data. Obtaining valid confidence intervals in this setting is more challenging, since standard asymptotic analyses break down due to the maximum and minimum operators causing nondifferentiability. This problem can lead the standard nonparametric bootstrap to have problematic theoretical properties (Andrews and Han 2009; Fang and Santos 2019).

Furthermore, confidence intervals for the bounds tend to be overly conservative when the target of inference is the parameter δ rather than the bounds themselves. As pointed out by Imbens and Manski (2004), this occurs because the true parameter cannot simultaneously be close to the upper and lower bound at the same time. This fact allows us to narrow the confidence intervals by a data-driven amount while maintaining nominal coverage for the parameter of interest.

Our approach to inference combines the Imbens and Manski (2004) approach with estimated standard errors of the bounds from the the nonparametric bootstrap. While the bootstrap may have theoretical problems, we find in simulations that this approach produces conservative confidence intervals that have slightly higher-than-nominal empirical coverage. In Supplementary Material B, we provide further details of our approach to estimation. In Supplementary Material C, we also develop a Bayesian model-based approach to incorporate additional pre-treatment covariates that might be available to researchers.

6. Empirical Example

To illustrate how our approach can be applied to each of these designs, we return to the example from Horowitz and Klaus (2020) introduced in Section 2. Recall that the interaction in this case shows how the effect of land-based appeals varies by a respondent's level of land security.

6.1. Setup and Assumptions

First, it is worth discussing the assumptions beyond randomization in this context. Assumptions 4 (moderator monotonicity) and 5 (stability) are not guaranteed by the design and require a substantive justification. In this case, moderator monotonicity requires the placement of the land insecurity question after treatment shift perceived land insecurity in the same direction for all respondents (or have no effect). We assume a positive (or zero) individual effect on land insecurity, consistent with the estimate ATE—though it is substantively small and not statistically significant ($\hat{\beta} = 0.04$, $p = 0.23$). Given that one part of the active treatment was intended to induce fear of "land grabbing," it seems somewhat plausible that measuring perceived land rights after treatment would only increased feelings of insecurity.

The assumption of stability means that hearing the generic campaign speech (with no appeals to the land issue) has no individual-level effect on perceived land insecurity when it is measured post-treatment. Again, this assumption cannot be conclusively tested with the observed data since we cannot estimate individual-level effects. However, with the randomized placement design, we can estimate the ATE of the pre/post randomization on the moderator among respondents in the control group. Our estimate of the ATE is very small and not statistically significant ($\hat{\beta} = -0.005$, $p = 0.91$), which is at least consistent with the assumption of a stable moderator under control.

6.2. Comparing the Sharp Bounds across Different Assumptions

Figure 2 displays the non parametric bounds (with 95% confidence intervals) for δ under different sets of assumptions applied to the post-test data (in grey), the pre-test (in blue), and the combined randomized placement design (labeled "Prepost" in black). We also include the naïve OLS estimate

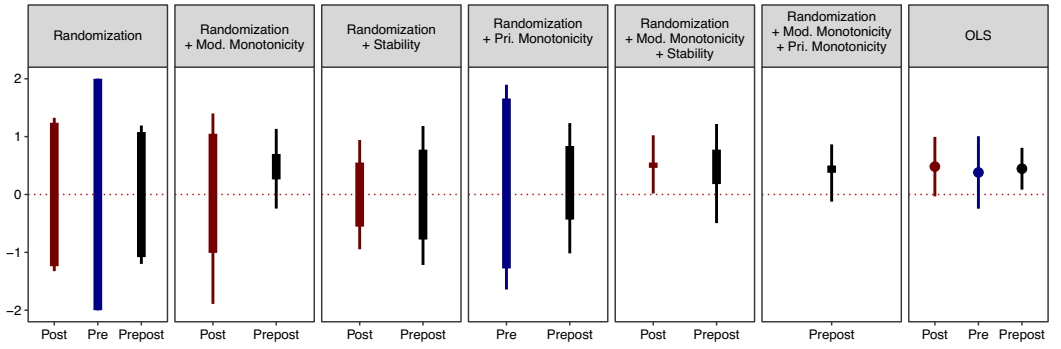

Figure 2. Estimated nonparametric bounds (thick bars) and 95% confidence intervals (thin bars) under different designs and assumptions. The final panel contains OLS point estimates and 95% confidence intervals.

with 95% confidence interval for comparison. Some of the designs are missing from panels because an assumption does not apply to that design (for example, moderator monotonicity under the pre-test design).

Assuming only randomization, the nonparametric bounds are uninformative of the sign of δ, and are much wider than the confidence interval of the naïve OLS estimate for all of the designs. Adding the assumption of moderator monotonicity reduces the width of the bounds, especially on the combined prepost data. Unfortunately, the confidence intervals become considerably larger, especially for the post-test data, in which the $M_i = 1$ group is small.

The assumption of stability tightens the bounds to a similar degree for the post-only and pre-post data. Under the pre-test assumptions of randomization, moderator monotonicity, and stability, the bounds exclude zero for both the post-test and randomized placement designs, though the confidence interval for the latter contains zero. Furthermore, the bounds for this design are wider under all three assumptions than under just randomization and monotonicity since, in finite samples, the pre-test mean of M_i differs from the post-test mean of M_i in the control arm. We would expect a difference by random chance even when stability holds, but the bounds may widen slightly to accommodate this divergence between the population and sample quantities.

Priming monotonicity also narrows the bounds for the pre-test and random placement designs, though not by as much as moderator monotonicity. The combination of the two monotonicity assumptions, however, recovers bounds that are close to the OLS estimate and has confidence intervals that barely contain zero.

In sum, the nonparametric bounds do not support the hypothesis of a positive interaction effect under the randomization assumptions alone. It is only with a combination of additional substantive assumptions—moderator monotonicity, priming monotonicity, and stability—that the sharp bounds produce results qualitatively similar to the naïve OLS estimate of a positive interaction effect, though the designs differ on whether this is statistically significant or not.

6.3. *Implementing the Sensitivity Analysis*

We now apply our sensitivity analysis procedures to this experiment. These procedures involve varying the proportion of respondents for whom the placement of the moderator measure affects their moderator value (land security) or their outcome (candidate support), labeled γ and θ respectively. We can apply the γ and θ sensitivity analyses to the post-test and pre-test designs, respectively, and we can combine them in the randomized placement design.

Figure 3 shows the post-test bounds as a function of γ under just the randomization assumption. While γ can theoretically range up to 1, here we limit it to 0.5 to aid presentation since the bounds quickly stabilize. The black lines denote the upper and lower bounds, and the shaded ribbon denotes the

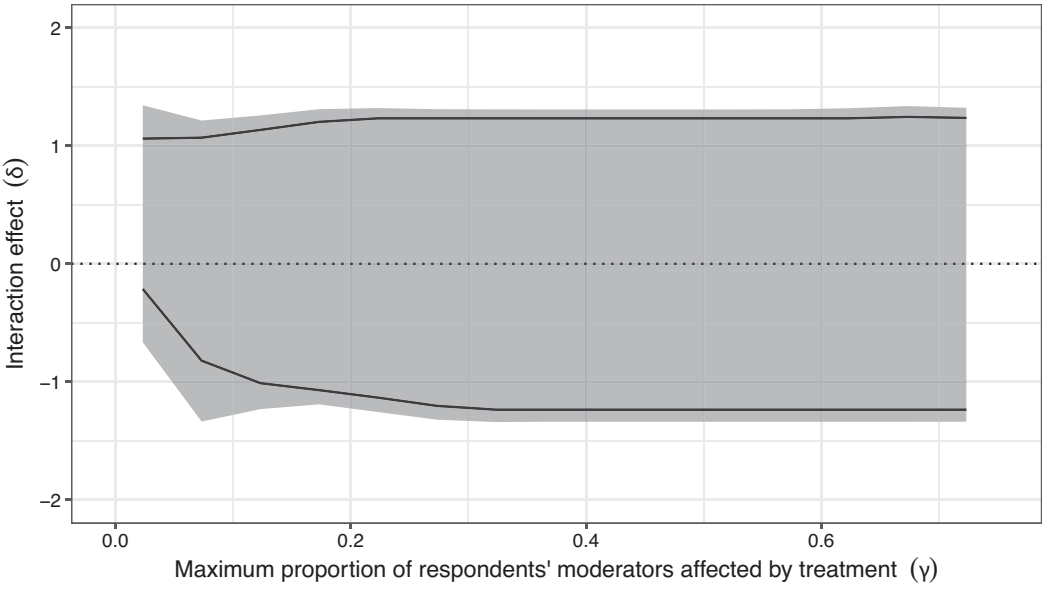

Figure 3. Post-test sensitivity analysis. Nonparametric bounds (black lines) with 95% confidence intervals (grey ribbon) as a function of γ, the proportion of respondents whose value of the moderator variable (land insecurity) is affected by post-test measurement.

95% confidence intervals around the bounds. The lower bound crosses 0 when $\gamma = 0.07$—that is, when no more than 7% of respondents are affected by the post-treatment measurement of the moderator. The 95% confidence interval contains 0 even for the minimum possible value of γ consistent with the observed data (0.02). The sensitivity analysis shows that the sharp bounds are highly sensitive to changes in the degree of post-treatment bias for small values of γ.

To interpret this sensitive analysis, researchers will need to draw on their substantive knowledge to assess the plausible range of γ. The estimated ATE on the post-treatment moderator is 0.02 ($p = 0.59$), but without a monotonicity assumption this may include respondents with positive and negative effects that offset their effects. We might assume that most of the effect of the land rights prime would accrue to respondents with less certain responses such as feeling "somewhat secure" in their land rights and that had been personally affected by prior ethnic conflict. Using the pre-test data, we find that this is 12% of the sample, which we might consider a reasonable value of γ. While it may seem like a relatively minor problem if only 12% of the sample is affected, our sensitivity analysis shows that the nonparametric bounds would be about eight times wider ($[-1.02, 1.14]$) compared to the case where γ is at its minimum ($[0.38, 0.64]$).

For priming bias in the pre-test design, we can assess sensitivity as a function of θ, the proportion of respondents whose outcome value is affected by when the moderator is measured. Figure 4 shows the bounds as a function of θ under priming monotonicity. When $\theta = 0$, we are assuming that priming bias does not exist so the true interaction is point identified in the pre-test arm, though the confidence interval at $\theta = 0$ already includes zero. The estimated lower bound crosses zero at approximately 0.2, when up to 20% of respondents are primed, which would be a rather large effect of priming.

Finally, in the randomized placement design, we combine these two tests by restricting both γ and θ. Figure 5 shows the prepost bounds as a function of γ under two different assumption about the amount of priming: limited priming $\theta \leq 0.25$ and unrestricted priming $\theta \leq 1$. In both settings, the bounds initially shrink as γ increases, most likely to due to the low values of γ being inconsistent with the data. The bounds then begin to widen, though they widen much more for the unrestricted priming compared the limited priming setting.

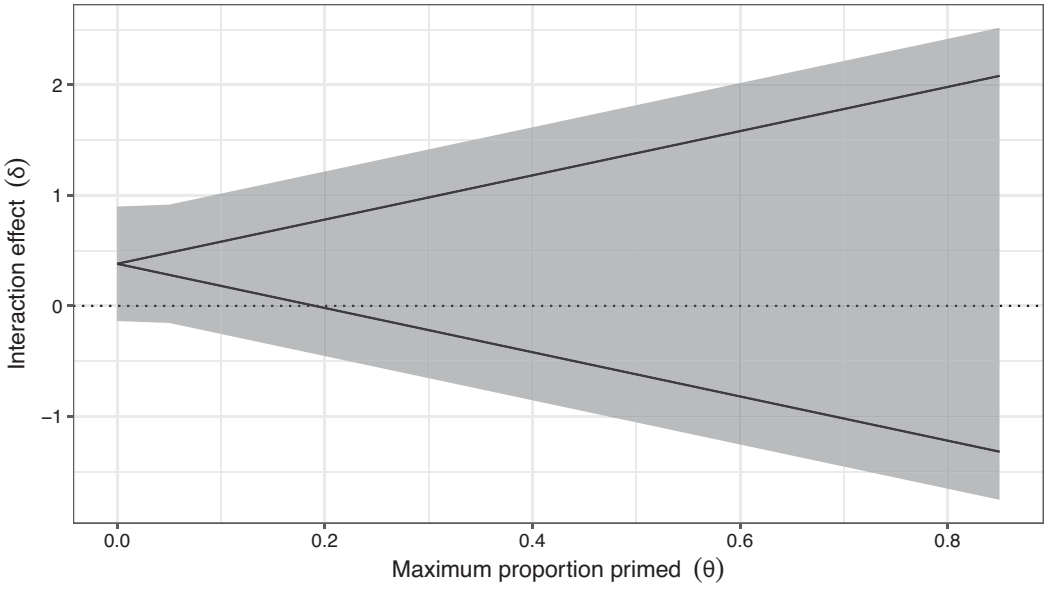

Figure 4. Pre-test sensitivity analysis. Nonparametric bounds (black lines) with 95% confidence intervals (grey ribbon) as a function of θ, the proportion of respondents who are primed by asking the moderator before treatment, under priming monotonicity assumption.

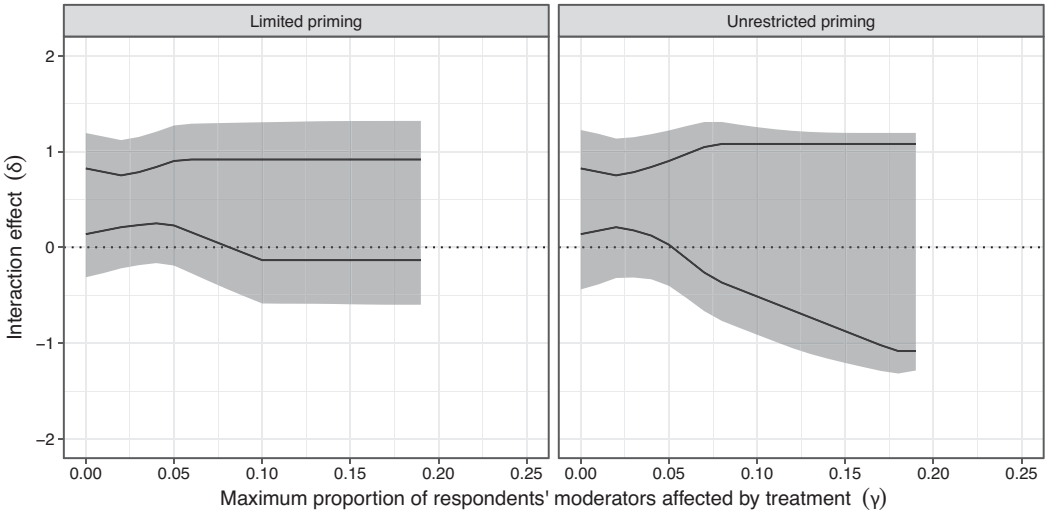

Figure 5. Randomized placement design sensitivity analysis. Nonparametric bounds (black lines) and 95% confidence intervals (grey ribbons) as a function of the post-test effect on the moderators and the amount of priming. The limited priming assumption assumes $\theta \leq 0.25$ and the unrestricted priming has $\theta \leq 1$.

7. Practical Guidance

What does all of this mean for the practice of experimental design? We recommend the following steps when attempting to estimate a CATE or interaction effect in an experimental context.

- **Tailor the experimental design to the target moderator.** Moderators that are fixed demographic traits about a respondent are less likely to be affected by treatment but might induce priming. Other moderators might work in reverse. Reasoning through the potential biases can provide guidance on which design to select. One can even adjust the relative size of the pre-test and post-test groups

in the randomized placement design to tailor the approach to address the most likely bias. Finally, researchers can use the randomized placement design in a pilot study to assess the potential average priming and post-treatment effects.

- **When possible, use indirect or distantly measured moderators.** Priming bias occurs because the act of measuring a moderator changes how a respondent reacts to treatment. We can safely ignore priming bias for indirectly measured moderators such as information about a respondent from an external database (e.g., voter files or census information about their local community). If sufficient resources are available, an earlier survey on the same respondents to measure the moderators might provide a long enough wash-out period to avoid priming bias, albeit with the possibility that the values of the moderators might change over the period. Similarly, many survey vendors provide basic demographic information on their panelists that was collected in earlier surveys.
- **Use the above bounds and sensitivity analyses to explore robustness.** Short of indirectly measured moderators, we can never fully rule out priming or post-treatment bias. The bounds and sensitivity analyses about can help researchers and readers understand when inferences are robust to small amounts of these biases. This transparency will help researchers understand the strengths and weaknesses of the scientific evidence.

8. Concluding Remarks

This paper addresses a central tension in survey methodology: how should researchers assess priming bias versus post-treatment bias when designing a survey experiment? The pre-test design avoids post-treatment bias but may suffer from priming bias. In contrast, the post-test design is free of priming bias and yet possibly leading to post-treatment bias. We conduct a formal analysis to show that neither design is informative about the moderation effect of interest without additional assumptions. We also analyze the randomized placement design, which is a mix of pre-test and post-test designs.

Our analysis derives sharp bounds for the moderation effect and shows how these bounds vary under additional substantive assumptions. We also provide sensitivity analyses for priming and post-treatment biases by varying the proportion of respondents whose moderator value changes in the post-test design and the proportion of respondents for whom the pre-test measurement of the moderator would prime their responses. We demonstrate how these tools can be used to diagnose and assess the severity of post-treatment bias and priming bias by applying them to a survey experiment regarding the effect of land-based appeals by politicians on electoral support in Kenya.

Open questions remain from our approach here. In particular, future work could optimize the randomized placement design to balance the priming and post-treatment bias concerns. In addition, we could consider how integrating separate pre-test surveys, often given weeks or months before treatment, might allow for a different set of plausible assumptions and identification. Lastly, the analytic approach used in this paper can be applied to other problems in survey experiments. For example, our bounds can be applied to cases where the causal moderation effects are estimated conditional on other pre-treatment covariates (Bansak 2021). Beyond moderation, it is also of interest to investigate how the use of attention and manipulation checks can affect the validity of causal inference in survey experiments using the proposed approach (Aronow, Baron, and Pinson 2019; Berinsky, Margolis, and Sances 2014; Kane and Barabas 2019; Varaine 2023).

Acknowledgments. Thanks to Dean Knox, Fredrick Sävje, and two anonymous reviewers from the Alexander and Diviya Magaro Peer Pre-Review at Harvard's Institute for Quantitative Social Science for helpful comments and feedback.

Competing Interests. The authors declare no competing interests.

Funding. Imai acknowledges financial support from the National Science Foundation (SES–0752050).

Data Availability Statement. Open source software to implement the method of this paper are included in the `prepost` R package, available at https://github.com/mattblackwell/prepost. Data and code to replicate all analyses can be found in the Dataverse replication repository at https://doi.org/10.7910/DVN/JZ55TF (Brown *et al.* 2024).

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2025.3.

References

Acharya, A., M. Blackwell, and M. Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110 (3): 512–529.

Albertson, B., and S. Jessee. 2023. "Moderator Placement in Survey Experiments: Racial Resentment and the "Welfare" versus "Assistance to the Poor" Question Wording Experiment." *Journal of Experimental Political Science* 10 (3): 448–454. https://doi.org/10.1017/XPS.2022.18

Andrews, D. W. K., and S. Han. 2009. "Invalidity of the Bootstrap and the m Out of n Bootstrap for Confidence Interval Endpoints Defined by Moment Inequalities." *The Econometrics Journal* 12: S172–S199.

Aronow, P. M., J. Baron, and L. Pinson. 2019. "A Note on Dropping Experimental Subjects Who Fail a Manipulation Check." *Political Analysis* 27 (4): 572–589. https://doi.org/10.1017/pan.2019.5

Balke, A., and J. Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association* 92: 1171–1176.

Bansak, K. 2021. "Estimating Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 184 (1): 65–86.

Berinsky, A. J., M. F. Margolis, and M. 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.

Brown, J., M. Blackwell, S. Hill, K. Imai, and T. Yamamoto. 2024. "Replication Data for: Priming Bias Versus Post-Treatment Bias in Experimental Designs." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/JZ55TF

Chong, D., and J. N. Druckman. 2010. "Dynamic Public Opinion: Communication Effects over Time." *American Political Science Review* 104 (4): 663–680.

Fang, Z., and A. Santos. 2019. "Inference on Directionally Differentiable Functions." *The Review of Economic Studies*. 86 (1): 377–412. https://doi.org/10.1093/restud/rdy049

Frangakis, C. E., and D. B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–29.

Holland, P. W. 1986. "Statistics and Causal Inference (with Discussion)." *Journal of the American Statistical Association* 81: 945–960.

Horowitz, J., and K. Klaus. 2020. "Can Politicians Exploit Ethnic Grievances? An Experimental Study of Land Appeals in Kenya." *Political Behavior* 42 (1): 35–58.

Imai, K., and T. Yamamoto. 2010. "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54 (2): 543–560.

Imbens, G. W., and C. F Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72 (6): 1845–1857.

Kane, J. V., and J. Barabas. 2019. "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science* 63 (1): 234–249.

Klar, S. 2013. "The Influence of Competing Identity Primes on Political Preferences." *The Journal of Politics* 75 (4): 1108–1124.

Klar, S., T. J. Leeper, and J. Robison. 2020. "Studying Identities with Experiments: Weighing the Risk of Post-Treatment Bias Against Priming Effects." *Journal of Experimental Political Science* 7 (1): 56–60.

Manski, C. F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C. F. 1997. "Monotone Treatment Response." *Econometrica* 65 (6): 1311–1334.

Montgomery, J. M., B. Nyhan, and M. Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62 (3): 760–775.

Morris, M. W., E. Carranza, and C. R. Fox. 2008. "Mistaken Identity: Activating Conservative Political Identities Induces "Conservative" Financial Decisions." *Psychological Science* 19 (11): 1154–1160.

Rosenbaum, P. R. 1984. "The Consquences 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.

Schiff, K. J., B . Pablo Montagnes, and Z. Peskowitz. 2022. "Priming Self-Reported Partisanship: Implications for Survey Design and Analysis." *Public Opinion Quarterly* 86 (3): 643–667.

Sheagley, G., and S. Clifford. 2025. "No Evidence that Measuring Moderators Alters Treatment Effects." American Journal of Political Science, 69: 49–63. https://doi.org/10.1111/ajps.12814

Transue, J. E. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51 (1): 78–91.

Valentino, N. A., V. L. Hutchings, and I. K. White. 2002. "Cues that Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96 (1): 75–90.

Varaine, S. 2023. "How Dropping Subjects Who Failed Manipulation Checks Can Bias Your Results: An Illustrative Case." *Journal of Experimental Political Science* 10 (2): 299–305.

Cite this article: Blackwell, M., Brown, J. R., Hill, S., Imai, K. and Yamamoto, T., (2025). Priming Bias Versus Post-Treatment Bias in Experimental Designs. *Political Analysis*, 361–377. https://doi.org/10.1017/pan.2025.3