Priming bias versus post-treatment bias in experimental designs*

Matthew Blackwell† Jacob R. Brown‡ Sophie Hill§
Kosuke Imai¶ Teppei Yamamoto‖

June 1, 2023

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 in 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 either source of bias. We extend the basic framework in two ways. First, we apply the framework to the case of post-treatment attention checks and bound how much inattentive respondents can attenuate estimated treatment effects. Second, we develop a parametric Bayesian approach to incorporate pre-treatment covariates in the analysis to sharpen our inferences and quantify estimation uncertainty. We apply these methods to a survey experiment on electoral messaging. We conclude with practical recommendations for scholars designing experiments.

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

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

†Associate Professor, Department of Government, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: mblackwell@gov.harvard.edu, URL: https://mattblackwell.org

‡Post-Doctoral Fellow, Center for the Study of Democratic Politics, Princeton University. 236 Corwin Hall, Princeton, NJ 08540. Email: jrbrown@princeton.edu, URL: https://jacobrbrown.com

§PhD Student, Department of Government, Harvard University. 1737 Cambridge Street, Cambridge MA 02138. Email: sophie_hill@g.harvard.edu, URL: sophie-e-hill.com

¶Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu URL: https://imai.fas.harvard.edu

‖Associate Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

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 (e.g., Valentino, Hutchings and White, 2002). Alternatively, we may want to know how an individual's land security affects how they respond to land-based appeals by political candidates (e.g., 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, at some point during the survey experiment. The question of *when* we measure these covariates, 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., Transue, 2007; Morris, Carranza and Fox, 2008; Klar, 2013; Klar, Leeper and Robison, Forthcoming; Schiff, Montagnes and Peskowitz, 2022). 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 a biased causal effect estimate. Several studies have documented priming effects from a range of different covariates (see Klar, Leeper and Robison, Forthcoming, for a review), and certain priming effects can last for weeks (Chong and Druckman, 2010).

On the other hand, the practice of measuring covariates after treatment, what we call a *post-test design*, has come under scrutiny due to the possibility for *post-treatment bias* (Rosenbaum, 1984; Acharya, Blackwell and Sen, 2016; Montgomery, Nyhan and Torres, 2018). In particular, if covariates are affected by the treatment, then conditioning on those covariates—as is typical when assessing effect heterogeneity—can bias the estimation of conditional average treatment effect and thus any interactions that compare such effects (though, see Albertson and Jessee, 2022, for a case with little evidence of post-treatment bias). Thus, researchers face a dilemma about when to

measure covariates when designing an experiment.

In this paper, we formally analyze the trade-off between priming and post-treatment biases and propose solutions to some of the difficulties. First, we derive nonparametric bounds to show that neither the pre-test nor post-test design provides much information about conditional average treatment effects without additional assumptions (see, e.g., Imai and Yamamoto, 2010, for a similar analysis of measurement error). Next, we show how two potentially plausible assumptions can narrow the bounds in the post-test design. The first is the *monotonicity of the post-test effect*, which assumes that measuring the covariates after treatment can move those covariates only in one direction. The second assumption is *stable moderator under control*, whereby the covariate under the control condition cannot be affected by the timing of treatment. Neither of these assumptions can point identify the interaction between the treatment and a moderator, but they can significantly narrow the bounds and sometimes be informative about the sign of such an interaction.

We also derive a sensitivity analysis procedure, where we vary the maximum magnitude of the discrepancy between the pre- and post-test moderator of interest that can exist and assess how the bounds change as a function of this sensitivity parameter. The proposed sensitivity analysis provides both the researcher and readers with the ability to gauge the credibility of a post-test result in light of their own assumptions.

We explore two possible avenues to further sharpen our inference from these standard designs. First, we consider a *randomized placement design*, where the experimenter randomly assigns respondents to either the pre-test or post-test design. We apply our nonparametric bounding approach and sensitivity analysis to this alternative design to examine how combining information from the pre-test and post-test arms might improve our ability to identify heterogeneous causal effects from a survey experiment.

Second, we extend the proposed methodology in a practically useful fashion by developing a model-based approach to incorporate additional pre-treatment covariates that might be available to researchers. One limitation of nonparametric bounds and sensitivity analyses is the difficulty of incorporating a large number of covariates, which may provide additional information and sharpen

our inference. We develop a parametric Bayesian approach to accommodate the covariates in a flexible manner and provide estimation uncertainty (Mealli and Pacini, 2013). Our proposed algorithm takes advantage of a recent advance in Bayesian analysis for binary and multinomial logistic regression models, which allows for fast, computationally efficient estimation of our causal quantities of interest (Polson, Scott and Windle, 2013).

This paper proceeds as follows. First, we 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. We then consider incorporating pre-treatment covariates in our analysis by developing a parametric Bayesian model. For illustration, we apply the proposed methods to the empirical example of how land insecurity moderates the effectiveness of land-based appeals by politicians from Horowitz and Klaus (2020). Finally, we conclude with some advice for practitioners and 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). In this study, the authors investigate whether politicians can use land-based grievances to increase their electoral support in Kenya's Rift Valley. Participants were randomly assigned to one of three treatment conditions. In the "control" condition, participants heard a generic campaign speech with no direct reference to the land issue. In the first treatment condition (T1), the candidate additionally states that the land issue is their top priority. In the second treatment condition (T2), the researchers added another sentence to the speech in which the candidate references an ethnic grievance, blaming "migrants and land grabbers". To preserve statistical power, we collapse the distinction between the conditions T1 and T2 and focus on the effect of a speech referencing the land issue (either T1 or T2) versus not (the "control" speech). The outcome is the participant's reported likelihood of supporting the candidate, measured on a 5-point Likert scale.[1]

[1] We dichotomize the outcome (likely to support the candidate vs. not) to apply 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 conditions (T1 and T2) and a dummy variable for land insecurity, in line with the authors' expectations (with the outcome on a 5-point numeric scale, $\hat{\beta}_{T1:LI} = 1.77$, $p = 0.01$; $\hat{\beta}_{T2:LI} = 1.71$, $p = 0.005$).

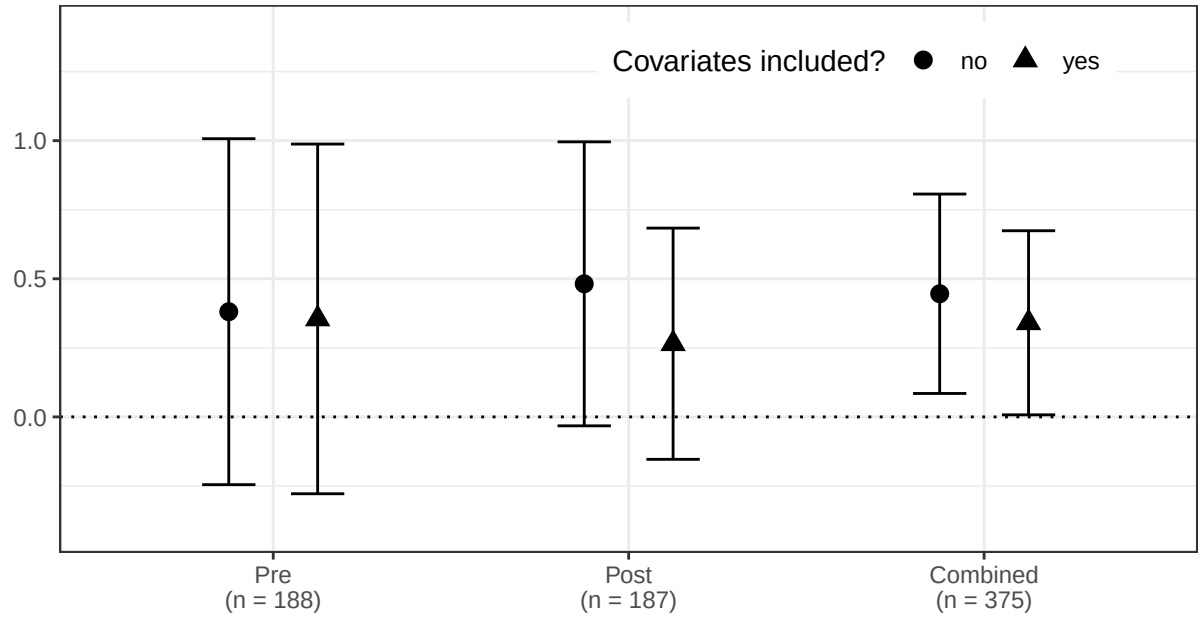
Both priming bias and post-treatment bias could be a concern in this context. Asking respondents to evaluate their land security *before* treatment may raise the salience of the land issue and thus attenuate any treatment effect of hearing a land-based appeal by a politician, which is a form of priming bias. Conversely, if the experiment asks respondents about land security *after* treatment, then their responses may be affected by the content of the speech,[2] yielding a post-treatment bias.

To address these concerns, Horowitz and Klaus (2020) use (what we call) the *random placement design*, in which questions relating to the respondent's land rights were randomly assigned to occur before or after the treatment. With this design, researchers can compare estimates of the quantity of interest produced using the pre-test or post-test data only, as well as the combined sample. As shown in Figure 1, the "naïve estimates" of the treatment-moderator interaction obtained from a linear probability model on different subsets of the data are qualitatively similar. All the point estimates are positive and substantively large, implying that the effect of a land-based appeal on electoral support is 25 – 50 percentage points larger for respondents who are land insecure compared to those who are not. However, these estimates only reach conventional levels of statistical significance with the increased statistical power of the combined pre/post sample. The point estimates from the post-test data are particularly sensitive to the inclusion of covariates.

Empirical investigations into post-treatment and priming biases for this study reveal mixed results. To investigate possible post-treatment bias, we can estimate the average treatment effect on perceived land insecurity among respondents who were asked this question after treatment. Com-

[2]Respondents were asked to rate the security of their land rights as not secure, somewhat secure, or very secure. The authors define land insecurity as corresponding to "not secure".

Figure 1: Estimates of treatment-moderator interaction using pre-test, post-test, and combined data.



Notes: Figure shows point estimates and 95% confidence intervals for the interaction between treatment and land insecurity on candidate support from a linear probability model, using the pre-test data (moderator measured before treatment), the post-test data (moderator measured after treatment), and the combined sample from Horowitz and Klaus (2020). Estimates are shown from models with and without covariates. We follow the original authors in using age, gender, education, and closeness to own ethnic group as covariates.

pared to the control condition, respondents who saw the land-based appeal tended to report higher levels of perceived land security (on a 3-point scale, $\hat{\beta} = 0.18$, $p = 0.03$), which raises the possibility of post-treatment bias.[3]

The original authors also examine whether estimates of the overall ATE from the pre-test data suffered from priming bias by interacting the treatment variables with a dummy variable for the random placement of the covariate measurement, but they find null results.[4] This analysis is feasible because the post-test design identifies the overall ATE. However, the observed data cannot shed light on whether priming bias is present in estimates of the interaction since the interaction is not

[3]In the authors' main analysis, this variable is dichotomized based on whether respondents reported the lowest level of security in their land rights. With this recoding, we observe no statistically significant average treatment effect on the post-test moderator. See Horowitz and Klaus (2020), Table A13. However, as we show in Section 4.2, ruling out post-treatment bias requires a stronger condition of no *individual* treatment effect on the moderator.

[4]See Horowitz and Klaus (2020), Table A14.

identified in either the pre-test or post-test design (see Sections 4.1 and 4.2). In Section 7, 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.

3 Experimental Designs and Causal Quantities of Interest

We now lay out the formal notation of our setting and describe the quantities of interest. Suppose we have a simple random sample of size n from a population of interest. Let T_i represent the binary treatment variable for unit i, indicating a certain experimental manipulation received by the unit, Y_i be an outcome of interest, and D_i be an observed binary covariate and moderator of interest. That is, one goal of our experimental design is to understand how the effect of T_i on Y_i varies as a function of D_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 when they measure the potential moderators. In the pre-test design, the experimenter measures the covariate, D_i, before treatment assignment, ensuring that these measurements are unaffected by treatment. In the post-test design, the experimenter measures moderators after treatment assignment. We let Z_i be an indicator for whether the experiment measures D_i before ($Z_i = 0$) or after treatment ($Z_i = 1$). The randomized placement design randomly assigns Z_i along with treatment, so only a random subset of units can have treatment affect their moderator.

We will analyze these experimental designs using the potential outcomes framework for causal inference (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. We assume that the potential outcomes are binary variables fixed for each unit i, i.e., $Y_i(t,z) \in \{0,1\}$ for all $(t,z) \in \{0,1\}^2$, though many of the methods below could be modified to handle more general outcomes. We make the following consistency assumption, $Y_i = Y_i(T_i, Z_i)$.

We assume that the potential outcome of interest is $Y_i(t,1)$—that is, the potential outcome under a particular treatment value in the post-test design. In general, however, the potential outcomes for

the pre- and post-test designs will differ so that $Y_i(t,0) \neq Y_i(t,1)$. This difference corresponds to priming bias.

Similarly, the measures of the effect modifier can be affected by the experimental design and the administered treatment. Let $D_i(t,z)$ be the potential value of the moderator that would be observed for unit i when the treatment is set to t and the variable is measured in design z. We assume this moderator is binary so that $D_i(t,z) \in \{0,1\}$, and make a consistency assumption such that $D_i = D_i(T_i, Z_i)$. Under the pre-test design, the treatment cannot affect the moderator, so we refer to $D_i(0,0) = D_i(1,0) \equiv D_i(0)$ as the true moderator for unit i at the time of the experiment. On the other hand, under the post-test design, it is possible that the treatment affects respondents' moderator, so that $D_i(0,1) \neq D_i(1,1)$, which can lead to post-treatment bias.

Given the above notation, the causal moderation effects can be explored by estimating the following quantities of interest,

$$\tau(d) \equiv \mathbb{E}(Y_i(1,1) - Y_i(0,1) \mid D_i(0) = d), \tag{1}$$

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

where $d \in \{0,1\}$. The first quantity, $\tau(d)$, characterizes the post-test conditional average treatment effect (CATE) as a function of the pre-test effect modifier. The second quantity of interest, δ, compares the CATE between two subpopulations with different levels of the pre-test moderator, which we sometimes refer to as the *interaction*.

These two quantities of interest formalize the dilemma about the choice of experimental designs researchers must make because the potential outcomes of interest ($Y_i(t,1)$) and moderator ($D_i(0)$) can never be jointly observed for the same unit. 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, even the overall ATE cannot be identified for the "true" outcome of interest despite the randomization of the treatment. This overall ATE can be written as,

$$\tau \equiv \mathbb{E}(Y_i(1,1) - Y_i(0,1)). \tag{3}$$

Thus, under the pre-test design $\tau(d)$ and δ are also not identified. In contrast, under the post-test design, the overall ATE can be estimated without bias, and yet this design may result in post-test bias for $\tau(d)$ and δ when the treatment affects the moderator. In sum, the dilemma is that we are interested in how the effects on the post-test outcome vary by pre-test moderator levels, but we cannot simultaneously observe both designs for the same individual.

4 Nonparametric Analysis of the Three Experimental Designs

We first analyze the pre- vs. post-test designs without parametric assumptions. We begin by showing that, without further assumptions, neither the pre-test nor post-test design is informative of the CATEs or the differences between CATEs. We then derive sharp bounds for the interaction between the treatment and the moderator under the post-test design and show how to narrow these bounds with additional substantive assumptions. We also develop a sensitivity analysis procedure, which allows researchers to vary the strengths of such assumptions and assess their implications. Finally, we examine the random placement design using similar approaches.

4.1 Uninformativeness of the Pre-test Designs

We first consider the identifying power of the pre-test design, where we observe (Y_i, D_i, T_i) among the units for whom $Z_i = 0$. Under this design, the randomization of the treatment guarantees the following statistical independence between potential outcomes and the treatment variable.

Assumption 1 (Pre-test Randomization)**.**

$$\{Y_i(t,z), D_i(t,1)\} \perp\!\!\!\perp T_i \quad | \quad D_i(0) = d, Z_i = 0, \tag{4}$$

for $t = 0, 1$, $d \in \{0, 1\}$ *and all* i.

This assumption allows for the possibility of conditioning our randomization on the pre-treatment moderator, though this assumption also holds when randomization is unconditional.

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

$$\begin{aligned}\tau_{pre}(d) &\equiv \mathbb{E}(Y_i \mid T_i = 1, D_i = d, Z_i = 0) - \mathbb{E}(Y_i \mid T_i = 0, D_i = d, Z_i = 0) \\ &= \mathbb{E}(Y_i(1,0) - Y_i(0,0) \mid D_i(0) = d), \end{aligned} \tag{5}$$

where the equality follows from Assumption 1 and the standard consistency assumption. Equation (5) does not generally equal the true CATE, $\tau(d)$, since the conditional distribution of $Y_i(t,0)$ given $D_i(0) = d$ may differ from that of $Y_i(t,1)$ given $D_i(0) = d$ due to priming. Similarly, δ is not generally equal to $\delta_{pre} \equiv \tau_{pre}(1) - \tau_{pre}(0)$. Indeed, the sampling process and the experimental design do not contain any direct information about the causal quantity of interest without further assumptions.

Thus, the randomization-only bounds under the pre-test design are the same as the original bounds,

$$\tau(d) \in [-1,\ 1], \tag{6}$$

$$\delta \in [-2,\ 2]. \tag{7}$$

To narrow the bounds, we require additional assumptions about the joint distribution of $Y_i(t,0)$ and $Y_i(t,1)$. Of course, such assumptions are not testable from the data under the pre-test design. This implies that, in theory, the priming bias can be arbitrarily large, and that the observed data do not contain any direct information about the magnitude of the priming bias.

4.2 Uninformativeness of the Post-test Design

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

Assumption 2 (Post-Test Randomization)**.**

$$\{Y_i(t,z), D_i(t,1), D_i(0)\} \perp\!\!\!\perp T_i \mid Z_i = 1,$$

for $t = 0, 1$.

How informative is Assumption 2 alone about the causal moderation effects, without making any other assumption? Let $P_{tzd} = \mathbb{P}(Y_i = 1 \mid T_i = t, Z_i = z, D_i = d)$ and $Q_{tz} = \Pr(D_i = 1 \mid T_i = t, Z_i = z)$ be the conditional outcome and moderator distributions. Then, the standard CATE estimator would be unbiased for $\tau_{post}(d) \equiv P_{11d} - P_{01d}$. Under the post-test design, we can connect this quantity to the following counterfactual contrast,

$$\tau_{post}(d) \;\equiv\; P_{11d} - P_{01d} = \mathbb{E}(Y_i(1,1) \mid D_i(1,1) = d) - \mathbb{E}(Y_i(0,1) \mid D_i(0,1) = d), \quad (8)$$

where the equality follows from Assumption 2 and the fact that $\Pr(Z_i = 1) = 1$ for all i.

Similar to the case of the pre-test design, $\tau_{post}(d)$ does not generally equal $\tau(d)$ because the conditional distribution of $Y_i(t,1)$ given $D_i(t,1) = d$ may differ from that of $Y_i(t,1)$ given $D_i(0) = d$. The equality of these two conditional distributions is not directly testable from the observed data. In fact, suppose that one observes that the moderator is not systematically affected by the treatment, i.e., $D_i(0,1) = D_i(1,1)$ for all $t, t' \in \mathcal{T}$. Even in this situation, $\tau_{post}(d) = \tau(d)$ does not necessarily follow from the assumption that $\Pr(D_i(t,1) = d) = \Pr(D_i(0) = d)$ for all $d \in \mathcal{D}$ and all $t \in \mathcal{T}$, because the effect heterogeneity is a function of the joint distribution of the counterfactual outcomes and moderators. Thus, under the post-test design, neither $\tau(d)$ nor δ is nonparametrically identified.

To derive the bias and bounds under the post-test design, we rely on a principal stratification approach in which we stratify the units in the study according to their potential responses to treatment on various outcomes (Frangakis and Rubin, 2002). In particular, we create strata based on how the moderator responds to both the treatment and pre-post indicators. Let $\mu_s(t) = \mathbb{P}[Y_i(t,1) = 1 \mid S_i = s]$, where S_i represents the principal strata defined by the moderator, $\{D_i(1,1), D_i(0,1), D_i(0)\}$, which under no assumptions can take any of the 2^3 values in

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

Let $\rho_s = \Pr[S_i = s]$ 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 $Q_0 = \mathbb{P}(D_i(0) = 1)$.

With these in hand, we can characterize the bias in the "naïve estimate" of $\tau(1)$, $\tau_{post}(1)$, as

$$\begin{aligned}
&\tau_{post}(1) - \tau(1) \\
&= (\mu_{111}(1)\rho_{111} + \mu_{101}(1)\rho_{101}) \left(\frac{Q_0 - Q_{11}}{Q_0 Q_{11}} \right) - (\mu_{111}(0)\rho_{111} + \mu_{011}(0)\rho_{011}) \left(\frac{Q_0 - Q_{01}}{Q_0 Q_{01}} \right) \\
&\quad - \left(\frac{\mu_{011}(1)\rho_{011} + \mu_{001}(1)\rho_{001}}{Q_0} \right) + \left(\frac{\mu_{110}(1)\rho_{110} + \mu_{100}(1)\rho_{100}}{Q_{11}} \right) \\
&\quad + \left(\frac{\mu_{101}(0)\rho_{101} + \mu_{001}(0)\rho_{001}}{Q_0} \right) - \left(\frac{\mu_{110}(0)\rho_{110} + \mu_{010}(0)\rho_{010}}{Q_{01}} \right)
\end{aligned}$$

From this expression, we see that the naïve estimator of the CATE will result in bias in an unknown direction except under the following two unlikely scenarios. First, the bias will be zero if the moderator is unaffected by either treatment status or when it is measured, i.e., $\Pr[S_i \notin \{111, 000\}] = 0$. Note that this requires no effect at the individual level, so there could still be bias even if there is no average effect of treatment on the moderator. Second, the naïve estimator will be unbiased if the distribution of potential outcomes does not depend on the moderator under either the treatment or control condition, i.e., $\overline{\mu}_{11}(t) = \overline{\mu}_{01}(t) = \overline{\mu}_{10}(t)$. By repeating this exercise for $\tau_{post}(0)$, we can arrive at similar conclusions for the bias of the naïve estimator of the interaction δ, δ_{post}.

Even though the post-test design does not point-identify the CATEs or the interaction, the design can still be informative.

Proposition 1. *Let $P_t = \mathbb{P}(Y_i = 1 \mid T_i = t, Z_i = 1)$. Under Assumptions 2, we have $\delta \in [\delta_L, \delta_U]$, where*

$$\begin{aligned}
\delta_L &= \max \left\{ -2,\ -\frac{1 - P_1 + P_0}{Q_0},\ -\frac{1 + P_1 - P_0}{1 - Q_0},\ -\frac{P_0}{Q_0} - \frac{P_1}{1 - Q_0},\ -\frac{1 - P_1}{Q_0} - \frac{1 - P_0}{1 - Q_0} \right\}, \\
\delta_U &= \min \left\{ 2,\ \frac{1 + P_1 - P_0}{Q_0},\ \frac{1 - P_1 + P_0}{1 - Q_0},\ \frac{1 - P_0}{Q_0} + \frac{1 - P_1}{1 - Q_0},\ \frac{P_1}{Q_0} + \frac{P_0}{1 - Q_0} \right\}.
\end{aligned} \tag{9}$$

Furthermore, these bounds are sharp.

The proofs of this result and the bounds for each CATE are given in Section A.1 and rely on a standard linear programming approaches often used in bounding causal quantities (e.g., Balke

and Pearl, 1997; Imai and Yamamoto, 2010). Sharpness here means that these bounds are the shortest possible bounds without additional assumptions. Unfortunately, these bounds are often quite wide in practice. The reason is that the observed data under the post-test design are completely uninformative about the true moderator. This means that Q_0 can take on any value within the unit interval. Thus, the sharp bounds under the post-test design can be quite wide and sometimes cover the entire possible range, $[-2, 2]$, for δ.

4.3 Narrowing the Post-test Bounds under Additional Assumptions

While the bounds only using the randomization can be reasonably wide in practice, we may be willing to entertain other assumptions on the causal structure that will allow us to narrow the bounds. We focus here on the post-test design since it has the benefit of ensuring identification of the ATE, but similar assumptions and bounds can be constructed for the pre-test design (see, for example, the priming assumptions we derive below for the random placement design).

4.3.1 Monotonicity and Stable Moderator Assumptions

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 3 (Monotonicity of the Post-Test Effect)**.** $D_i(t,1) \geq D_i(0)$ *or* $D_i(t,1) \leq D_i(0)$ *for all* $t = 0, 1$.

In the context of the motivating example, this would require hearing a land-based appeal by a politician to shift perceived land insecurity in the same direction for all respondents (or have no effect). Obviously, the plausibility of this assumption will depend on the experimental context. In most of the paper, we focus on the positive version of this monotonicity assumption. This 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. While we present the monotonicity assumption in a particular direction for both treatment levels, it is possible to derive bounds under a monotonicity assumption with differing directions for each treatment condition.

Proposition 2. *Under Assumptions 2 and 3, we have sharp bounds* $\delta \in [\delta_{L2}, \delta_{U2}]$*, where*

$$\begin{aligned}\delta_{L2} &= P_{111}\left(\frac{Q_{11}}{Q_0}\right) - P_{011}\left(\frac{Q_{01}}{Q_0}\right) - P_{110}\left(\frac{1-Q_{11}}{1-Q_0}\right) + P_{010}\left(\frac{1-Q_{01}}{1-Q_0}\right) \\ &\quad + \frac{\max\{P_{011}Q_{01} - Q_0, 0\} - \min\{P_{111}Q_{11}, Q_{11} - Q_0\}}{Q_0(1-Q_0)}, \\ \delta_{U2} &= P_{111}\left(\frac{Q_{11}}{Q_0}\right) - P_{011}\left(\frac{Q_{01}}{Q_0}\right) - P_{110}\left(\frac{1-Q_{11}}{1-Q_0}\right) + P_{010}\left(\frac{1-Q_{01}}{1-Q_0}\right) \\ &\quad + \frac{\min\{0, Q_0 - P_{111}Q_{11}\} + \min\{P_{011}Q_{01}, Q_{01} - Q_0\}}{Q_0(1-Q_0)}.\end{aligned}$$

We provide the derivation of these bounds (and those in the next proposition) in Section A.2. Inspection of these functions reveals that we can find the maximum of the upper bound by comparing when Q_0 makes one of the two minimum statements into equalities, $Q_0 = P_{111}Q_{11}$ or $Q_0 = (1 - P_{011})Q_{01}$.

These bounds will differ from the above randomization bounds in two ways. 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 places no restriction on the relationship between the pre-test and post-test versions of the moderator. Under monotonicity (Assumption 3), we can leverage P_{tzd} and Q_{tz} to narrow the bounds. Second, monotonicity places bounds on the true value of Q_0 since it must be less than $\min(Q_{11}, Q_{01})$.

While the post-test 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 4 (Stable Moderator under Control)**.** $D_i(0) = D_i(0, 1)$

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 3 and 4, the only values that principal strata that S_i can take are $\{111, 100, 000\}$.

Proposition 3. *Under Assumptions 2, 3, and 4, we have* $Q_0 = Q_{01}$ *and sharp bounds* $\delta \in [\delta_{L3}, \delta_{U3}]$, *where*

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

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

These bounds demonstrate how the magnitude of treatment effect on the moderator affects identification in the post-test design. A unit's moderator is affected by treatment whenever $D_i(1,1) \neq D_i(0,1)$, 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 narrow if the post-treatment average effect on the moderator is small.

4.3.2 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. To this end, we now consider an alternative approach to bounds that does not rule out any particular principal strata, but rather places restrictions on what proportion of units have moderators that 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 values, regardless of the treatment condition (contrast this with Assumption 4 which applies to the control condition only). We operationalize this via the following constraint,

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

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 the second step of the above constrained optimization procedure. 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.

4.3.3 Statistical Inference

The above bounding procedures assume that we observe the conditional probabilities P_{tzd} perfectly when, in fact, we only have estimates of these quantities. Thus, the bounds derived above do not account for estimation uncertainty. To resolve this, we rely on the approach of Imbens and Manski (2004) to derive confidence intervals for the underlying parameters of interest that take into account their partial identification. These confidence intervals are designed to provide nominal coverage for the partially identified parameter rather than nominal coverage for the unobserved bounds. This procedure requires estimates of the bounds' variance, which we obtain via the nonparametric bootstrap.

4.4 Analysis of the Randomized Placement Design

Finally, we consider a combined pre/post design called the *randomized placement design*, where in addition to treatment, the timing of covariate measurement, Z_i, is also randomized. Under the randomized placement design, the bounds from the post-test alone can be tightened because we can identify the marginal probability of the true moderator as

$$Q_0 \;=\; \Pr(D_i(0) = 1) \;=\; \Pr(D_i = 1 \mid Z_i = 0).$$

Unfortunately, for the randomization bounds in Equation (9), the sign of δ cannot be identified for any value of $Q_0 \in (0, 1)$. However, under the assumption of monotonicity for the moderator (Assumption 3), the bounds can be informative.

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 have the following:

$$\psi_{y_1 y_0 s}(t) \;=\; \Pr(Y_i(t,1) = y_1, Y_i(t,0) = y_0 \mid S_i = s)$$

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. These values characterize the joint distribution of the pre-test and post-test potential outcomes for a given treatment level, t, and principal strata, s. Furthermore, let $\mathcal{S}_d^* = \{s : s \in \mathcal{S}, D_i(0) = d\}$ be the set of principal strata with the true value of the moderator equal to d.

Given these, we can write the interaction between the treatment and the moderator as

$$\delta \;=\; \sum_{y_0=0}^{1} \left\{ \sum_{s_1 \in \mathcal{S}_1^*} \frac{\rho_{s_1}}{Q_0} (\psi_{1 y_0 s_1}(1) - \psi_{1 y_0 s_1}(0)) - \sum_{s_0 \in \mathcal{S}_0^*} \frac{\rho_{s_0}}{(1 - Q_0)} (\psi_{1 y_0 s_0}(1) + \psi_{1 y_0 s_0}(0)) \right\},$$

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

$$P_{t1d} Q_{t1} = \Pr(Y_i = 1, D_i = d \mid T_i = t, Z_i = 1) \;=\; \sum_{y_0=0}^{1} \sum_{s \in \mathcal{S}(t,1,d)} \psi_{y y_0 s}(t) \rho_s,$$

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

for all values of y, d,, t, and d^*.

4.4.1 Bounds under Priming Monotonicity

Without further assumptions, the pre-test data only identifies the probability distribution of the moderator, Q_0, but additional assumptions about the connection between the pre- and post-test data can be informative. We consider a priming monotonicity assumption that states that the effect of asking the moderator before treatment can only move the outcome in a single direction.

Assumption 5 (Priming Monotonicity)**.** $Y_i(t,0) \geq Y_i(t,1)$ *for all* $t = 0, 1$.

That is, 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. As stated, this assumption holds across levels of

treatment, though it is possible to assume the reverse direction ($Y_i(t,0) \leq Y_i(t,1)$) or even to have a different effect direction for each level of treatment.

As with moderator monotonicity, priming monotonicity implies restrictions on the principal strata that help narrow the bounds on the CATEs and interactions. In particular, priming monotonicity implies that $\psi_{10s} = 0$ for any value of $s \in \mathcal{S}$. To obtain bounds under this assumption, we again can numerically solve the above linear programming problem subject to these restrictions. Furthermore, combining this priming monotonicity with moderator monotonicity to narrow the bounds even further is possible.

4.4.2 Sensitivity Analysis

With this setup, we can also develop a sensitivity analysis procedure based on substantive assumptions about how much the pre-test vs. post-test measurement of the moderator affects the moderator itself and the outcome. The first of these is the sensitivity analysis described above for the post-test design now adapted to the principal strata of the randomized placement design:

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

Again, this assumption constrains the proportion of the respondents whose moderator value is affected by Z_i in either treatment arm. If $\gamma = 0$, then the moderator is unaffected by the measurement timing and $D_i(t,1) = D_i(0)$. We take a similar approach for the issue of priming for the outcomes. In particular, we add a substantive constraint on the proportion of respondents primed by the pre-test measurement of D_i, or more precisely, the proportion of respondents whose value of $Y_i(t,1)$ is different from $Y_i(t,0)$.

Thus, the sensitivity analysis would use the following restriction:

$$\begin{aligned} &\Pr(Y_i(t,1) = 1, Y_i(t,0) = 0 \mid D_i(0) = d_*) + \Pr(Y_i(t,1) = 0, Y_i(t,0) = 1 \mid D_i(0) = d_*) \leq \theta \\ &\iff \frac{\sum_{s \in \mathcal{S}^*_{d_*}} (\psi_{10s}(t) + \psi_{01s}(t))\, \rho_s}{\sum_{s \in \mathcal{S}^*_{d_*}} \rho_s} \leq \theta, \end{aligned}$$

for all $t, d^* \in \{0,1\}$. Note that if $\theta = 0$, then we have $\text{corr}(Y_i(t,1), Y_i(t,0) \mid D_i(0) = d_*) = 1$ and pre-test data alone identify the quantity of interest. Like the γ-based sensitivity analysis, this sensitivity analysis gradually restricts the severity of the priming effects in the empirical setting.

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 (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. The value of this approach is that it allows researchers to encode substantive assumptions and allows for more agnostic assessments of the results.

5 Extension to post-treatment attention checks

We can extend the proposed approach to other problems of post-treatment bias. We briefly discuss an important application of our methodology to post-treatment attention checks in survey experiments.

5.1 Problem

A common concern in survey experiments is inattentive respondents who fill out answers without regard to the actual survey questions. Inattentiveness is especially concerning with the rise of online platforms for respondent recruitment like Mechanical Turk or Lucid. To combat these types of respondents, there has been an explosion in attention or manipulation checks in surveys to identify which respondents are paying attention and which are not.

These checks can come before treatment, in which case they are usually called attention checks, or after treatment, in which they are usually called *manipulation checks* because these questions ask

the respondent to recall some features of the experimental manipulation. Analysts will then usually drop respondents who fail either of these checks. As pointed out by Aronow, Baron and Pinson (2018), dropping units based on post-treatment manipulation checks can lead to biased estimates of the CATE on attentive respondents if treatment affects the ability to pass the check.

While pre-treatment attention checks appear to be a viable alternative, these checks are, by their nature, less specific to the study at hand. Moreover, standard attention checks have become well-known to "professional" survey-takers, and so may not be a good proxy for actual attentiveness. Post-treatment manipulation checks are harder for these types of adversarial respondents to game since they depend on the specifics of the experimental manipulation. Of course, a respondent might fail a manipulation check even if they read the prompt carefully and the experiment affected them. Furthermore, recall could be different for different treatments if those treatments differed in how specific or vivid they were.

5.2 Nonparametric approach

This setting is the same as the above problem of estimating an interaction based on a post-treatment moderator. Our goal is to estimate the CATE of a post-treatment manipulation check. In this setting, $D_i(0)$ measures if a respondent was attentive before treatment was assigned, and $D_i(t,1)$ is an indicator for passing the post-treatment manipulation check under treatment t.

We make two plausible behavioral assumptions about attentiveness to derive realistic bounds in this setting.

Assumption 6 (Attention monotonicity)**.** *For all i and $t \in \{0,1\}$, $D_i(t,1) \leq D_i(0)$.*

Assumption 7 (Inattentive exclusion restriction)**.** $\mu_{000}(1,1) - \mu_{000}(0,1) = 0$

The first assumption states that respondents cannot recall facts to which they were not paying attention. This is a negative version of the monotonicity from earlier, and in this context means that all of those who would pass attention checks later in the survey were attentive to the treatment prompts themselves. The second assumption states that there is no average treatment effect for re-

spondents who are never attentive, which seems plausible if inattentive respondents fill out surveys randomly.

Proposition 4. *Under randomization, attention monotonicity, and the inattentive exclusion restriction, we can bound the CATE as*

$$\tau(1) \in \left[P_1 - P_0, \frac{P_1 - P_0}{\max(Q_{01}, Q_{11})} \right], \tag{10}$$

and these bounds are sharp.

Proposition 4 establishes an *inattentive attenuation* result: the estimated post-test treatment effect, $P_1 - P_0$, is a lower bound for the true conditional average treatment effect among the attentive. Its proof is in Section A.3. Inattentive attenuation is an intuitive result, but our infrastructure allows us to develop the assumptions needed to justify it and derive an upper bound for the relevant CATE based on the observed data. This result holds even when treatment affects the attention checks so long as it does so in a way that is consistent with our monotonicity assumption. If we define true attentiveness $D_i(0) = 1$ as paying attention to the treatment presented, then monotonicity appears to be a mild assumption that respondents cannot recall information to which they never paid attention. The crucial assumption for our result is the lack of an average treatment effect for those that do not pay attention. This assumption could be violated if subtle information is conveyed to inattentive respondents that is not captured by the attention checks but does influence the outcome. In that case, one could use the more general bounds for $\tau(1)$ presented in Appendix A.3.

6 Incorporating Covariates to Sharpen Inference

The nonparametric bounds above are sharp in the sense that they leverage all information about the outcome, moderator, treatment, and question order. Researchers, however, often have additional data in the form of covariates that may help reduce the uncertainty of their estimates. Here, we consider a Bayesian parametric model of the principal strata approach to the pre-test, post-test, and random placement designs, building on the work of Mealli and Pacini (2013) (see also Imbens and

Rubin, 1997; Hirano et al., 2000). Unlike the nonparametric bounds approach, a Bayesian model allows us to incorporate prior information about the data-generating process in a smooth and flexible manner.

6.1 The Model

Our approach focuses on a data augmentation strategy that models the joint distribution of the outcomes and the principal strata, the latter of which are not directly observable. We allow the distribution of the potential outcomes and principal strata conditional on those strata to further depend on covariates via a binomial and multinomial logistic model, respectively:

$$\mathbb{P}(Y_i = 1 \mid T_i = t, Z_i = z, S_i = s, \boldsymbol{X}_i) = \mu_{is}(t,z) = \text{logit}^{-1}(\alpha_{tz|s} + \boldsymbol{X}_i'\boldsymbol{\beta}),$$
$$\mathbb{P}(S_i = s \mid \boldsymbol{X}_i) = \rho_{is} = \frac{\exp(\boldsymbol{X}_i'\boldsymbol{\psi}_s)}{\sum_{j \in \mathcal{S}} \exp(\boldsymbol{X}_i'\boldsymbol{\psi}_j)},$$

where the strata probabilities do not depend on T_i and Z_i due to randomization. We gather the parameters as $\boldsymbol{\alpha} = \{\alpha_{tz|s}\}$ and $\boldsymbol{\psi} = \{\boldsymbol{\psi}_s\}$. We can easily incorporate assumptions like monotonicity and stable moderators by simply restricting the space of possible principal strata $\mathcal{S}$.

Our goal is to make inferences about the posterior distribution of these parameters and the ultimate quantity of interest, δ. There are two ways to represent δ under this parametric model, resulting in two different posterior distributions. The first is based on *population* inference and derives an expression for δ purely in terms of the parameters of the model. The second is based on *in-sample* inference and derives an expression for δ in terms of potential outcomes in a particular sample.

For the population inference approach, we first note that due to consistency and randomization, we have $\mu_{is}(t,z) = \mathbb{P}(Y_i(t,z) = 1 \mid S_i = s, \boldsymbol{X}_i)$. Thus, we can write the value of the interaction of interest for a given unit as,

$$\begin{aligned}\delta_i &= \mathbb{E}\left[Y_i(1,1) - Y_i(0,1) \mid D_i(0) = 1, X_i\right] - \mathbb{E}\left[Y_i(1,1) - Y_i(0,1) \mid D_i(0) = 0, X_i\right] \\ &= \left(\sum_{s \in \mathcal{S}_1^*} (\mu_{is}(1,1) - \mu_{is}(0,1))\rho_{is}\right) - \left(\sum_{s \in \mathcal{S}_0^*} (\mu_{is}(1,1) - \mu_{is}(0,1))\rho_{is}\right),\end{aligned}$$

where we omit the implied dependence on $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi})$ and remember that $\mathcal{S}_d^*$ is the set of strata levels such that $D_i(0) = d$. Using the empirical distribution of covariates, the average of these conditional interactions will equal the overall quantity of interest, i.e., $\delta = \sum_{i=1}^n \delta_i / n$.

The in-sample version of the interaction is more straightforward, since it is just the interaction among the units in the sample

$$\delta_s = \frac{\sum_{i=1}^n D_i(0) \left\{ Y_i(1,1) - Y_i(0,1) \right\}}{\sum_{i=1}^n D_i(0)} - \frac{\sum_{i=1}^n (1 - D_i(0)) \left\{ Y_i(1,1) - Y_i(0,1) \right\}}{\sum_{i=1}^n (1 - D_i(0))}.$$

Obviously, across repeated samples, we can relate this to the population quantity by $\mathbb{E}[\delta_s] = \delta$. In Supplemental Materials B, we define a Markov Chain Monte Carlo (MCMC) algorithm to take draws from the posterior and then calculate each value.

This Bayesian approach has the advantage of easily incorporating covariates, but it does require us to select prior distribution for the model parameters, some of which are unidentified in the frequentist sense. Thus, the identification of these parameters will depend on the prior. To investigate this, we take draws of the prior predictive distribution under different prior structures, which we show in Figure A.5. All the priors we considered are symmetric, but uniform priors on the model parameters lead to somewhat informative priors on the ultimate quantity of interest. Thus, we rely on more dispersed priors for the simulations and the application. We discuss the choice of prior distribution more fully in Appendix B.

We conduct two simulation studies to demonstrate the gains in efficiency from the monotonicity and stability assumptions and the incorporation of covariates in the Bayesian approach. The first simulation varies the assumptions of the data-generating process and compares the posterior variance of these distributions across combinations of our two assumptions. The second simulation varies the predictive power of the covariates on the outcome and the strata in the data-generating process and compares the variance of the posterior distributions from Gibbs run on each simulated data set with and without incorporating covariates. We present these results in Appendix B.1.

7 Empirical Example

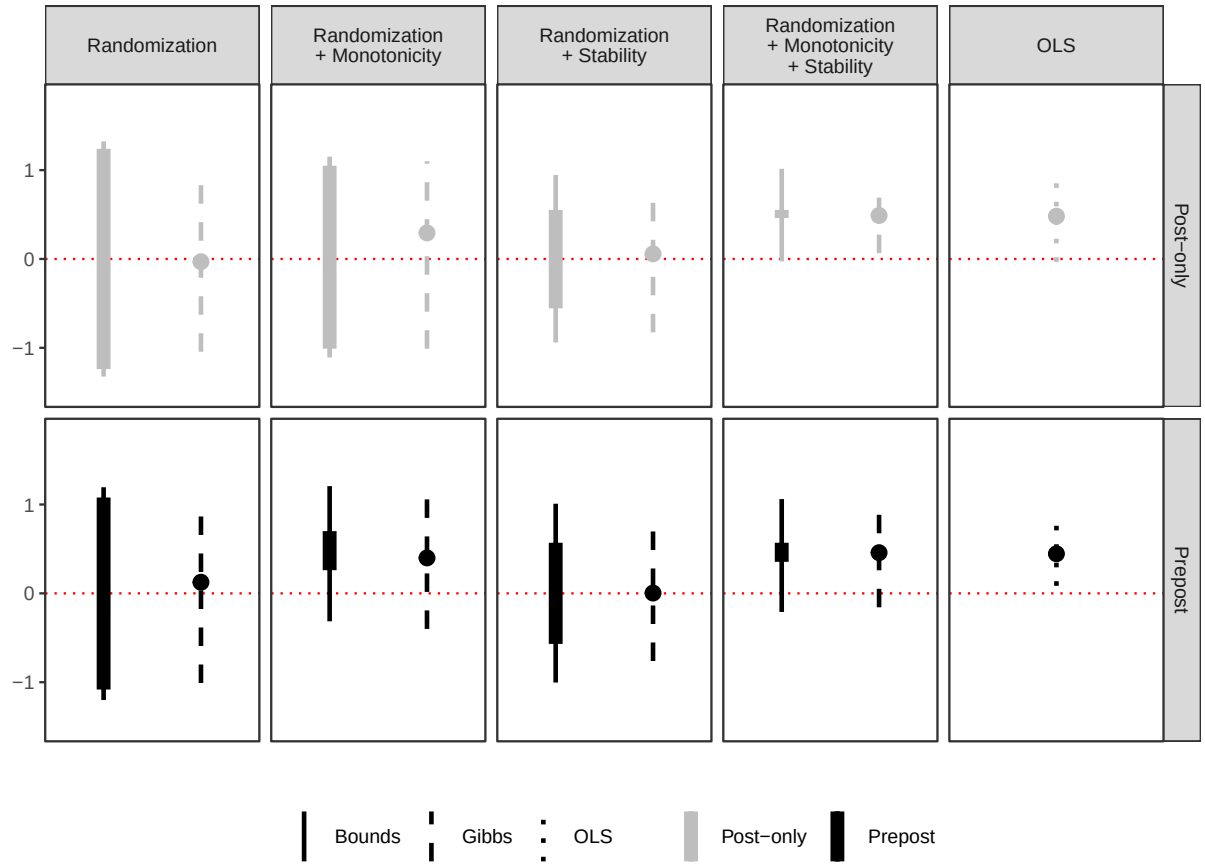
To illustrate how the sharp bounds and Bayesian approach can be applied to both the post-test and randomized placement design, we return to the example from Horowitz and Klaus (2020) introduced in Section 2.

7.1 The Setup

First, it is worth discussing our assumptions in this context. Recall that Assumption 2 (randomization) holds due to the randomization of treatment. Assumptions 3 (monotonicity) and 4 (stability), however, are not guaranteed by the design and require a substantive justification. In this case, the assumption of monotonicity implies that hearing a land-based appeal by a politician must 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 our estimate of the average treatment effect – though it is not statistically significant and is substantively small ($\hat{\beta} = 0.02$, $p = 0.59$). One theoretical justification for this assumption is based on a simple cueing mechanism: listening to a speech in which a politician discusses the importance of the land issue could prompt respondents to think about past conflict over land and consequently report a higher level of perceived land 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.

Figure 2: Comparing non-parametric bounds and Bayesian estimates for δ under different assumptions



Notes: The figure shows nonparametric bounds and Bayesian estimates of the quantity of interest under different sets of assumptions applied to either the post-only data (grey) or the combined pre-post data (black). The thick bars denote the width of the bounds, and thinner lines denote the 95% confidence intervals around the bounds. Across the first four panels, the thin lines with dots denote the Bayesian posterior mean and 95% credible interval. This estimate included no covariates to facilitate comparison with the nonparametric bounds. For the final panel ("OLS"), the thin lines with dots denote the OLS estimate and 95% confidence interval.

7.2 Comparing the sharp bounds and Bayesian approach

Figure 2 displays the non-parametric bounds (with 95% confidence intervals) and Bayesian estimates (posterior means with 95% credible intervals) for δ under different sets of assumptions applied to the post-only data (in grey) and to the combined pre-post data (in black). We also include the naïve OLS estimate with 95% confidence interval for comparison in the final panel.

Assuming only randomization, the nonparametric bounds and the Bayesian credible intervals are not informative of the sign of δ, and are much wider than the confidence interval of the naïve OLS estimate. Without incorporating information from covariates or imposing substantive assumptions, there is little evidence for the theorized interaction, where respondents who are land insecure respond more positively to land-based appeals by politicians. Adding the assumption of monotonicity reduces the width of the bounds and the Bayesian credible intervals, especially on the combined pre-post data. Adding this assumption also shifts the posterior mean from being close to zero to being positive and substantively large (0.28 for the post-only data, 0.49 for the pre-post). These estimates are closer in magnitude to the naïve OLS estimates, which suggest that the treatment effect of land-based appeals is 30-40 percentage points larger among respondents who are land insecure compared to those who are not.

Adding the assumption of stability tightens the bounds and credible intervals to a similar degree for the post-only and pre-post data. With all three assumptions (randomization, monotonicity, and stability), the width of the nonparametric bounds and Bayesian credible intervals is roughly similar to the width of the confidence intervals around the naïve OLS estimate. The 95% credible interval for the post-only data just excludes zero ($[0.04, 0.90]$), while the 95% credible interval for the combined pre-post data just includes it ($[-0.05, 0.87]$).

Thus, while the original authors correctly noted that the average treatment effect on the post-measured moderator is insignificant and close to zero, this is not sufficient to ensure that the naïve OLS estimate of the treatment-moderator interaction is free from post-treatment bias. In addition, the observed data cannot rule out the possibility of priming bias for the CATE. With only the design-based assumption of randomization, the nonparametric bounds and estimates from the Bayesian approach do not support the hypothesis of a positive interaction effect. It is only with two additional substantive assumptions – monotonicity and stability – that the sharp bounds and Bayesian approach produce results qualitatively similar to the naïve OLS estimate, of a positive interaction effect that is verging on statistical significance at the 5% confidence level.

7.3 Implementing the sensitivity analysis

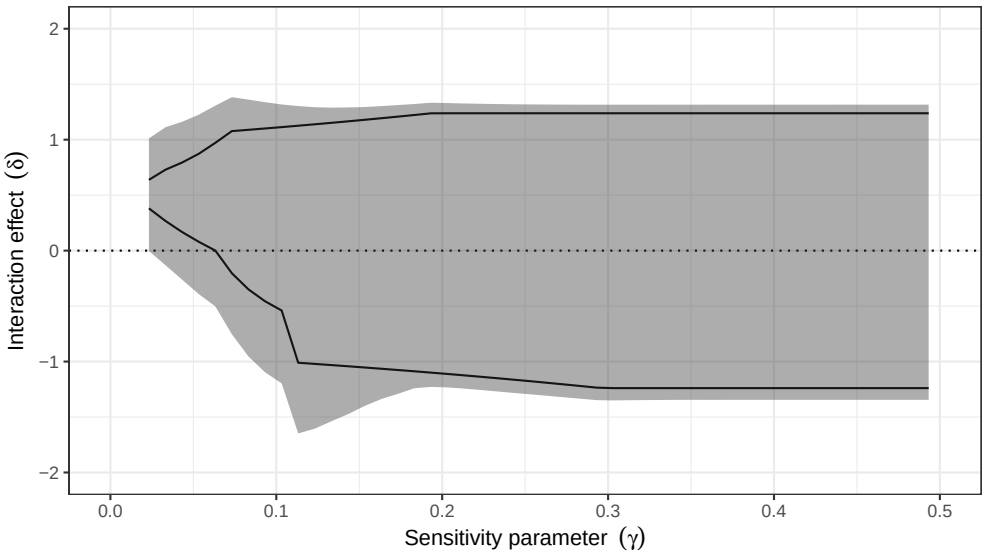
While we cannot directly observe the degree of individual-level post-treatment bias, it is possible to investigate how the nonparametric bounds change by varying an assumption about the degree to which this bias is present in the sample. Here, we apply the sensitivity analysis described in Section 4.3.2 to our empirical example. Recall that the parameter γ bounds the proportion of respondents who would respond differently to the land insecurity question if asked pre- or post-treatment, and that the minimum feasible value of γ is given by $|Q_{11} - Q_{01}|$.

Figure 3 shows how the nonparametric bounds vary as a function of γ. In this example, we impose Assumptions 3 (monotonicity) and 4 (stability). 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 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. When we incorporate the estimation uncertainty, we can see that the shaded ribbon already includes 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. For example, in this case we may assume that some respondents experience such a high degree of land security that they are unlikely to change their response from "very secure" to "not at all secure" due to post-test measurement. If any respondents are affected by the post-test measurement, it is likely to be those who would respond "somewhat secure" in the pre-test and would change their answer to "not at all secure" in the post-test.

We can estimate the true distribution of the moderator using the pre-test data, in which about 30% of respondents choose the middle category ("somewhat secure"). Out of those respondents, about 40% also reported that they had personally been affected by prior ethnic conflict. Therefore, we might assume that this subset, which comprises about 12% of the full sample, is most likely to

Figure 3: Sensitivity Analysis under Limited Effects on the Moderator



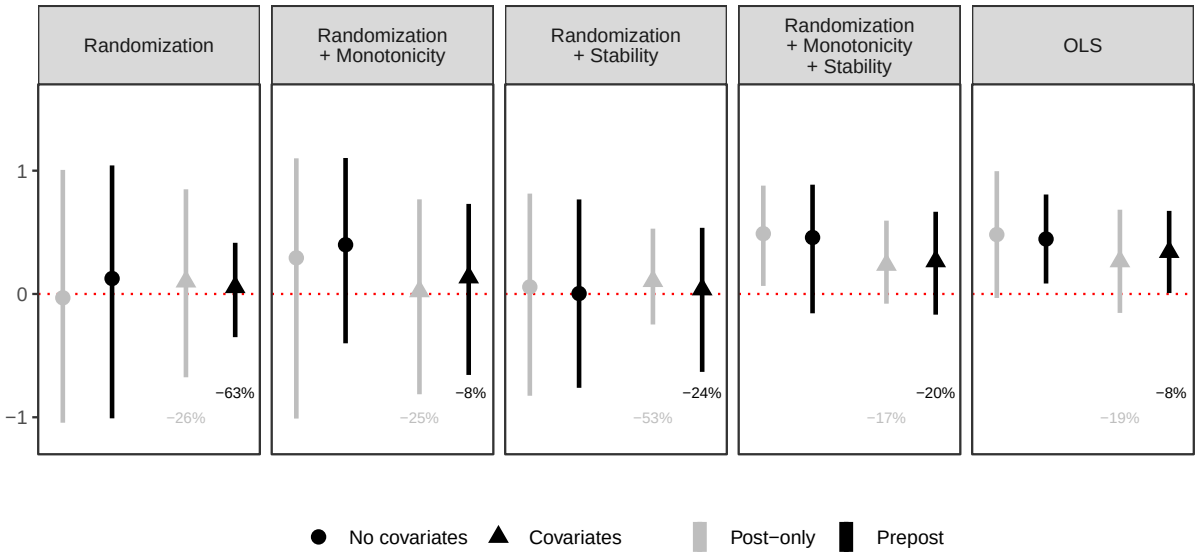
Notes: The figure shows nonparametric bounds (black lines) with 95% confidence intervals (grey ribbon) as a function of γ, the proportion of respondents whose value of the moderator variable (land insecurity) is affected by post-test measurement. The assumptions of monotonicity and stability are both imposed. While γ can vary up to 1, we limit the maximum to 0.5 to aid presentation since the bounds plateau beyond this point.

be affected by the cueing mechanism and change their perceived land insecurity when measured after treatment. 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 ($[0.38, 0.64]$).

7.4 Incorporating covariates into the Bayesian approach

Thus far, our Bayesian estimates have omitted covariates to aid comparison with the nonparametric bounds. However, as discussed above, a key attraction of the Bayesian approach is the ease with which we can incorporate additional information. Figure 4 presents posterior means with 95% credible intervals, both with and without covariates, under different assumptions. We follow the original authors in using age, gender, education, and closeness to one's ethnic group as covariates. Including covariates significantly tightens the credible intervals, especially when fewer assumptions are imposed. For example, when only randomization is assumed, the width of the 95% credible

Figure 4: Comparing Bayesian estimates for δ with and without covariates



Notes: Figure shows posterior means and 95% credible intervals for δ under different sets of assumptions applied to either the post-only data (grey) or the combined pre-post data (black). Estimates are shown with and without the inclusion of covariates (denoted by triangles and circles, respectively), and the numbers indicate the reduction in the width of the credible intervals due to the inclusion of covariates for the post-only data (in grey) and the combined pre-post data (in black). We follow the original authors in using age, gender, education, and closeness to one's ethnic group as covariates. The naïve OLS estimates are included for comparison.

interval shrinks by more than 50% when including covariates. While including covariates does not alter our substantive conclusions in this case, it does show that incorporating additional information can lead to large gains in precision. Since researchers often include a wide range of control variables in the design of a survey experiment, flexibly leveraging this information is a key advantage of the Bayesian approach.

8 Concluding Remarks

This paper addresses a central tension in survey methodology: how should researchers weigh up priming bias versus post-treatment bias when designing a survey experiment? We provide sharp bounds for interactions for covariates measured post-treatment and show how these bounds vary

under additional substantive assumptions. We also provide sensitivity analyses for both types of bias 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 develop a Bayesian parametric model to incorporate pre-treatment covariates into our analysis. 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 random placement design to balance the priming and post-treatment bias concerns. In addition, it would be interesting to 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.

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):512–529.

Albertson, Bethany and Stephen Jessee. 2022. "Moderator Placement in Survey Experiments: Racial Resentment and the "Welfare" versus "Assistance to the Poor" Question Wording Experiment." *Journal of Experimental Political Science* pp. 1–7. Forthcoming.

Aronow, Peter M, Jonathon Baron and Lauren Pinson. 2018. "A note on dropping experimental subjects who fail a manipulation check." *Political Analysis* pp. 1–18.

Balke, Alexander and Judea Pearl. 1997. "Bounds on treatment effects from studies with imperfect compliance." *Journal of the American Statistical Association* 92:1171–1176.

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

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

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin and Xiao-Hua Zhou. 2000. "Assessing the effect of an influenza vaccine in an encouragement design." *Biostatistics* 1(1):69–88.

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

Horowitz, Jeremy and Kathleen Klaus. 2020. "Can politicians exploit ethnic grievances? An experimental study of land appeals in Kenya." *Political Behavior* 42(1):35–58.

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

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

Imbens, Guido W. and Donald B. Rubin. 1997. "Bayesian inference for causal effects in randomized experiments with noncompliance." *The Annals of Statistics* 25(1):305–327.

Klar, Samara. 2013. "The influence of competing identity primes on political preferences." *The Journal of Politics* 75(4):1108–1124.

Klar, Samara, Thomas J. Leeper and Joshua Robison. Forthcoming. "Studying Identities with Experiments: Weighing the Risk of Post-Treatment Bias Against Priming Effects." *Journal of Experimental Political Science* .

Mealli, Fabrizia and Barbara Pacini. 2013. "Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance." *Journal of the American Statistical Association* 108(503):1120–1131.

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

Morris, Michael W, Erica Carranza and Craig R Fox. 2008. "Mistaken identity: Activating conservative political identities induces "conservative" financial decisions." *Psychological Science* 19(11):1154–1160.

Polson, Nicholas G., James G. Scott and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108(504):1339–1349.

Press, Daryl G., Scott D. Sagan and Benjamin A. Valentino. 2013. "Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-Use of Nuclear Weapons." *American Political Science Review* 107(1):188–206.

Rosenbaum, Paul 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, Kaylyn Jackson, B Pablo Montagnes and Zachary Peskowitz. 2022. "Priming Self-Reported Partisanship: Implications for Survey Design and Analysis." *Public Opinion Quarterly* 86(3):643–667.

Transue, John 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, 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.

A Proofs

A.1 Randomization bounds

Proof of Proposition 1. Under Assumption 2, the information about the parameter of interest comes from $P_t = \mathbb{P}(Y_i = 1 \mid T_i = t, Z_i = 1)$ alone. This is because the distribution of the post-test moderators provides no information about the pre-test moderator. Recall that

$$P_t = \pi_{t1} Q_0 + \pi_{t0}(1 - Q_0), \tag{11}$$

where $\pi_{td} = \mathbb{P}[Y_i(t,1) = 1 \mid D_i(0) = d]$ and $Q_0 = \mathbb{P}[D_i(0) = 1]$.

Below, we show how to derive the upper bound for δ. The derivation of the lower bound is similar. Conditional on Q_0, we can define the following linear program:

$$\begin{aligned} &\max \ \pi_{11} - \pi_{01} - \pi_{10} + \pi_{00} \\ &\text{subject to } \pi_{t1} Q_0 + \pi_{t0}(1 - Q_0) = P_t \quad \text{for } t = 0, 1, \\ &0 \leq \pi_{td} \leq 1 \quad \forall (t,d) \in \{0,1\}^2 \end{aligned}$$

We can convert this to an augmented form by adding slack variables,

$$\begin{aligned} &\max \ \pi_{11} - \pi_{01} - \pi_{10} + \pi_{00} \\ &\text{subject to } \pi_{t1} Q_0 + \pi_{t0}(1 - Q_0) = P_t \quad \text{for } t = 0, 1 \\ &\pi_{td} + s_{td} = 1 \quad \forall (t,d) \in \{0,1\}^2 \\ &\{\pi_{11}, \pi_{01}, \pi_{10}, \pi_{00}, s_{11}, s_{01}, s_{10}, s_{00}\} \geq 0. \end{aligned}$$

The feasibility of various basic solutions here will depend on the relationship between the observed probabilities and Q_0. In Table A.1, we show basic feasible solutions for the four different conditions relating P_1 and P_0 to Q_0. Under each condition, it is straightforward to determine that the basic feasible solution is also optimal since there is no entering variable that can increase the value of the quantity of interest. Thus, we know that

$$\delta \leq \min \left\{ \frac{1 + P_1 - P_0}{Q_0}, \ \frac{1 - P_1 + P_0}{1 - Q_0}, \ \frac{1 - P_0}{Q_0} + \frac{1 - P_1}{1 - Q_0}, \ \frac{P_1}{Q_0} + \frac{P_0}{1 - Q_0} \right\}$$

A similar derivation shows that

$$\delta \geq \max \left(-\frac{1 - P_1 + P_0}{Q_0}, \ -\frac{1 + P_1 - P_0}{1 - Q_0}, \ -\frac{P_0}{Q_0} - \frac{P_1}{1 - Q_0}, \ -\frac{1 - P_1}{Q_0} - \frac{1 - P_0}{1 - Q_0} \right).$$

Table A.1: Optimal solutions to the linear program under different conditions

Condition	π_{11}	π_{10}	π_{01}	π_{00}	s_{11}	s_{01}	s_{10}	s_{00}	Value
$P_1 > Q_0, P_0 > 1-Q_0$	$\frac{P_1}{Q_0}$	0	0	$\frac{P_0}{1-Q_0}$	$1-\frac{P_1}{Q_0}$	1	1	$1-\frac{P_0}{1-Q_0}$	$\frac{P_1}{Q_0}+\frac{P_0}{1-Q_0}$
$P_1 < Q_0, P_0 > 1-Q_0$	1	$\frac{P_1-Q_0}{1-Q_0}$	0	$\frac{P_0}{1-Q_0}$	0	1	1	$1-\frac{P_0}{1-Q_0}$	$\frac{1-P_1+P_0}{1-Q_0}$
$P_1 > Q_0, P_0 < 1-Q_0$	$\frac{P_1}{Q_0}$	0	$\frac{P_0-1+Q_0}{Q_0}$	1	$1-\frac{P_1}{Q_0}$	1	1	0	$\frac{P_1-P_0+1}{Q_0}$
$P_1 < Q_0, P_0 < 1-Q_0$	1	$\frac{P_1-Q_0}{1-Q_0}$	$\frac{P_0-1+Q_0}{Q_0}$	1	0	1	1	0	$\frac{1-P_1}{1-Q_0}+\frac{1-P_0}{Q_0}$

Sharpness of these bounds is implied by the linear nature of the optimization function and the convexity of the feasible set. If these bounds were not sharp, this would imply that that there are bounds sharper than these that contain all values of δ consistent with the data and maintained assumptions. But this is clearly contradicted by the fact that the solutions in Table A.1 are feasible and would fall outside these supposedly sharper bounds. □

A.2 Bounds under additional assumptions

To derive bounds under additional assumptions, we first derive bounds conditional on the strata probabilities.

Lemma A.1. *The bounds on δ for given values of $\boldsymbol{\rho}$ are $\delta \in [\delta_L(\boldsymbol{\rho}), \delta_U(\boldsymbol{\rho})]$, where*

$$\delta_U(\boldsymbol{\rho}) = P_{111}\left(\frac{Q_{11}}{Q_0}\right) - P_{011}\left(\frac{Q_{01}}{Q_0}\right) - P_{110}\left(\frac{1-Q_{11}}{1-Q_0}\right) + P_{010}\left(\frac{1-Q_{01}}{1-Q_0}\right) + \frac{1}{Q_0(1-Q_0)}\min\left\{\begin{matrix} P_{110}(1-Q_{11}) \\ \rho_{011}+\rho_{001} \\ Q_0 - P_{111}Q_{11}\end{matrix}\right\} + \frac{1}{Q_0(1-Q_0)}\min\left\{\begin{matrix} P_{011}Q_{01} \\ \rho_{110}+\rho_{010} \\ 1-Q_0-P_{010}(1-Q_{01})\end{matrix}\right\},$$

and,

$$\delta_L(\boldsymbol{\rho}) = P_{111}\left(\frac{Q_{11}}{Q_0}\right) - P_{011}\left(\frac{Q_{01}}{Q_0}\right) - P_{110}\left(\frac{1-Q_{11}}{1-Q_0}\right) + P_{010}\left(\frac{1-Q_{01}}{1-Q_0}\right) + \frac{1}{Q_0(1-Q_0)}\max\left\{\begin{matrix} -P_{111}Q_{11} \\ -\rho_{110}-\rho_{100} \\ -1+Q_0+P_{110}(1-Q_{11})\end{matrix}\right\} + \frac{1}{Q_0(1-Q_0)}\max\left\{\begin{matrix} -P_{011}Q_{01} \\ -\rho_{001}-\rho_{101} \\ -Q_0+P_{011}Q_{01}\end{matrix}\right\}.$$

Proof. Conditional on ρ_s and Q_0, deriving the bounds on δ is a standard linear programming problem. We now describe the process for deriving these bounds at a general level. Without any as-

sumptions, we are interested in maximizing or minimizing the objective function,

$$\begin{aligned}\delta =& P_{111}\left(\frac{Q_{11}}{Q_0}\right) - P_{011}\left(\frac{Q_{01}}{Q_0}\right) - P_{110}\left(\frac{1-Q_{11}}{1-Q_0}\right) + P_{010}\left(\frac{1-Q_{01}}{1-Q_0}\right) \\ &+ \mu_{011}(1,1)\frac{\rho_{011}}{Q_0(1-Q_0)} + \mu_{001}(1,1)\frac{\rho_{001}}{Q_0(1-Q_0)} \\ &- \mu_{110}(1,1)\frac{\rho_{110}}{Q_0(1-Q_0)} - \mu_{100}(1,1)\frac{\rho_{100}}{Q_0(1-Q_0)} \\ &+ \mu_{110}(0,1)\frac{\rho_{110}}{Q_0(1-Q_0)} + \mu_{010}(0,1)\frac{\rho_{010}}{Q_0(1-Q_0)} \\ &- \mu_{101}(0,1)\frac{\rho_{101}}{Q_0(1-Q_0)} - \mu_{001}(0,1)\frac{\rho_{001}}{Q_0(1-Q_0)},\end{aligned}$$

subject to the constraints

$$\begin{aligned}P_{111}Q_{11} &= \mu_{111}(1,1)\rho_{111} + \mu_{101}(1,1)\rho_{101} + \mu_{110}(1,1)\rho_{110} + \mu_{100}(1,1)\rho_{100} \\ P_{011}Q_{01} &= \mu_{111}(0,1)\rho_{111} + \mu_{011}(0,1)\rho_{011} + \mu_{110}(0,1)\rho_{110} + \mu_{010}(0,1)\rho_{010} \\ P_{110}(1-Q_{11}) &= \mu_{011}(1,1)\rho_{011} + \mu_{110}(1,1)\rho_{110} + \mu_{010}(1,1)\rho_{010} + \mu_{000}(1,1)\rho_{000} \\ P_{010}(1-Q_{01}) &= \mu_{101}(0,1)\rho_{101} + \mu_{001}(0,1)\rho_{001} + \mu_{100}(0,1)\rho_{100} + \mu_{000}(0,1)\rho_{000} \\ 0 \leq \mu_s(t) &\leq 1, \quad \forall s,t.\end{aligned}$$

For this step, we do not need to specify constraints on ρ_s because we consider them fixed (and Q_0 is a linear function of ρ_s). The simplex tableau method yields the given bounds. □

Proof of Proposition 2. Recall the constraints on the strata probabilities:

$$\begin{aligned}Q_{11} &= \rho_{111} + \rho_{101} + \rho_{110} + \rho_{100} \\ Q_{01} &= \rho_{111} + \rho_{011} + \rho_{110} + \rho_{010} \\ Q_0 &= \rho_{111} + \rho_{011} + \rho_{101} + \rho_{001},\end{aligned}$$

Under monotonicity, we only have strata $S_i \in \{111, 110, 010, 100, 000\}$, so we have $Q_0 = \mathbb{P}[D_i(0) = 1] = \rho_{111}$ and $\rho_{110} + \rho_{010} = Q_{01} - Q_0$ and $\rho_{110} + \rho_{100} = Q_{11} - Q_0$. Plugging these values into the bounds from Lemma A.1, we obtain

$$\begin{aligned}\delta_U(Q_0) = & P_{111}\left(\frac{Q_{11}}{Q_0}\right) - P_{011}\left(\frac{Q_{01}}{Q_0}\right) - P_{110}\left(\frac{1-Q_{11}}{1-Q_0}\right) + P_{010}\left(\frac{1-Q_{01}}{1-Q_0}\right) \\ &+ \frac{1}{Q_0(1-Q_0)}\min\begin{Bmatrix} P_{110}(1-Q_{11}) \\ 0 \\ Q_0 - P_{111}Q_{11}\end{Bmatrix} + \frac{1}{Q_0(1-Q_0)}\min\begin{Bmatrix} P_{011}Q_{01} \\ Q_{01} - Q_0 \\ 1 - Q_0 - P_{010}(1-Q_{01})\end{Bmatrix},\end{aligned}$$

and,

$$\delta_L(Q_0) = P_{111}\left(\frac{Q_{11}}{Q_0}\right) - P_{011}\left(\frac{Q_{01}}{Q_0}\right) - P_{110}\left(\frac{1-Q_{11}}{1-Q_0}\right) + P_{010}\left(\frac{1-Q_{01}}{1-Q_0}\right)$$
$$+ \frac{1}{Q_0(1-Q_0)} \max\left\{\begin{array}{c} -P_{111}Q_{11} \\ -Q_{11}-Q_0 \\ -1+Q_0+P_{110}(1-Q_{11}) \end{array}\right\} + \frac{1}{Q_0(1-Q_0)} \max\left\{\begin{array}{c} -P_{011}Q_{01} \\ 0 \\ -Q_0+P_{011}Q_{01} \end{array}\right\}.$$

We further simplify the upper bound expression by noting that $P_{110}(1-Q_{11}) \geq 0$ and $Q_{01} - Q_0 \leq 1 - Q_0 - P_{010}(1-Q_{01})$. The lower bound simplifies because $Q_{11} - Q_0 \leq 1 - Q_0 - P_{110}(1-Q_{11})$ and $P_{011}Q_{01} \geq 0$. Removing these extraenous conditions gives the result in the text.

□

Proof of Proposition 3. Under the maintained assumptions, $Q_0 = Q_{01}$, which we plug into the expression of Proposition 2. Then, the result is immediate upon noting that $P_{011}Q_{01} - Q_{01} \leq 0$, $P_{011}Q_{01} \geq 0$ and rearranging terms. □

For calculating the bounds under the sensitivity constraints, we can take the bounds from Lemma A.1 and solve a corresponding linear programming problem to optimize them with respect to the principal strata probabilities. For example, depending the observed data, the upper bound will depend on $\rho_{011} + \rho_{001}$, $\rho_{110} + \rho_{010}$, or $\rho_{011} + \rho_{001} + \rho_{110} + \rho_{010}$. To find the upper bound across values of $\boldsymbol{\rho}$, we apply the linear programming machinery to finding the upper bound for each of these quantities subject to the constraints that

$$Q_{11} = \rho_{111} + \rho_{101} + \rho_{110} + \rho_{100}$$
$$Q_{01} = \rho_{111} + \rho_{011} + \rho_{110} + \rho_{010}$$
$$Q_0 = \rho_{111} + \rho_{011} + \rho_{101} + \rho_{001},$$

where $0 \leq \rho_s \leq 1$ for all s and $\sum_{s\in\mathcal{S}} \rho_s = 1$. Note that for the sensitivity analysis, we may impose additional constraints on ρ_s in this step. As an example, for the objection function of $\rho_{011} + \rho_{001}$, we have the upper bound

$$\min\left\{1 - Q_{11}, Q_0, 1 - Q_{01} + Q_0, \frac{1}{2}(1 - Q_{11} + Q_0), 1 + Q_{01} - Q_{11}\right\}.$$

Plugging these bounds into the upper bound $\delta_U(\boldsymbol{\rho})$ will yield an upper bound purely as a function of observed parameters and Q_0 and γ (the sensitivity parameter). Under some of our assumptions, inspection of the resulting functions reveals that the maximum of these functions can only occur at a handful of critical values of Q_0 which can be evaluated and compared quickly. Otherwise, we use a standard optimization routine to find the value of Q_0 that maximizes the upper bound or minimizes the lower bound.

A.3 Post-treatment attention checks

Proof of Proposition 4. First, we write the CATE as follows:

$$\begin{aligned}\tau(1) = \frac{P_1 - P_0}{Q_0} &- (\mu_{110}(1,1) - \mu_{110}(0,1))\,\frac{\rho_{110}}{Q_0} - (\mu_{010}(1,1) - \mu_{010}(0,1))\,\frac{\rho_{010}}{Q_0} \\ &- (\mu_{100}(1,1) - \mu_{100}(0,1))\,\frac{\rho_{100}}{Q_0} - (\mu_{000}(1,1) - \mu_{000}(0,1))\,\frac{\rho_{000}}{Q_0}\end{aligned}$$

Under the attention monotonicity assumption, we have

$$\rho_{110} = \rho_{010} = \rho_{100} = 0,$$

and combined with the assumption of no inattentive effects implies

$$\tau(1) = \frac{P_1 - P_0}{\rho_{111} + \rho_{011} + \rho_{101} + \rho_{001}}.$$

To find the bounds for this function, we need to find the minimum and maximum values of $\rho_{111} + \rho_{011} + \rho_{101} + \rho_{001}$ subject to the constraints implied by the assumptions that $Q_{11} = \rho_{111} + \rho_{101}$, $Q_{01} = \rho_{111} + \rho_{011}$, and $\rho_{111} + \rho_{011} + \rho_{101} + \rho_{001} + \rho_{000} = 1$. Using the simplex tableau method, we find the minimum value of this function is $\max(Q_{01}, Q_{11})$ and the maximum value it could take is 1. Thus, we have

$$\tau(1) \in \left[P_1 - P_0, \frac{P_1 - P_0}{\max(Q_{01}, Q_{11})}\right].$$

□

If we want to weaken the assumptions of Proposition 4, to focus solely on randomization, we simply use the simplex tableau method in the proof of Proposition 1 with the objective function $\pi_{11} - \pi_{01}$. Doing so leads to the bounds

$$\pi_{11} - \pi_{01} \in \left[\max\left(-1, \frac{P_1 - P_0 - (1 - Q_0)}{Q_0}, -\frac{P_0}{Q_0}, -\frac{1 - P_1}{Q_0}\right), \right. \\ \left. \min\left(1, \frac{P_1 - P_0 + (1 - Q_0)}{Q_0}, \frac{P_1}{Q_0}, \frac{1 - P_0}{Q_0}\right) \right]. \tag{12}$$

B MCMC Algorithm

In this section we describe our MCMC algorithm for the Bayesian model of Section 6. Our goal is to sample from the joint distribution of the parameters and the principal strata indicator,

$$\mathbb{P}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{S} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{D}) \propto \\ \prod_{i=1}^{n} \left(\sum_{s \in \mathcal{S}_i} \left[\mathbb{P}(Y_i \mid T_i, Z_i, S_i = s, \boldsymbol{X}_i') \mathbb{P}(S_i = s \mid \boldsymbol{X}_i') \right]^{\mathbb{I}(S_i = s)} \right) \mathbb{P}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi}),$$

where $\mathcal{S}_i = \mathcal{S}(T_i, Z_i, D_i)$ are the set of principal strata to which unit i could possibly belong. Traditionally, Bayesian inference for logistic regression models has been complicated and challenging due to a lack of a simple Gibbs sampling algorithm. Recently, Polson, Scott and Windle (2013) introduced a simple data-augmentation strategy based on the Pólya-Gamma (PG) distribution, obviating the need for approximate methods or precise tuning of a Metropolis-Hastings algorithm. We use this approach for both the binary and multinomial logistic regression models for the outcome and principal strata, respectively. This allows a simple Gibbs structure where the full conditional posterior distributions of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\boldsymbol{\psi}$ are Normal conditional on specific draws from the PG distribution. Conditional on the other parameters, then the full conditional posterior of the principal strata follows a similar form to Hirano et al. (2000),

$$\mathbb{P}(S_i = s \mid Y_i, \boldsymbol{X}_i, T_i, Z_i, D_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi}) = \frac{\mu_{is}(T_i, Z_i)^{Y_i}(1 - \mu_{is}(T_i, Z_i))^{1 - Y_i} \rho_{is}}{\sum_{k \in \mathcal{S}_i} \mu_{ik}(T_i, Z_i)^{Y_i}(1 - \mu_{ik}(T_i, Z_i))^{1 - Y_i} \rho_{ik}},$$

where we suppress the dependence of μ_{is} and ρ_{is} on the model parameters. Repeatedly drawing from these full conditional posterior distributions should provide a sample from the above joint

posterior and allow for posterior inference in the usual manner. In each iteration, $r \in \{1, \ldots, R\}$, of the algorithm, we have draws

$$\left(\left\{ \widehat{S}_i^{(r)} \right\}_{i=1}^{n}, \widehat{\boldsymbol{\psi}}^{(r)}, \widehat{\boldsymbol{\alpha}}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)} \right).$$

We can use these draws to generate draws of the population and in-sample versions of the quantity of interest. Given that $\widehat{S}_i^{(r)}$ is the imputed principal strata imputed for unit i in the rth draw from the posterior, we let

$$\widehat{\mu}_i^{(r)}(t, z) = \widehat{\mu}_{i, \widehat{S}_i^{(r)}}(t, z, \widehat{\boldsymbol{\alpha}}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)})$$

be the mean of the potential outcomes conditional on that imputed principal strata. Furthermore, let $\widehat{\rho}_{is}^{(r)}$ be the rth draw of the predicted probabilities of each principal strata for each unit. Then, we can calculate the population quantity as

$$\widehat{\delta}_p^{(r)} = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{s \in \mathcal{S}_1^*} (\widehat{\mu}_{is}^{(r)}(1,1) - \widehat{\mu}_{is}^{(r)}(0,1)) \widehat{\rho}_{is}^{(r)} \right) - \left(\sum_{s \in \mathcal{S}_0^*} (\widehat{\mu}_{is}^{(r)}(1,1) - \widehat{\mu}_{is}^{(r)}(0,1)) \widehat{\rho}_{is}^{(r)} \right),$$

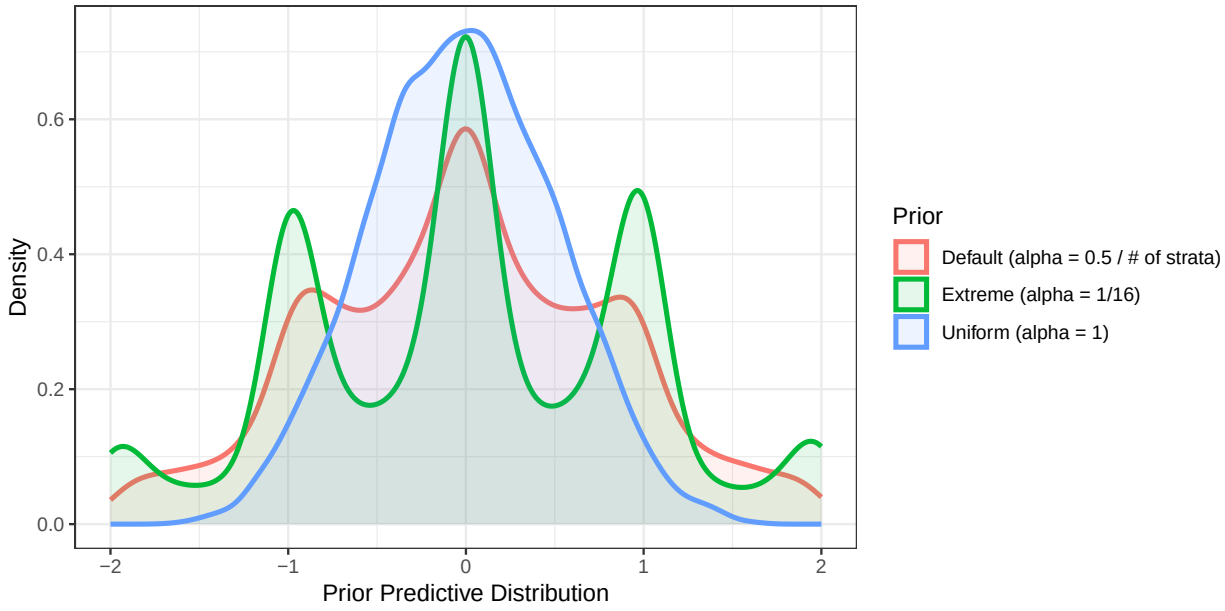
For the in-sample quantity, we can then draw *imputed* values of the missing potential outcomes themselves $\widehat{Y}_i^{(r)}(1,1) \sim \text{Bin}(\widehat{\mu}_i^{(r)}(1,1))$ and $\widehat{Y}_i^{(r)}(0,1) \sim \text{Bin}(\widehat{\mu}_i^{(r)}(0,1))$. We can combine this with the imputed value of $D_i(0)$, which mechanically derives from $\widehat{S}_i^r$, to get the rth draw from the posterior of δ_s,

$$\widehat{\delta}_s^{(r)} = \frac{\sum_{i=1}^{n} \widehat{D}_i^{(r)}(0) \left\{ \widehat{Y}_i^{(r)}(1,1) - \widehat{Y}_i^{(r)}(0,1) \right\}}{\sum_{i=1}^{n} \widehat{D}_i^{(r)}(0)} - \frac{\sum_{i=1}^{n} \left(1 - \widehat{D}_i^{(r)}(0)\right) \left\{ \widehat{Y}_i^{(r)}(1,1) - \widehat{Y}_i^{(r)}(0,1) \right\}}{\sum_{i=1}^{n} \left(1 - \widehat{D}_i^{(r)}(0)\right)}.$$

Broadly speaking, we would not expect very large differences between these two targets, except for slightly less posterior variance for the in-sample version.

As discussed in the main text, the priors need careful attention because they drive the identification of the parameters that are unidentified by the likelihood. One additional complication comes from how the ultimate quantity of interest is a function of the parameters so we cannot directly place, for example, a uniform prior on δ. Figure A.5 shows the prior predictive distribution for interaction with three different priors when monotonicity and stable moderator under control are

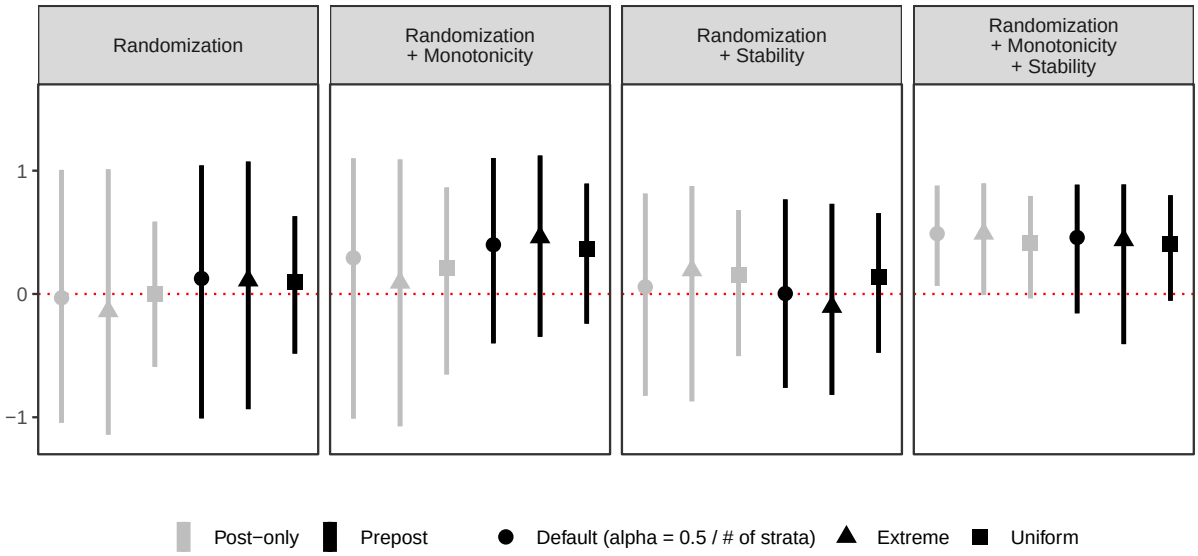
Figure A.5: Prior predictive distribution of the parameter under three different prior distributions: (red) the default priors that scales a Jeffreys prior by the number of principal strata; (blue) a uniform prior on all parameters; and (green) a more extreme prior that has $\alpha = 1/16$.



assumed and there are no covariates. The uniform prior on all parameters results in a prior on δ that has more density in the center of identified range than we might expect. This result is similar to how sums of uniform random variables are not themselves uniform. We can counteract this issue by reducing the Dirichlet and Beta hyperparameters below 1 to put more density at extreme values of the parameters compared to the center. Dropping these parameters down to 1/16 (in green) leads to more mass on strata means closer to 0 or 1 and strata probabilities closer to 0 and 1. In terms of the interaction, this leads to more mass at the values -2, -1, 0, 1, and 2. Our default prior (red) is one that scales the hyperparameters by the inverse of the number of strata to achieve something closer to a uniform distribution.

Additionally, we re-ran the Gibbs empirical analysis of the Horowitz and Klaus (2020) study, adjusting the priors to the extreme values or uniform values in the previous simulation. The results are displayed in Figure A.6, demonstrating the general consistency of the point estimates across starting priors, although there is some fluctuations in the variance of the results.

Figure A.6: Comparing Bayesian estimates for δ for default, extreme, and uniform priors



Notes: Figure shows posterior means and 95% credible intervals for δ under different sets of assumptions, applied to either the post-only data (grey) or the combined pre-post data (black). Estimates are shown with default, extreme, and uniform priors, (denoted by circles, triangles, and squares, respectively). We follow the original authors in using age, gender, education, and closeness to one's ethnic group as covariates. The naïve OLS estimates are included for comparison.

B.1 Simulation Evidence for the Bayesian Approach

Simulation Study I. In the first simulation study, we generate simulated data with $n = 1000$ constructed using a data generating process that matches the Bayesian posterior, pre-specifying coefficient values for the outcome and principal strata models, randomly drawing values of Z, T, and three covariates X_1, X_2, and X_3, and generating values of Y and D from the models. Tables A.2 and A.3 in the Appendix display the β coefficients for the outcome and ψ coefficients for the for the true data generating process (DGP). The DGP assumes that monotonicity and stable moderator under control both hold so that there are three feasible strata ($\mathcal{S} = \{000, 100, 111\}$). Thus, in this setting it would be most appropriate to incorporate both assumptions into the MCMC algorithm for sampling from the posterior distribution. Since these assumptions narrow the nonparametric bounds, we expect the assumptions to reduce variance of the posterior distribution of δ.

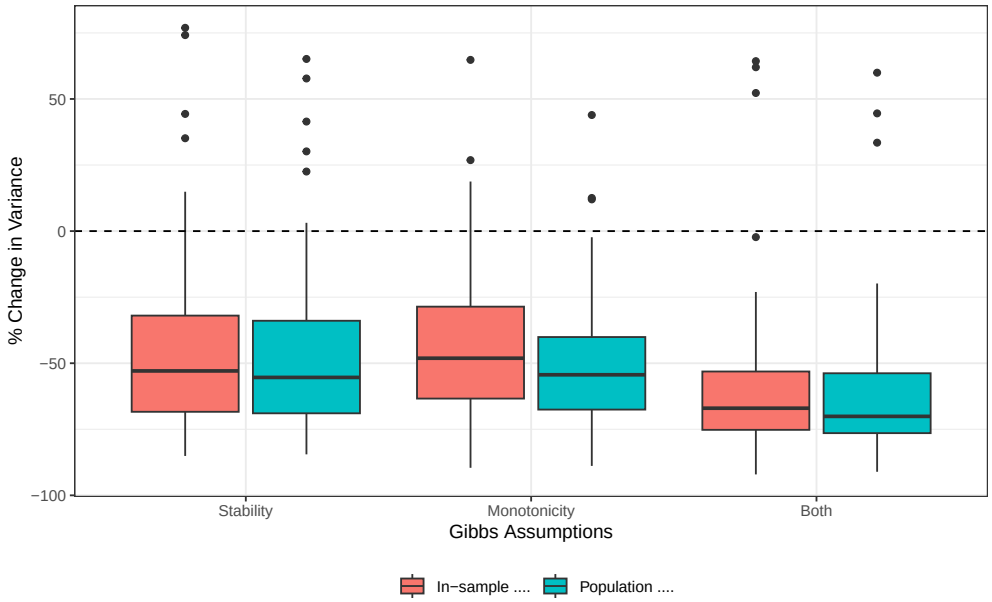


Figure A.7: Variance reduction from different combinations of assumptions. Boxplots present distribution of % variance reduction of δ from the MCMC algorithm with the labeled assumptions compared to same algorithm with no assumptions, across 1,000 draws of simulated data. For each simulation iteration, 4 chains were run for each combination of assumptions. MCMC parameters: 2,000 iterations, 200 burn-in, 2 thinning parameter, simulated data $n = 1,000$.

To test this, we perform a Monte Carlo simulation with 1,000 iterations. For each iteration, we calculate the posterior distribution of δ with the same data across four different versions of the MCMC algorithm: enforcing just the monotonicity assumption, enforcing just the stable moderator under control assumption, enforcing neither assumption, and enforcing both assumptions. Each run of our MCMC algorithm consists of 4 chains with 2,000 iterations each, 200 burn-in (or warm-up) iterations, and a thinning parameter of 2. Both in-sample and population δ values are calculated at each iteration and the variance of the posterior is calculated from a sample of 1,000 draws from the posterior. This is done for each of the 1,000 simulated datasets, and for each dataset we compute the percent reduction in variance compared to the MCMC algorithm with no assumptions when using the algorithm with the monotonicity assumption, the stable moderator under control assumption, or both assumptions.

Figure A.7 presents boxplots for the distribution of reductions in variance for each combination of assumptions. Both the monotonicity and stable moderator assumptions on their own reduce the

variance compared to no assumptions, while making both assumptions reduces the variance even further. The monotonicity assumption showed a median posterior variance reduction of 40.8% for the in-sample δ and 42.0% for the population δ. The stable moderator under control assumption on reduced the posterior variance by a median reduction of 47.1% (in-sample δ) and 50.4% (population δ). The MCMC algorithm with both assumptions exhibited a posterior variance reduction of 59.3% (in-sample δ) and 61.7% (population δ).

Simulation Study II. In the second simulation study, we drew a series of simulated datasets under different conditions where the covariates had a weak, medium, or strong correspondence with the outcome and principal strata in the data generating processes. Thus, there were six total conditions: Weak, Medium, and Strong influence in the outcome DGP; and Weak, Medium, and Strong influence in the principal strata DGP. The values of the coefficients for these conditions are β and ψ values of 0, 0.25, and 0.5, respectively. When varying the influence of covariates in the outcome DGP, the influence of covariates on the strata was held constant, and the influence in the outcome model was similarly held constant when varying influence in the strata DGP. Fixed values of the β's and ψ's are shown in the Appendix in Tables A.4 and A.5.

For each condition, we drew 1,000 simulated datasets and ran the MCMC algorithm twice: one time incorporating covariates and one time omitting them. Each MCMC run consisted of the same iterations, burn-in, and thinning parameters as in the previous simulation study. We again calculate in-sample and population δ values for each iteration of the Gibbs and calculate the variance of the posterior distribution and the % variance reduction comparing the Gibbs with covariates to that without. Figure A.8 presents boxplots for the distribution in variance reduction. When we vary the influence of covariates on the outcome, we see a clear variance reduction in all conditions, and we observe a larger reduction as the influence of covariates on the outcome in the DGP increases. When testing the impact of incorporating covariates across different levels of influence in the DGP on the strata, the pattern is less pronounced, with overall reduction increases in all conditions but slightly lower reductions in the Medium than Weak condition. The Strong condition still has the largest variance reduction overall, however, so in general the efficiency gains from incorporating

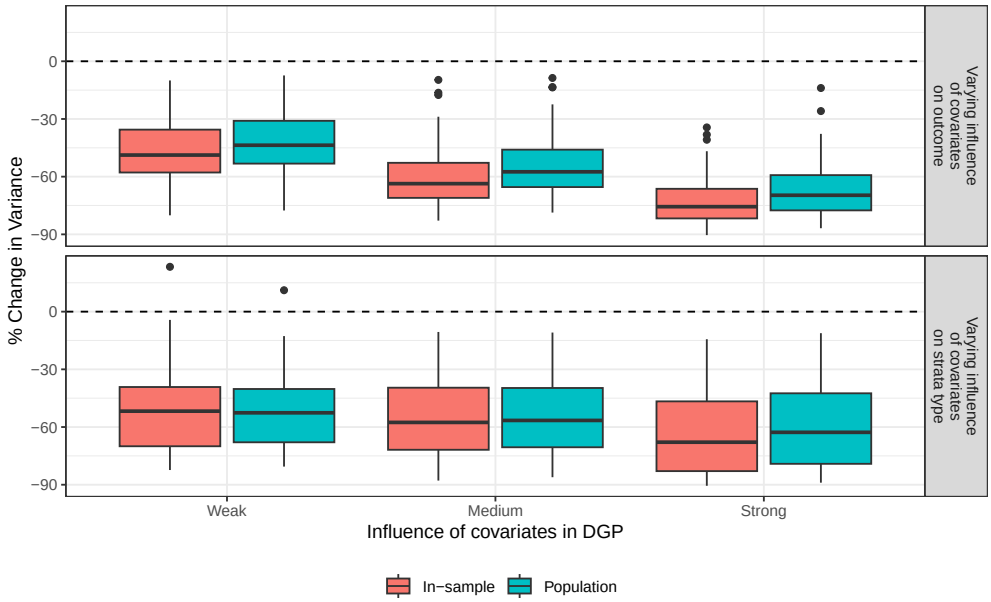


Figure A.8: Variance reduction from incorporation of covariates. Boxplots present distribution of % variance reduction of δ from Gibbs with covariates compared to MCMC without covariates, across 1,000 draws of simulated data. For each draw 4 MCMC chains were run for each combination of assumptions. MCMC parameters: 2,000 iterations, 200 burn-in, 2 thinning parameter, simulated data $n = 1,000$.

covariates are increasing as the influence of covariates on strata in the data increases.

C Additional Simulation Details

Table A.2: β Values for DGP in Bayesian Assumptions Simulation

Variable	β
(Intercept)	-2.00
X1	1.00
X2	0.15
X3 (Medium)	0.24
X3 (Large)	0.28
T	0.83
Z	-0.01
T:Z	0.11
S111	0.41
S100	0.62
T:S111	0.01
T:S100	0.23
Z:S111	0.20
Z:S100	-0.02
T:Z:S111	-0.90
T:Z:S100	0.09

Table A.3: ψ Values for DGP in Bayesian Assumptions Simulation

	S111	s100	s000
(Intercept)	-2.06	-1.00	0.00
X1	2.00	1.50	0.00
X2	0.50	0.17	0.00
X3 (Medium)	1.35	-0.28	0.00
X3 (Large)	1.75	-1.01	0.00

Table A.4: Fixed β Values in Covariate Simulation

Variable	β
(Intercept)	-1.00
X1	1.00
X2	0.50
X3 (Medium)	0.50
X3 (Large)	0.28
T	0.83
Z	-0.01
T:Z	0.11
S111	0.41
S100	0.62
T:S111	2.00
T:S100	-0.13
Z:S111	0.50
Z:S100	0.10
T:Z:S111	0.05
T:Z:S100	0.01

Table A.5: Fixed ψ values in Bayesian Covariate Simulation

	S111	S100	S000
(Intercept)	-2.06	-1.00	0.00
X1	2.00	1.50	0.00
X2	0.50	0.17	0.00
X3 (Medium)	1.35	-0.28	0.00
X3 (Large)	1.75	-1.01	0.00

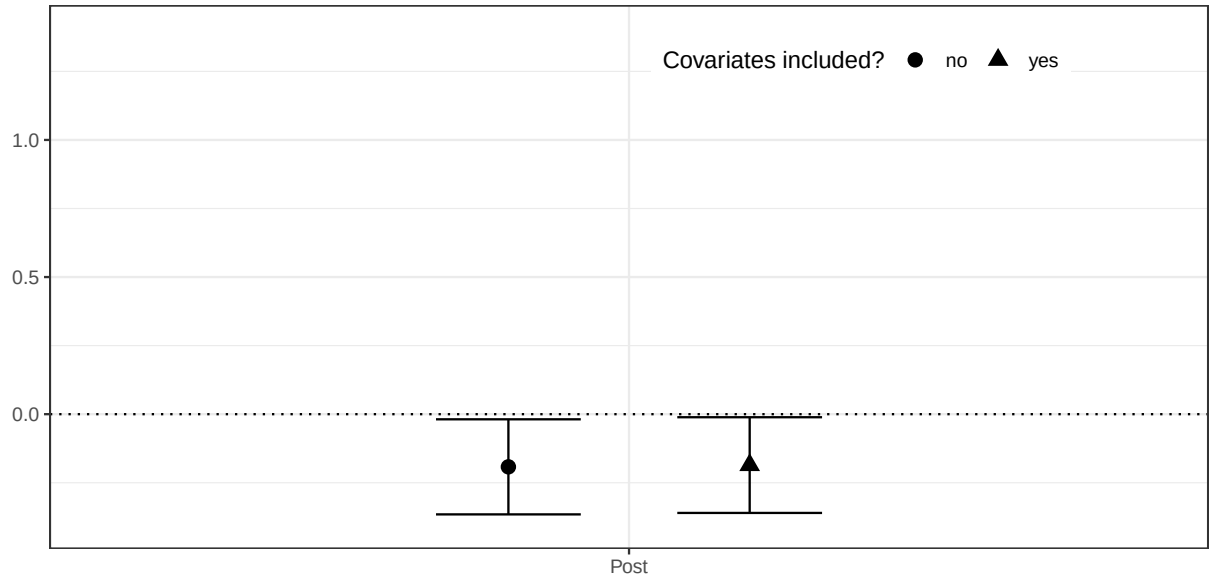
D Empirical Application: Post-treatment Attention Checks

Here, we present a second empirical application demonstrating how our methodology can be used to diagnose how dropping experimental subjects who fail post-treatment attention or manipulation checks influences experimental estimates. To do so, we use data collected by Aronow, Baron and Pinson (2018) in their replication of Press, Sagan and Valentino (2013). In both the original study and the replication, the researchers conduct a survey experiment where respondents are randomly assigned to read a vignette describing situations where Al Qaeda in Lebanon and Syria obtain weapons-grade uranium to make nuclear weapons for use against the United States. The vignettes vary in the reported likelihood of a nuclear versus conventional weapons strike on this Al Qaeda facility. Following Aronow, Baron and Pinson (2018) , we focus on the three conditions which report the likelihood of nuclear versus conventional weapons strike success at 90%-90%, 90%-70%, and 90%-45%. For the outcome, we focus on reported preference for a nuclear strike over a conventional weapons strike. Prior to reading the vignette treatments, subjects were asked a manipulation check asking them to verify which probability values they had just seen in their assigned vignette. In the original study, subjects that failed the manipulation check in the analysis, while in Aronow, Baron and Pinson (2018) the impact of dropping these subjects was the focus of the replication.

In our analysis, we code treatment as binary, coded as 1 if the subjects were assigned the 90%/90% vignette and 0 if they were assigned the other two vignettes which present a higher relative probability of success for a nuclear strike relative to a conventional strike. The quantity of interest in our analysis is the interaction between treatment and passing the manipulation check on preference for a nuclear strike. Figure A.9 presents the OLS estimates for this interaction, which show that passing the manipulation check is associated with a larger decreased preference for a nuclear strike in the 90%/90% condition relative to the other two conditions. We include a simple OLS interaction model and one with binary covariates for whether subjects are interested in politics, interested in news, and are Republicans.

Figure A.10 displays the non-parametric bounds (with 95% confidence intervals) and Bayesian estimates (posterior means with 95% credible intervals) for δ under different sets of assumptions,

Figure A.9: Estimates of treatment-moderator interaction using pre-test, post-test, and combined data.

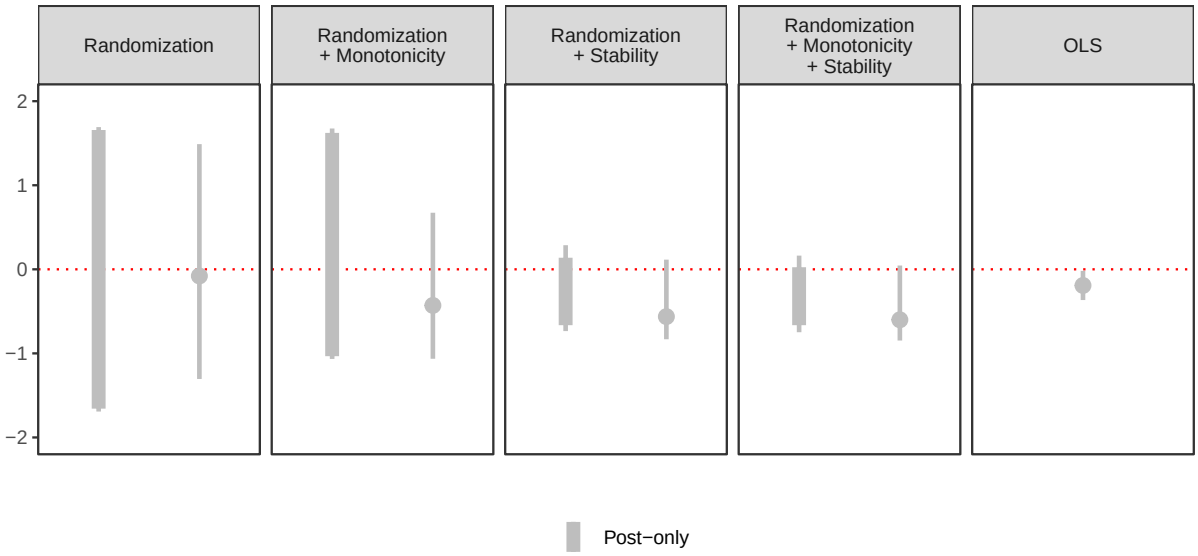


Notes: Figure shows point estimates and 95% confidence intervals for the interaction between treatment and passing the manipulation check on preference for a nuclear strike from a linear probability model, using the data (moderator measured after treatment)from Aronow, Baron and Pinson (2018). Estimates are shown from models with and without covariates. We use political interest, news interest, and party identification as covariates.

applied to the data. We also include the naïve OLS estimate with 95% confidence interval for comparison in the final panel.

Assuming only randomization, the nonparametric bounds and the Bayesian credible intervals are not informative of the sign of δ, and are much wider than the confidence interval of the naïve OLS estimate. Without incorporating information from covariates or imposing substantive assumptions, there is little evidence for the theorized interaction, where respondents who fail the manipulation check express lower preference for a nuclear strike in the treatment condition compared to the other two vignettes. Adding the assumption of monotonicity reduces the width of the Bayesian credible intervals shifts the posterior mean from zero to negative and closer to the estimates from the OLS models. Adding the assumption of stability further tightens the bounds and credible intervals, and with all three assumptions (randomization, monotonicity, and stability), the width of the nonparametric bounds and Bayesian credible intervals come close to excluding zero under conventional

Figure A.10: Comparing non-parametric bounds and Bayesian estimates for δ under different assumptions



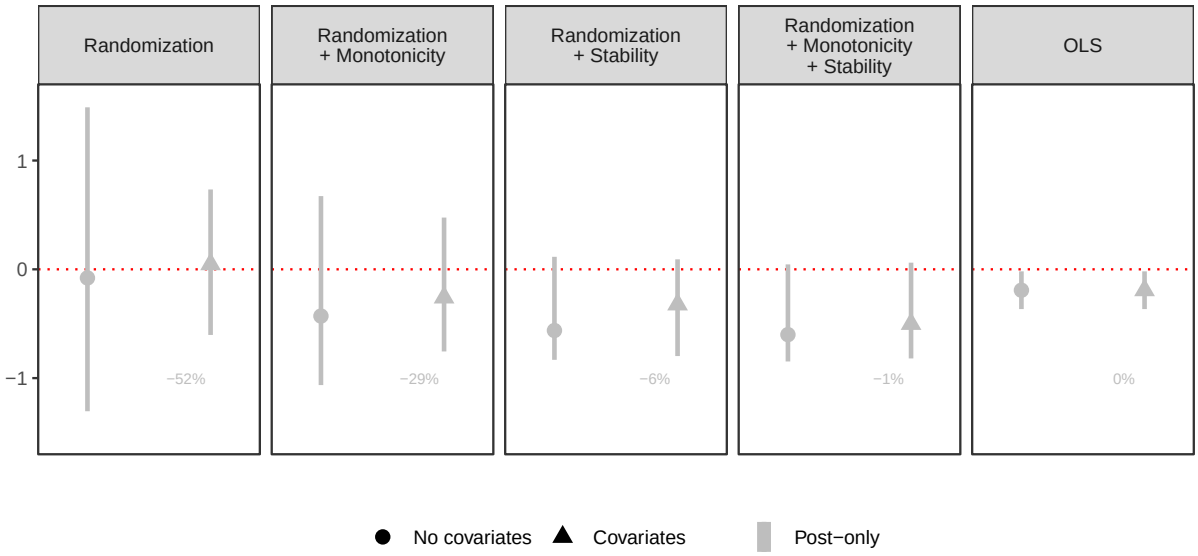
Notes: Figure shows non-parametric bounds and Bayesian estimates of the quantity of interest under different sets of assumptions, applied to the data (moderator measured after treatment)from Aronow, Baron and Pinson (2018). The thick bars denote the width of the bounds, and thinner lines denote the 95% confidence intervals around the bounds. Across the first four panels, the thin lines with dots denote the Bayesian posterior mean and 95% credible interval. No covariates were included in this estimates, to facilitate comparison with the nonparametric bounds. For the final panel ("OLS"), the thin lines with dots denote the OLS estimate and 95% confidence interval, which is included for reference.

(95% bounds and credible intervals) threshholds.

Thus, in our setup, we find evidence consistent with Aronow, Baron and Pinson (2018) that dropping subjects who fail manipulation checks may alter the results from Press, Sagan and Valentino (2013), although the variance around our estimates even under the full set of assumptions do not allow for overly strong takeaways about the impact of dropping subjects in this context. Still, this application demonstrates how our methodology could be applied to diagnose the measurement challenge of manipulation checks, and further demonstrates how different assumptions narrow the bounds and clarify the posterior means in our methodology.

Figure A.11 plots the Gibbs results with and without covariates, which further demonstrate how covariates can be used, in some case, to reduce the width of the bounds and credible intervals.

Figure A.11: Comparing Bayesian estimates for δ with and without covariates



Notes: Figure shows posterior means and 95% credible intervals for δ under different sets of assumptions, applied to either the post-only data (grey) or the combined pre-post data (black). Estimates are shown with and without the inclusion of covariates (denoted by triangles and circles, respectively), and the numbers indicate the reduction in the width of the credible intervals due to the inclusion of covariates for the post-only data (in grey) and the combined pre-post data (in black). We follow the original authors in using XYZ as covariates. The naïve OLS estimates are included for comparison.

We see the largest reductions under the assumptions of just randomization or randomization and monotonicity, and smaller reductions with stability or the full set of assumptions.