Estimating Controlled Direct Effects with Panel Data: An Application to Reducing Support for Discriminatory Policies*

Matthew Blackwell[†] Adam Glynn[‡] Hanno Hilbig[§]
Connor Halloran Phillips[¶]

February 27, 2025

Abstract

Recent experimental studies in the social sciences have demonstrated that perspective-taking conversations are effective at reducing prejudicial attitudes and support for discriminatory policies. We ask if such interventions can directly affect policy views without changing prejudice. Unfortunately, the identification of the controlled direct effect—the natural causal quantity of interest for this question—has required strong selection-on-observables assumptions for any mediator. We leverage a recent experimental study with multiple survey waves of follow-up to identify and estimate the controlled direct effect using the changes in the outcome and mediator over time assuming parallel trends in the potential outcomes. This design allows us to weaken the identification assumptions to allow for linear and time-constant unmeasured confounding between the mediator and the outcome. We develop a semiparametrically efficient and doubly robust estimator for these quantities along with a sensitivity analysis for the key identifying assumption of parallel trends. Contrary to what traditional methods find, our approach estimates a controlled direct effect of perspective-taking conversations when subjective feelings are neutral but not positive or negative, and this result is robust to moderate departures from parallel trends.

Keywords: causal inference, controlled direct effects, difference-in-differences, panel data, political science, discrimination

*Thanks to David Broockman, Kosuke Imai, Josh Kalla, Dean Knox, Soichiro Yamauchi, and Yiqing Xu for comments and suggestions. Working paper, comments welcome. Software to implement the methods in this paper can be found in the `DirectEffects` R package. This paper was originally titled "Difference-in-differences Designs for Controlled Direct Effects."

[†]Department of Government and Institute for Quantitative Social Science, Harvard University. web: http://www.mattblackwell.org, email: mblackwell@gov.harvard.edu

[‡]Department of Political Science and Institute for Quantitative Theory and Methods, Emory University. email: aglynn@emory.edu

[§]Department of Political Science, University of California, Davis. email: hhilbig@ucdavis.edu

[¶]Department of Political Science and Center for Effective Lawmaking, Vanderbilt University. email: connor.phillips@vanderbilt.edu

1 Introduction

Many scholars and citizens view discrimination on the basis of identity as incompatible with an egalitarian society. Thus, a large literature in the social sciences has explored the relationship between prejudicial views about identity groups and support for policies that discriminate against those groups (Kinder and Sears, 1981; Sniderman et al., 1991; Sidanius, Pratto and Bobo, 1996; Kinder and Sanders, 1996; Nelson, 1999; Krysan, 2000; Rabinowitz et al., 2009). Across a wide variety of settings, this work has shown that people who hold negative views about a group tend to be more supportive of policies that discriminate against that group. This robust finding has led to a somewhat implicit hope that improving the dominant group's prejudicial feelings about a group would decrease support for discriminatory policies and ultimately lead to a more equal society.

Given the theory that prejudice causes discriminatory policy, the social sciences have focused on theories that can explain how societies might reduce prejudice (Paluck and Green, 2009). One of the most prominent of these theories, often called the *contact hypothesis*, holds that intergroup contact can reduce prejudice (Allport, 1954; Pettigrew and Tropp, 2006). The contact hypothesis has been impactful in the policy world, influencing policies such as desegregation and international peace-building efforts (see Paluck, Green and Green, 2019, and references therein for a review). However, support for the contact hypothesis is mixed, especially when focusing on high-quality evidence from randomized control trials (Paluck, Green and Green, 2019). If interventions may or may not be able to reduce prejudicial feelings, or if they reduce prejudicial feelings under specific conditions, it becomes essential to understand if and in what circumstances these interventions might be able to reduce support for discriminatory policies *directly*, without necessarily affecting prejudicial views about the group.

In this paper, we focus on the effects of interventions that encourage respondents to take the perspective of a minority group (Dovidio et al., 2004; Paluck, 2009; Broockman and Kalla, 2016; Simonovits, Kézdi and Kardos, 2018). These interventions mimic the core theoretical mechanisms of contact theory by encouraging participants to better identify with the discrimination that the minority group faces. Several studies in political science have shown that these perspective-taking interventions can improve general attitudes toward those groups *and* increase support for politicians or policies that benefit the group (Broockman and Kalla, 2016; Simonovits, Kézdi and Kardos, 2018; Adida, Lo and Platas, 2018). In this context, we seek to understand if perspective-taking interventions can have a direct effect on support for anti-discrimination policies for fixed values of subjective feelings about a group. Previous studies of perspective-taking have attempted to estimate these direct effects. For example, Adida, Lo and Platas (2018) showed that perspective-taking toward Syrian refugees can increase support for admitting those refugees to the United States and explored how the direct effect varied by a subjective rating of the refugees in the intervention. Paluck (2009) showed that a soap opera with prejudice-reducing messages affected perceptions of social norms and behavior toward outgroups without changing their personal prejudicial beliefs. Unfortunately, inferring direct effects from these studies requires strong "no unmeasured confounders" assumptions that may not hold.

We focus on the experimental setting of Broockman and Kalla (2016), who found that a door-to-door perspective-taking canvassing intervention reduced support for legal discrimination against transgender people (those who identify with a gender different from their sex assigned at birth). The intervention consisted of a 10-minute conversation that encouraged respondents to think about a time when they were judged negatively for being different and asked them to reflect on if and how the conversation changed their minds. We aim to determine if this intervention has a *direct* effect on policy views for fixed feelings of subjective warmth toward transgender people. These direct effects are crucial for understanding persuasion in diverse democracies since they show whether or not personal tolerance of outgroups is required to increase support for *legal* tolerance of those same groups.

1.1 Methodological challenges

We focus on estimating the controlled direct effect (CDE) of the perspective-taking intervention, which is a popular quantity of interest since it relies on weaker identification assumptions than traditional mediation quantities like the natural direct effect (Robins, 1986, 1999; Petersen, Sinisi and van der Laan, 2006; Goetgeluk, Vansteelandt and Goetghebeur, 2008; VanderWeele, 2015; Frölich and Huber, 2017; Zhou and Wodtke, 2019; Blackwell and Strezhnev, 2022) and has been targeted by previous studies of perspective-taking interventions (Adida, Lo and Platas, 2018). Unfortunately, extant methods for estimating controlled direct effects require a strong and often implausible assumption of "no unmeasured confounders" for the mediator-outcome relationship. For example, there are likely unmeasured factors such as cultural beliefs that influence both subjective feelings about disadvantaged groups *and* views on policies about those groups, even conditional on covariates. These unmeasured confounders would bias the standard methods for estimating CDEs even in an experimental setting when researchers randomize treatment.

Fortunately, the Broockman and Kalla study featured a multiwave design, allowing us to relax the no unmeasured confounders assumption. The researchers recruited respondents from a list of registered voters with a mailer for a baseline survey that measured both the mediator and the outcome and then conducted posttreatment surveys at three days, three weeks, six weeks, and three months after administering treatment. With this design, we can focus on the *changes* in the mediator and outcome before and after treatment, which will purge any linear, time-constant, unmeasured confounding between the mediator and the outcome. In this way, our approach is similar to the parallel trends assumptions in difference-in-differences (DID) designs (Heckman, Ichimura and Todd, 1997; Abadie, 2005).

This paper shows how to identify two conditional versions of the controlled direct effect leveraging different flavors of parallel trends assumptions. The two identification results differ in what additional assumptions are required to establish them. For both estimands, we develop doubly robust, semiparametrically efficient estimators for these effects, leveraging propensity score and outcome regression modeling. Furthermore, we show that the cross-fitting strategy of Chernozhukov et al. (2018) is valid in this setting, allowing for weaker conditions on the estimators for the nuisance parameters, easy incorporation of flexible machine learning estimators, and a simple variance estimator.

Given the importance of the parallel trends assumption for identification of the CDE, we also develop a sensitivity analysis for departures from this assumption (Robins, Rotnitzky and Scharfstein, 1999). We bound the strength of unmeasured confounding for the trends in the outcome and mediator, which implies an interpretable set of bounds on the CDE. We can vary the strength of the bound to see how the amount of unmeasured confounding affects our estimated effects. We then compare these results to the amount of unmeasured confounding that would be implied by omitting observed covariates from the study.

We use these tools to show that there is a controlled direct effect of the perspective-taking intervention on support for anti-discrimination policies fixing the value of subjective or prejudicial feelings about transgender people, at least for the group who begins with neutral feelings about that population. Naive methods of estimating average controlled direct effects (ACDEs), such as simply adding the mediator and intermediate confounder to a difference-in-differences model, fail to find this same result. In general, we find that there is considerable variation in the overall and direct effect of treatment across baseline feelings about transgender people and that there is some evidence that the direct effects vary by race. Finally, we find that direct effects appear to persist until the end of the three-month follow-up survey. These results imply that direct effects on policy not through prejudicial feelings might be most potent on respondents with less polarized feelings about the target group. The sensitivity analysis shows that this result is robust to a moderate amount of unmeasured confounding for the trends in the outcome.

A handful of other studies have connected DID designs to direct effects more broadly. Deuchert, Huber and Schelker (2019), Huber, Schelker and Strittmatter (2022), and Holm and Breen (2024) use a principal stratification approach to identify and estimate different mediation quantities under monotonicity and parallel trends assumptions without intermediate covariates. Crucially, those settings focus on the natural direct and indirect effects, the traditional mediation quantities that allow for a decomposition of the overall effect. Instead, we focus on the controlled direct effect, which can generally be identified under weaker conditions, allowing us to sidestep monotonicity assumptions, incorporate nonbinary mediators, and allow baseline and intermediate confounders. Concurrently with our work, Shahn et al. (2022) developed estimation techniques for estimating the parameters of a structural nested mean model under a DID-style assumption similar to ours. Their setting differs from ours in that they focus on estimating the effects of time-varying treatments on outcomes measured between each treatment, using that "in-between" outcome to differentiate between the effects of the treatment and mediator. In our setting, we only observe the outcome after the realization of both the treatment and the mediator.

1.2 Roadmap

The paper proceeds as follows. Section 2 introduces the data and the primary quantities of interest and establishes the core identification results. We introduce the primary estimation strategy in Section 3. In this section we also show how we leverage cross-fitting and develop our sensitivity analysis. Section 4 presents the results of our empirical application, and Section 5 concludes with a discussion.

2 Data, Estimands, and Assumptions

We now describe the empirical setting of the Broockman and Kalla (2016) experiment. The experimenters recruited respondents from a list of registered voters in Miami, Florida ($n = 68,378$) and sent participants a link to an (ostensibly unrelated) online baseline survey via postal mail. The authors then randomly assigned subjects that responded to the survey ($n = 1825$) to receive door-to-door canvassing from a nonprofit organization, either the perspective-taking intervention (treatment, $D_i = 1$) or information about recycling (control, $D_i = 0$). After the experiment, they sent subjects who came to their doors under either condition ($n = 501$) to complete online follow-up surveys three days, three weeks, six weeks, and three months after the intervention.

Our goal is to estimate the direct effect of a treatment (a perspective-taking intervention) on an outcome (views of policies about transgender people) with a mediating variable (subjective feelings about transgender people) set to a particular value. The primary outcome of interest is a seven-point scale measuring support for transgender nondiscrimination laws six weeks after treatment (wave 3), which we denote as Y_{i2}. We denote the baseline outcome as Y_{i1}.

To construct our mediator, we focused on the transgender feeling thermometer, measured on

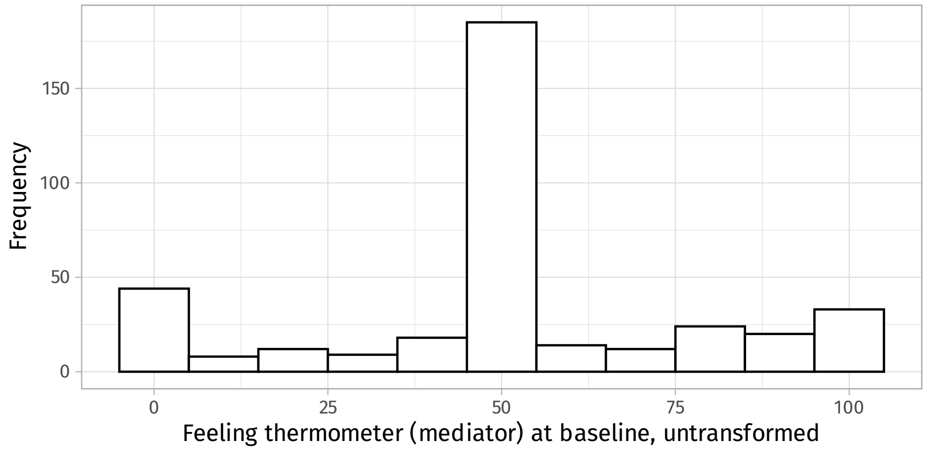

Figure 1: Distribution of the mediator at baseline, prior to discretizing

	Wave 2		
Baseline	Cool	Neutral	Warm
Cool	53 (54.64%)	33 (34.02%)	11 (11.34%)
Neutral	17 (9.83%)	98 (56.65%)	58 (33.53%)
Warm	6 (5.50%)	11 (10.09%)	92 (84.40%)

Table 2: Joint distribution of the baseline and posttreatment mediator after binning into three categories. Row percentages in parentheses.

a scale of 0–100, where higher values indicate "warmer" feelings toward the group. As we show in Figure 1, thermometer scores often show a significant amount of clumping at "even" numbers such as 0, 50, and 100, and much of the informational content could be summarized as a person feeling coolly, warmly, or neutral about a group. Thus, we create our main mediator of interest, M_{i2}, by transforming the three-week posttreatment scores into a three-level discrete variable such that $M_{i2} = 1$ for participants who score below 50 on the thermometer (cooler feelings), $M_{i2} = 2$ for participants who score exactly 50 (neutral feelings), and $M_{i2} = 3$ for participants who score above 50 points (warm feelings). We define the baseline mediator, M_{i1}, similarly. While we focus on a three-level mediator here, our theoretical results are valid for any discrete mediator. Table 2 shows the joint distribution of the baseline and posttreatment mediator after binning. Below, we investigate an alternative categorization that accounts for clumping at 0 and 100, but the results are broadly similar.

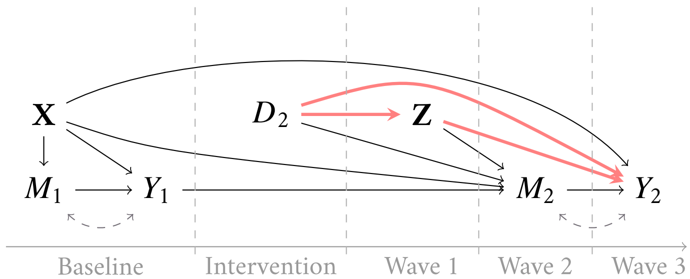

Figure 2: Directed acyclic graph indicating the causal structure of the application and timeline of measurements. Thick red arrows are components of the quantity of interest, the controlled direct effect. Bi-directed arrows between the mediators and outcomes indicate that we allow for unmeasured confounding between those variables.

The multiwave surveys also provide a host of covariates. We define a set of baseline/pretreatment covariates as $\mathbf{X}_i$, which include basic demographics (for example, age, race, and gender), political leanings, and gender-related attitudes. We additionally have posttreatment covariates, $\mathbf{Z}_i$, measured in wave 1 (3 days after treatment), that include changes in various political and gender attitudes from baseline. These posttreatment covariates are measured before the mediator, M_{i2}. We present a complete list of all covariates in Table SM.3 in the Supplemental Materials. We assume the observed data $\mathbf{O}_i = (\mathbf{X}_i, \mathbf{Z}_i, D_i, M_{i1}, M_{i2}, Y_{i1}, Y_{i2})$ is independent and identically distributed across i. Figure 2 shows a directed acyclic graph with the causal structure and measurement timing for this application.

Let $Y_{it}(d, m)$ be the potential outcome for a unit with treatment set to $D_i = d$ and mediator set to $M_{it} = m$. We assume the usual consistency assumption that we observe the potential outcome of the observed treatment and mediator, or $Y_{it} = Y_{it}(D_i, M_{it})$. There are potential versions of the intermediate covariates and posttreatment mediator as well, $\mathbf{Z}_i(d)$ and $M_{i2}(d)$, with similar consistency assumptions. We make a standard no anticipation assumption that M_{i1} and Y_{i1} are unaffected by future values of D_i or M_{i2}. Given the DID setup, we have $Y_{i1} = Y_{i1}(0, M_{i1})$. A key feature of differences-in-differences designs is analyzing changes in the outcome over time to adjust for time-constant confounding. To that end, let $\Delta Y_i(d, m) = Y_{i2}(d, m) - Y_{i1}(0, M_{i1})$ be the potential outcome changes, where we connect this to the observed changes over time as $\Delta Y_i = Y_{i2}(D_i, M_{i2}) - Y_{i1}$. We could allow Y_{i2} to depend on the baseline mediator explicitly, but our estimands will always condition on the baseline mediator, so we make that dependence implicit for notational simplicity.

2.1 Quantities of interest

Our goal is to estimate the direct effect of treatment, fixing the value of the posttreatment mediator to a particular value. We introduce a few different estimands to this end. The first is the controlled direct effect conditional on the mediator taking that value at baseline:

$$\tau_m = \mathbb{E}\{Y_{i2}(1, m) - Y_{i2}(0, m) \mid M_{i1} = m\},$$

for some $m \in \mathcal{M}$. We refer to this as the baseline-conditional average controlled direct effect or ACDE-BC. In the context of our application, this is the effect of the perspective-taking intervention for a fixed level of subjective feelings about transgender people for units with that same level of subjective feelings at baseline. Conditioning on the baseline level of the mediator creates a separate DID-like setting for each stratum of M_{i1}, since within each stratum, all observations begin in period 1 with $D_{i1} = 0$ and $M_{i1} = m$, and so $\Delta Y_i(d, m)$ represents changes over time for a fixed value of the mediator. This approach also allows us to ignore any carryover effects of M_{i1} on Y_{i2}.

The ACDE-BC is useful when assessing the effect for a particular value of the mediator, but it is also helpful to have a summary measure of the direct effect at differing levels of the mediator. Letting $\rho_m = \mathbb{P}(M_{i1} = m)$, we can marginalize over the distribution of the baseline mediator with

$$\tau = \sum_{m \in \mathcal{M}} \tau_m \rho_m = \sum_{m \in \mathcal{M}} \mathbb{E}[Y_{i2}(1, m) - Y_{i2}(0, m) \mid M_{i1} = m]\rho_m,$$

which we call the marginalized ACDE-BC. This estimand treats each level of the baseline mediator as a separate DID study and aggregates them based on their size. In this way, it is similar to a conditional version of the average factorial effect in a factorial experiment or the average marginalized component effect in conjoint studies.

We also investigate the controlled direct effect on those who were treated *and* hold their value of the mediator constant over time,

$$\gamma_m = \mathbb{E}\{Y_{i2}(1, m) - Y_{i2}(0, m) \mid D_i = 1, M_{i1} = m, M_{i2} = m\},$$

which is similar to the average treatment effect on the treated in single-treatment settings. We call this the path-conditional average controlled direct effect or ACDE-PC, and we can marginalize it similarly to τ_m and τ. In our application, this estimand has the same interpretation as the ACDE-BC, except that the effect is only among those units who would (and do) remain at their baseline subjective feelings about transgender people before and after treatment. Below, we identify this quantity under an alternative set of assumptions that may be more plausible in some empirical settings. We note that, like causal mediation quantities, the ACDE-PC is a cross-world quantity because we can write it as

$$\gamma_m = \mathbb{E}\{Y_{i2}(1,m) \mid D_i = 1, M_{i1} = m, M_{i2}(1) = m\} - \mathbb{E}\{Y_{i2}(0,m) \mid D_i = 1, M_{i1} = m, M_{i2}(1) = m\},$$

where the second term is an average of the potential outcomes under control $Y_{i2}(0,m)$ with a conditioning statement that depends on a treated potential value of the mediator $M_{i2}(1)$. As we will see below, this cross-world property makes identifying this quantity incompatible with intermediate confounders.

2.2 Assumptions

We build our identification from two key features of the experimental design: randomization and panel data. Randomization allows us to identify the overall effect of treatment, and the panel nature of the data allows us to leverage the key identifying assumption of a difference-in-differences design that there are parallel trends in specific potential outcomes over time. Let $\mathbf{Y}(\bullet) = \{Y_{it}(d,m) : t = 1,2 \; d = 0,1 \; m \in \mathcal{M}\}$ be the set of all potential outcomes, with similar notation defined for $M_{i2}(d)$ and $\mathbf{Z}_i(d)$.

Assumption 1 (Treatment Randomization)**.** $\{\mathbf{Y}(\bullet), M_{i2}(\bullet), \mathbf{Z}_i(\bullet), M_{i1}\} \perp\!\!\!\perp D_i$.

Assumption 2 (Mediator Parallel Trends)**.** *For* $d \in \{0,1\}$, *and* $m, m', m'' \in \mathcal{M}$

$$\begin{aligned}\mathbb{E}\{Y_{i2}(0,m) - Y_{i1}(0,m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, \mathbf{Z}_i, M_{i2} = m'\} \\ = \mathbb{E}\{Y_{i2}(0,m) - Y_{i1}(0,m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, \mathbf{Z}_i, M_{i2} = m''\}.\end{aligned}$$

Assumption 1 comes from the design of the experiment, though it is possible to generalize this assumption to a selection-on-observables or parallel trends assumption for an observational study. Assumption 2 states that the over-time trends in the potential outcomes under control and mediator level m are mean-independent of the mediator value in period 2, conditional on some covariates that might be pretreatment ($\mathbf{X}_i$) or posttreatment ($\mathbf{Z}_i$). This assumption would identify the effect of M_{i2} in a DID study where we viewed M_{i2} as the treatment of interest, with D_i becoming a baseline covariate. For example, suppose a unit switches their subjective feelings about transgender people from neutral to positive (say, $M_{i1} = m$ to $M_{i2} = m'$) before and after treatment. The mediator parallel trends assumption states that those who switch would have had the same mean changes in the potential outcomes under control and mediator level m over time as those who do not switch their subjective feelings, conditional on $\mathbf{X}_i$ and $\mathbf{Z}_i$. Note that this places no restrictions on the baseline mediator, so we allow for unmeasured confounding between the outcome and the baseline mediator. Thus, we allow pretreatment subjective feelings to be arbitrarily related to baseline attitudes about laws relating to transgender people. Unfortunately, this assumption is not enough to identify the ACDE-BC.

Assumption 3 (No Direct Effect Moderation by the Mediator)**.** *For $d \in \{0, 1\}$, and $m, m', m'' \in \mathcal{M}$.*

$$\mathbb{E}\{Y_{i2}(1, m) - Y_{i2}(0, m) \mid D_i = 1, \mathbf{X}_i, M_{i1} = m, \mathbf{Z}_i, M_{i2} = m'\}$$
$$= \mathbb{E}\{Y_{i2}(1, m) - Y_{i2}(0, m) \mid D_i = 1, \mathbf{X}_i, M_{i1} = m, \mathbf{Z}_i, M_{i2} = m''\}.$$

Assumption 3 imposes a homogeneity assumption on the treated group such that strata defined by the posttreatment mediator have the same ACDE-BC, conditional on all covariates. Intuitively, this assumption means that treatment cannot induce variation in the mediator that is correlated with the controlled direct effect. In our context, this assumption could be violated if, for example, feeling cold vs. warm toward transgender people in the treatment arm is related to having a stronger or weaker controlled direct effect of treatment on support for anti-discrimination policies. One way to view these violations is as time-varying confounding between M_{i2} and Y_{i2} induced by treatment. As is typical for DID designs, our approach can only handle time-constant unmeasured confounding. We do note that this mean-independence of the treatment effect is still significantly weaker than other no-interactions assumptions used to identify mediation effects that require no interaction between D_i and M_{i2} at the individual level (Robins, 2003). Shahn et al. (2022) avoid this assumption by using an intermediate outcome between D_i and M_{i2} that would allow them to make parallel trends assumptions with respect to the mediator in the treated group; that is, $Y_{i2}(1, m) - Y_{i1}(1, m)$. Given that M_{i2} has a less clear assignment time, we believe intermediate outcomes are less helpful in this context.

Assumptions 2 and 3 together imply that a "parallel trends" assumption holds for changes like $\Delta Y_i(d,m) = Y_{i2}(d,m) - Y_{i1}(0, M_{i1})$, which is a sufficient condition for our identification assumption below. (Recall that the random M_{i1} in the baseline potential outcome $Y_{i1}(0, M_{i1})$ will not impact our analysis since we will always condition on the baseline mediator.) While there may be special cases where this weaker condition holds and our assumptions do not, separating these into more primitive assumptions helps clarify them. Furthermore, these assumptions are implied by (but do not imply) the following sequential ignorability assumption with changes in the outcome as the dependent variable,

$$\Delta Y_i(d,m) \perp\!\!\!\perp M_{i2} \mid M_{i1} = m, D_i = d, \mathbf{X}_i, \mathbf{Z}_i. \tag{1}$$

Our assumptions are weaker since (a) they only restrict the averages of the potential outcomes rather than their entire distributions, and (b) they only restrict the potential outcomes for the same mediator status as the baseline mediator, m. This sequential ignorability version of the assumptions does retain the core benefit of a differences-in-differences design: both D_i and M_i can still be correlated with time-constant factors that affect both Y_{i1} and Y_{i2} in the same way. That is, they still allow for time-constant unmeasured confounding, albeit in a restricted, linear fashion.

As an example of how this unmeasured confounding might manifest, suppose a time-constant unmeasured confounder, U_i, is correlated with M_{i2}. Further, suppose we have the following models for our potential outcomes:

$$Y_{i1}(d,m) = f_{1dm}(\mathbf{X}_i) + g(U_i) + \varepsilon_{i1}, \qquad Y_{i2}(d,m) \quad = f_{2dm}(\mathbf{X}_i, \mathbf{Z}_i(d)) + g(U_i) + \varepsilon_{i2},$$

where we assume that ε_{it} are i.i.d. and independent of all variables $\mathbf{O}_i$. Here, the functions f_{1dm} and f_{2dm} capture the observed time-varying confounding and treatment effect heterogeneity, whereas g reflects the time-constant unmeasured confounding. Under this model, the usual sequential ignorability assumption for $Y_{i2}(d,m)$ and M_{i2} conditional on just $\mathbf{X}_i$ and $\mathbf{Z}_i$ would not hold because of the unmeasured confounder, U_i. However, because that confounder enters into the model for potential outcomes in a linear, additive, and time-constant manner, it will be unrelated to the *changes* in the potential outcomes over time.

The homogeneous effects restriction of Assumption 3 may be too strong for certain empirical applications. Furthermore, it requires a cross-sectional restriction at odds with the spirit of difference-in-differences. In settings where the posttreatment mediator might be correlated with the controlled direct effect, we propose an alternative identifying assumption based on parallel trends among controls only. This assumption will identify the ACDE-PC, γ_m.

Assumption 4 (Mediator Parallel Trends with No Intermediate Confounders)**.** *For all $m, m', m'' \in \mathcal{M}$ and $d \in \{0, 1\}$,*

$$\mathbb{E}\{\Delta Y_i(0, m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, M_{i2} = m'\} = \mathbb{E}\{\Delta Y_i(0, m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, M_{i2} = m''\}.$$

This assumption states that parallel trends holds for the mediator in the control group conditioning just on the pretreatment covariates and the baseline value of the mediator. In other words, conditional on pretreatment covariates, the observed value of the second-period mediator is unrelated to the trends in the potential outcomes in the control condition—or in the context of our application, that observed changes in subjective feelings are unrelated to counterfactual changes in support for laws under control. Combined with randomization of D_i, this implies that every group defined by their values of D_i and M_{i2} would have followed the same average trend if, possibly contrary to fact, they had remained at $M_{i2} = m$ and stayed in the control condition. Given the lack of intermediate covariates, this is similar to a standard difference-in-differences design with a multileveled treatment (combining D_i and M_{i2}).

The exclusion of posttreatment covariates in this identifying assumption is a significant limitation, so it is essential to consider why we must exclude them. To identify the ACDE, we will need to impute the trends for $(D_i = 0, M_{i2} = m)$ group among those with, say, $(D_i = 1, M_{i2} = m)$. We typically accomplish this by adjusting for covariates through weighting or regression, and those methods would require assumptions on both the treated and control potential outcome trends as in Assumptions 2 and 3. If we include posttreatment confounders, however, our adjustment would require information about the joint distribution of the potential outcomes of the posttreatment covariates, $\mathbf{Z}_i(1)$ and the potential outcomes $\Delta Y_i(0, m)$. Unfortunately, we cannot identify this joint distribution (without strong additional assumptions) due to the fundamental problem of causal inference. We could assume $\mathbf{Z}_i$ is unaffected by D_i, but then it ceases to be posttreatment. We could alternatively assume parallel trends holds for $\mathbf{Z}_i(1, m)$ with respect to $\Delta Y_i(0, m)$, conditional on $\mathbf{X}_i$, $M_{i1} = m$ and $D_i = 1$, but this seems to call into question why it would be needed to block confounding for M_{i2}. Thus, in this setting, we can either restrict our parallel trends assumption to the control treatment of D_i or allow for posttreatment confounders, but not both simultaneously.

2.3 Identification

We now describe the various functions of the observed data we will use in identification. First, let $\pi_{dm_2}(m_1, \mathbf{x}, \mathbf{z}) = \mathbb{P}(M_{i2} = m_2 \mid M_{i1} = m_1, D_i = d, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z})$ be the generalized propensity score for M_{i2}. We define W_{i1m} to be an indicator for the baseline mediator being equal to m so that $W_{i1m} = 1$ when $M_{i1} = m$ and 0 otherwise, with W_{i2m} defined similarly for M_{i2}. We use the convention that when $\pi_{dm}(\cdot)$ omits $\mathbf{Z}_i$, the function represents the propensity score just as a function of $\mathbf{X}_i$. Next, we define the regressions of the differenced outcome on the treatment, mediator, and covariates as

$$\mu_{dm_2}(m_1, \mathbf{x}, \mathbf{z}) = \mathbb{E}[\Delta Y_i \mid M_{i2} = m_2, M_{i1} = m_1, D_i = d, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}].$$

Since our focus is on estimands where the mediator is the same before and after treatment, we omit the baseline value from the function when it matches the posttreatment mediator: $\pi_{dm}(\mathbf{X}_i, \mathbf{Z}_i) = \pi_{dm}(m, \mathbf{X}_i, \mathbf{Z}_i)$ and $\mu_{dm}(\mathbf{X}_i, \mathbf{Z}_i) = \mu_{dm}(m, \mathbf{X}_i, \mathbf{Z}_i)$. Finally, let $\delta = \mathbb{P}(D_i = 1)$ and $\lambda_{dm} = \mathbb{P}(M_{i1} = m, D_i = d, M_{i2} = m)$ be the marginal probabilities of treatment and a particular path, respectively. We make the following standard positivity assumption:

Assumption 5 (Positivity)**.** *For all $(\mathbf{x}, \mathbf{z})$ in the support of $(\mathbf{X}_i, \mathbf{Z}_i)$, $m \in \mathcal{M}$, and $d \in \{0, 1\}$, we have $\pi_{dm}(\mathbf{x}, \mathbf{z}) > \varepsilon > 0$.*

We now state the main two identification results for each estimand. One relies on inverse propensity score weighting, and the other relies on outcome regressions. All proofs are in Supplemental Materials B.

Proposition 1. *Under Assumptions 1, 2, 3, and 5, we have*

$$\tau_m = \mathbb{E}\left[\left(\frac{D_i W_{i2m}}{\delta \pi_{1m}(\mathbf{X}_i, \mathbf{Z}_i)} - \frac{(1-D_i)W_{i2m}}{(1-\delta)\pi_{0m}(\mathbf{X}_i, \mathbf{Z}_i)}\right)\Delta Y_i \mid M_{i1} = m\right], \tag{2}$$

and,

$$\tau_m = \mathbb{E}\left[\frac{W_{i1m} D_i}{\rho_m \delta}\mu_{1m}(\mathbf{X}_i, \mathbf{Z}_i) - \frac{W_{i1m}(1-D_i)}{\rho_m(1-\delta)}\mu_{0m}(\mathbf{X}_i, \mathbf{Z}_i)\right]. \tag{3}$$

Under Assumptions 1, 4, and 5, we have

$$\gamma_m = \mathbb{E}\left[\frac{W_{i1m} W_{i2m}}{\lambda_{1m}}\left(D_i - (1-D_i)\frac{\delta \pi_{1m}(\mathbf{X}_i)}{(1-\delta)\pi_{0m}(\mathbf{X}_i)}\right)\Delta Y_i\right], \tag{4}$$

and

$$\gamma_m = \mathbb{E}\left[\frac{W_{i1m} D_i W_{i2m}}{\lambda_{1m}}\left(\mu_{1m}(\mathbf{X}_i) - \mu_{0m}(\mathbf{X}_i)\right)\right]. \tag{5}$$

These results resemble standard identification results based on weighting and outcome regressions with changes in the outcomes before and after treatment, ΔY_i, replacing levels of the outcome, Y_{i2}.

2.3.1 Connections to other identification results

While we have focused so far on the direct effects of treatment, mediation analyses often target indirect effects as well. The presence of posttreatment confounders, $\mathbf{Z}_i$, usually precludes the possibility of identifying mediation quantities like the natural indirect effect (Robins, 2003; Avin, Shpitser and Pearl, 2005) and this is true for our controlled direct effect parameter τ_m. For example, Deuchert, Huber and Schelker (2019) use a principal strata approach to provide sufficient conditions for identifying the natural direct and indirect effects for subgroups under parallel trends assumptions. In particular, their Theorem 1 on the direct effect for those whose mediators are unaffected by treatment is similar to our identification of γ_m. Our results go further in showing how to estimate the direct effects in the presence of intermediate confounding, in addition to allowing for non-binary mediators.

We can also compare the identification results to those based on a standard sequential ignorability assumption for outcome levels rather than outcome changes. This design would maintain that

$Y_{i2}(d,m) \perp\!\!\!\perp M_{i2} \mid D_i = d, \mathbf{X}_i, \mathbf{Z}_i$ and we would identify τ_m using an IPW approach as $\mathbb{E}[\omega_{im} Y_{i2}]$, where

$$\omega_{im} = \frac{W_{i1m} D_i W_{i2m}}{\rho_m \delta \pi_{1m}(\mathbf{X}_i, \mathbf{Z}_i)} - \frac{W_{i1m}(1 - D_i) W_{i2m}}{\rho_m (1-\delta) \pi_{0m}(\mathbf{X}_i, \mathbf{Z}_i)},$$

meaning that the difference between our IPW DID identification result and the sequential ignorability identification result is $\mathbb{E}[\omega_{im} Y_{i1}]$. This estimand represents the identified "controlled direct effect" of treatment on a pretreatment measurement of the outcome, which should be zero under sequential ignorability. Thus, one way to view the DID approach we present is leveraging the known null effect of treatment on the past to correct biases in standard sequential ignorability approaches—a technique referred to in the statistics literature as negative control (Lipsitch, Tchetgen Tchetgen and Cohen, 2010; Sofer et al., 2016).

This analysis assumes we use the same conditioning set under a parallel trends approach and a sequential ignorability approach, but what if we condition on the lagged dependent variable (LDV) in the latter? Several authors have shown a bracketing relationship between the DID approach and this LDV approach in the case of a single treatment variable (Angrist and Pischke, 2009; Ding and Li, 2019). In Supplemental Materials A, we derive the difference between the DID target of inference and LDV target of inference for the ACDE-BC and the ACDE-PC. In the latter case, we show that when either parallel trends or sequential ignorability with an LDV holds, the two approaches should bound the true ACDE-PC in the limit, as in Ding and Li (2019).

3 Estimation

We now turn to the estimation of the controlled direct effects. Given the identification results, we could construct plug-in estimators based on the IPW or outcome regression approaches where we model either π_{dm} or μ_{dm} and plug in our estimates into a sample version of the expectations in Section 2.3. However, both IPW and outcome regression approaches can be biased, unstable, or both when these models are incorrectly specified. We develop a set of doubly robust estimators to create efficient and stable estimators. These estimators are doubly robust in that we will specify two models—for the propensity scores and the outcome regression—and the resulting estimator will be consistent and asymptotically normal when one, but not necessarily both, of the models are correctly specified. Finally, we integrate a cross-fitting procedure into our estimation approach so that we can use data-adaptive machine learning models to make estimates less sensitive to particular functional form assumptions.

To derive a doubly robust and semiparametrically efficient estimator, we first derive the efficient influence functions (EIFs) for our parameters based on the maintained assumptions. We will build these EIFs from estimating equations $\psi_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau)$ for τ_m and $\phi_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma)$ for γ_m, where $\boldsymbol{\eta}_\tau = (\pi_{dm}, \mu_{dm}, \rho_m, \delta)$ is the vector of nuisance parameters for τ_m, $\boldsymbol{\eta}_\gamma = (\pi_{dm}, \mu_{dm}, \lambda_{dm}, \delta)$ is the vector of nuisance parameters for γ_m. Letting $\mathbf{V}_i = (\mathbf{X}_i, \mathbf{Z}_i)$, the estimating equations for the ACDE-BC have the form $\psi_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau) = \psi_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) - \psi_{0m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau)$, where

$$\psi_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) = \frac{W_{i1m}}{\rho_m}\left[\frac{D_i W_{i2m}}{\delta \pi_{1m}(\mathbf{V}_i)}(\Delta Y_i - \mu_{1m}(\mathbf{V}_i)) + \frac{D_i}{\delta}(\mu_{1m}(\mathbf{V}_i))\right]. \tag{6}$$

$$\psi_{0m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) = \frac{W_{i1m}}{\rho_m}\left[\frac{(1 - D_i) W_{i2m}}{(1-\delta) \pi_{0m}(\mathbf{V}_i)}(\Delta Y_i - \mu_{0m}(\mathbf{V}_i)) + \frac{1 - D_i}{1 - \delta}(\mu_{0m}(\mathbf{V}_i))\right]. \tag{7}$$

For the ACDE-PC, the estimating equations take the form $\phi_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) = \phi_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) - \phi_{0m}(\mathbf{O}_i; \boldsymbol{\eta}_\gamma)$, where

$$\phi_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) = \frac{W_{i1m} D_i W_{i2m}}{\lambda_{1m}}(\Delta Y_i - \mu_{0m}(\mathbf{X}_i)) \tag{8}$$

$$\phi_{0m}(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) = \left(\frac{W_{i1m}(1 - D_i) W_{i2m}}{\lambda_{1m}}\right)\left(\frac{\pi_{1m}(\mathbf{X}_i)\delta}{\pi_{0m}(\mathbf{X}_i)(1-\delta)}\right)(\Delta Y_i - \mu_{0m}(\mathbf{X}_i)) \tag{9}$$

These estimating equations combine imputation (via the outcome regressions) with weighting (via the propensity scores). They are generalizations of similar doubly robust approaches to estimating ACDEs under sequential ignorability (Murphy et al., 2001; Orellana, Rotnitzky and Robins, 2010; van der Laan and Gruber, 2012) and are similar to doubly robust approaches to the estimation of the effect of point exposures under difference-in-differences designs (Sant'Anna and Zhao, 2020). In Supplemental Materials B.3, we show that, under suitable regularity conditions, these functions are the efficient influence function after recentering.

We now present two estimators based on these EIFs. Doubly robust estimators like ours contain two steps: estimating the "nuisance" functions (propensity scores and outcome regressions) and then plugging these estimates into sample versions of the identification formulas to estimate the quantity of interest. The first step of the doubly robust estimator is to estimate these nuisance functions with what are often called "working models," a name that emphasizes that we do not necessarily assume they are correctly specified. For the propensity score estimates, which we refer to as $\widehat{\pi}_{dm_2}(m_1, \mathbf{x}, \mathbf{z})$, common approaches would be to use a logistic regression for a binary mediator or a multinomial logistic regression for more general discrete mediators. However, our setup allows for more flexible machine learning models. We assume another working model for the outcome regression, $\widehat{\mu}_{dm_2}(m_1, \mathbf{x}, \mathbf{z})$, which might be a simple ordinary least squares regression or something more complicated like the Lasso or a random forest.

Our final estimator for the ACDEs will be

$$\widehat{\tau}_m = \mathbb{P}_n\{\psi_m(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_\tau)\} \qquad \widehat{\gamma}_m = \mathbb{P}_n\{\phi_m(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_\gamma)\},$$

where $\widehat{\boldsymbol{\eta}}_\tau$ and $\widehat{\boldsymbol{\eta}}_\gamma$ are the estimated nuisance functions and $\mathbb{P}_n(f(\mathbf{O}_i)) = n^{-1}\sum_{i=1}^n f(\mathbf{O}_i)$. We first establish a doubly robust consistency result for these estimators.

Theorem 1. *(a) Under Assumptions 1, 2, 3, 5, and suitable regularity conditions, $\widehat{\tau}_m$, is consistent for τ_m when, for $d \in \{0, 1\}$, either $\widehat{\pi}_{dm} \xrightarrow{p} \pi_{dm}$ or $\widehat{\mu}_{dm} \xrightarrow{p} \mu_{dm}$. When all models are correctly specified, $\widehat{\tau}_m$ is semiparametrically efficient. (b) Under Assumptions 1, 4, and 5, $\widehat{\gamma}_m$ is consistent for γ_m when either $\widehat{\pi}_{dm} \xrightarrow{p} \pi_{dm}$ or $\widehat{\mu}_{dm} \xrightarrow{p} \mu_{dm}$. When all models are correctly specified, $\widehat{\gamma}_m$ is semiparametrically efficient.*

Theorem 1 ensures that our estimators will be consistent for their intended estimands when either the propensity score model for the posttreatment mediator or the outcome regressions are correctly specified. While we focus on the case with a randomized treatment, this result extends to handling a treatment that satisfies selection on observables or a parallel trends assumption. In that case, we would require an additional propensity score and outcome regression model that adjust the confounding between treatment and the outcome due to $\mathbf{X}_i$. In that case, one could employ the multiply robust estimators of Zhou (2022) and Zhou and Yamamoto (2023) with the changes over time as the dependent variable. When using parametric regression estimators for those approaches approach, covariate selection should be consistent across outcome regression specifications to ensure model compatibility.

3.1 Variance estimation and crossfitting

As we have shown, we first need to fit a series of working models to obtain estimators from the doubly robust estimator $\widehat{\tau}_m$. When we "double dip" and use the same observations to fit the outcome regressions and propensity scores as we use in the sample mean $n^{-1}\sum_{i=1}^{n}\psi_m(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}_\tau)$, our estimates can become less stable, and we must account for using the data twice in our uncertainty estimates.

We can avoid both of these issues using cross-fitting/sample-splitting (Chernozhukov et al., 2018), a simple way to make nuisance parameter estimates independent of the final estimates of the quantities of interest. We first randomly split the sample into K equal-sized groups and use $K-1$ of the groups to fit the working models, then use those fitted models to obtain predicted values for the remaining group to form an estimate of $\widehat{\tau}_m$. To recover the lost efficiency of estimating with $1/K$ of the original sample, we use cross-fitting to repeat this procedure for each of the K groups and average the resulting estimates. Finally, to account for the variability of this splitting process, we repeat this process several times and take the median of the estimates across the different splits as recommended by Chernozhukov et al. (2018).

To be specific, we randomly partition the data into K groups by drawing $(B_1,\ldots,B_n)$ independently of the data, where B_i is distributed uniformly over $\{1,\ldots,K\}$. We take $B_i=b$ to mean that unit i is split into group b. Let $\psi_m(\mathbf{O}_i,\widehat{\boldsymbol{\eta}}_{\tau,-b})$ be the value of the influence function when the nuisance parameters are estimated without the group $B_i=b$. We also let $\mathbb{P}_n^b$ denote the conditional empirical distribution for the group $B_i=b$,

$$\mathbb{P}_n^b\{f(\mathbf{O}_i)\} = \frac{\sum_{i=1} f(\mathbf{O}_i)\mathbb{I}(B_i=b)}{\sum_{i=1}\mathbb{I}(B_i=b)}$$

Then, we can define the cross-fitting estimator as

$$\widehat{\tau}_m = \sum_{b=1}^{K}\left\{\frac{1}{n}\mathbb{I}(B_i=b)\right\}\mathbb{P}_n^b\{\psi_m(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}_{\tau,-b})\} = \mathbb{P}_n\left\{\psi_m(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}_{\tau,-B_i})\right\},$$

with $\widehat{\gamma}_m$ defined similarly. Let $\|f\| = \left(\mathbb{E}[(f(\mathbf{O}_i))^2]\right)^{1/2}$ for any function f. Let $\widetilde{\psi}_m(\mathbf{O}_i;\boldsymbol{\eta}_\tau)$ and $\widetilde{\phi}_m(\mathbf{O}_i;\boldsymbol{\eta}_\gamma)$ be the centered EIFs for τ_m and γ_m, respectively. We describe these EIFs in detail in the Supplemental Materials.

Theorem 2. *(a) Let Assumptions 1, 2, 3, and 5 hold and suppose that (i) $\left\|\widehat{\boldsymbol{\eta}}_\tau - \boldsymbol{\eta}_\tau\right\| = o_p(1)$, (ii) $\|\widehat{\mu}_{dm} - \mu_{dm}\| \times \|\widehat{\pi}_{dm} - \pi_{dm}\| = o_p(n^{-1/2})$. Then, $\sqrt{n}(\widehat{\tau}_m - \tau_m)$ will converges in distribution to $\mathcal{N}(0, \mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i;\boldsymbol{\eta}_\tau)^2])$ and is thus semiparametrically efficient. (b) Under the same assumptions with Assumption 4 replacing Assumptions 2 and 3, $\sqrt{n}(\widehat{\gamma}_m - \gamma_m)$ will converge in distribution to $\mathcal{N}(0, \mathbb{E}[\widetilde{\phi}_m(\mathbf{O}_i;\boldsymbol{\eta}_\gamma)^2])$ and is thus semiparametrically efficient.*

An additional benefit of this cross-fitting procedure is that it allows for "plug-and-play" integration with machine learning algorithms so that we can replace, for example, a standard logistic regression model for a propensity score with a data-adaptive algorithm such as the logistic Lasso. Bradic, Ji and Zhang (2021) provide the rate conditions on these types of algorithms needed to ensure consistent and asymptotic normality of dynamic treatment effects like the ones we study here. Other work has shown that many algorithms will achieve the necessary rates, including regression trees, random forests, neural nets, and boosting in sparse linear models (see Chernozhukov et al., 2018, and citations therein). In our empirical application, we leverage a version of the Lasso for outcome regressions and a random forest approach to estimate the generalized propensity score for a three-level discrete mediator.

3.2 Sensitivity analysis

The parallel trends assumptions might be quite strong in applications like ours, even after controlling for covariates. It is crucial, then, to understand how robust our findings are to violations of these key assumptions. Following Robins, Rotnitzky and Scharfstein (1999) and Blackwell (2014), we can conduct a sensitivity analysis that allows for restricted violations of the parallel trends assumption. In particular, suppose we assume that the average trends in the potential outcomes do, in fact, vary by the value of the post-treatment mediator by some amount,

$$\left|\mathbb{E}\left[\Delta Y_i(0,m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m\right] - \mathbb{E}\left[\Delta Y_i(0,m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, M_{i2} = m\right]\right| \leq \Gamma. \quad (10)$$

This restriction controls the violation of parallel trends by allowing the group of respondents who stay at $M_{i1} = M_{i2} = m$ to have average potential outcome trends under control different from the overall population by Γ, conditional on treatment status and the baseline covariates.

Under Assumptions 1 and 5 and restriction (10), it is straightforward to show that the ACDE-PC can be bounded by $\gamma_m \in [L_{\gamma_m}, U_{\gamma_m}]$, where

$$L_{\gamma_m} = \mathbb{E}\left[\frac{W_{i1m}W_{i2m}}{\lambda_{1m}}\left(D_i - (1-D_i)\frac{\delta\pi_{1m}(\mathbf{X}_i)}{(1-\delta)\pi_{0m}(\mathbf{X}_i)}\right)\Delta Y_i\right] - 2\Gamma,$$
$$U_{\gamma_m} = \mathbb{E}\left[\frac{W_{i1m}W_{i2m}}{\lambda_{1m}}\left(D_i - (1-D_i)\frac{\delta\pi_{1m}(\mathbf{X}_i)}{(1-\delta)\pi_{0m}(\mathbf{X}_i)}\right)\Delta Y_i\right] + 2\Gamma.$$

We focus here on the ACDE-PC, but we could easily apply a similar procedure to the ACDE-BC as well. This result is presented in terms of the inverse probability weighting approach to identification, but a similar result holds for the regression imputation identification as well. Thus, a valid sensitivity analysis procedure will be to calculate bounds $\widehat{L}_{\gamma_m} = \widehat{\gamma}_m - 2\Gamma$ and $\widehat{U}_{\gamma_m} = \widehat{\gamma}_m + 2\Gamma$. To obtain a confidence interval for γ_m under these assumptions, we use the procedure of Imbens and Manski (2004), which accounts for how the true parameter cannot be close to both the upper and lower bound at the same time.

A key part of any sensitivity analysis is benchmarking how substantively large the deviations from parallel trends are in (10). We take the strategy of observing how much unmeasured confounding would result from omitting observed covariates. In particular, we pretend as though our analysis only has access to $\mathbf{X}_{i,-k}$, which is $\mathbf{X}_i$ omitting the kth. Under the assumption that parallel trends holds for $\mathbf{X}_i$, we can then estimate the bound Γ that would arise from omitting X_{ik} from

$$\begin{aligned}
&\mathbb{E}[\Delta Y_i(0,m) \mid D_i = 0, \mathbf{X}_{i,-k}, M_{i1} = m, M_{i2} = m] - \mathbb{E}[\Delta Y_i(0,m) \mid D_i = 0, \mathbf{X}_{i,-k}, M_{i1} = m] \\
&= \mathbb{E}[\Delta Y_i \mid D_i = 0, \mathbf{X}_{i,-k}, M_{i1} = m, M_{i2} = m] \\
&\quad - \mathbb{E}\left[\mathbb{E}[\Delta Y_i \mid D_i = 0, \mathbf{X}_i, M_{i1} = m, M_{i2} = m] \mid D_i = 0, \mathbf{X}_{i,-k}, M_{i1} = m\right].
\end{aligned}$$

To approximate this bound while taking into account sampling uncertainty, we estimate these regressions and find the 95th percentile of the absolute value of the differences between them as an estimate of how much of a violation of parallel trends would be caused by omitting X_{ik}. This will help us reason about how large the value of Γ is that could overturn a finding.

3.3 Simulation evidence

In Supplemental Materials C, we present results from a simulation experiment that show the finite-sample performance of our estimator compared with traditional DID estimators. We show that when our models are correctly specified, the doubly robust estimators all outperform those DID estimators in terms of bias and estimation error. These results also hold when the models are incorrectly specified, at least for large samples. The coverage rates of confidence intervals based on cross-fitting are very close to nominal levels under correctly specified models, but we do see undercoverage when the models are incorrect.

4 Results

We now apply these methods to estimate whether a pro-transgender intervention changes support for nondiscrimination laws, holding constant feelings of warmth towards transgender people. The main finding in Broockman and Kalla (2016) is that the canvassing intervention increased support for nondiscrimination laws in the third and fourth post-intervention periods. (The authors speculate that the absence of treatment effects in the first two periods could be due to respondents' lack of knowledge about the meaning of the term 'transgender' and so included a definition in the subsequent waves.) While Broockman and Kalla (2016) report treatment effects based on cross-sectional differences between treatment and control groups after the intervention, we instead use changes in the outcome ΔY_i.

For the ACDE-BC and ACDE-PC estimators, we use adaptive estimation for the nuisance functions with the lasso approach of Belloni and Chernozhukov (2013) from the `hdm` R package for the outcome regression and random forests from the `ranger` R package for the propensity scores for M_{i2}. For the lasso estimator, we pass all the covariates plus first-order interactions and squared terms for continuous variables (though we do not include these flexible terms for the standard DID estimates). The lasso in particular requires this kind of basis expansion to capture nonlinear functional forms. As dictated by our identifying assumptions, we omit intermediate covariates for estimating the ACDE-PCs. We restrict our sample to individuals for which all covariates are observed ($N = 369$). We use $K = 5$ folds for cross-fitting and repeat the cross-fitting procedure 20 times and combined the estimates using the median approach of Chernozhukov et al. (2018).

4.1 Main estimates

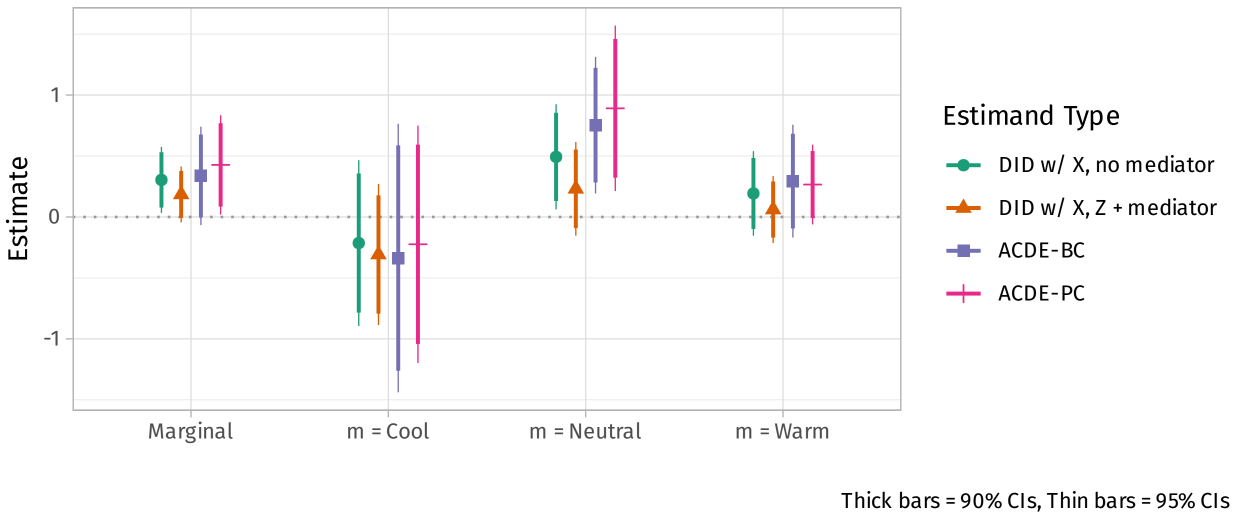

Figure 3: Controlled direct effect estimates using different estimation strategies.

We present the results in Figure 3. We group the estimates by the subset of the baseline mediator used for estimation, including the estimates marginalized over that variable. We present two standard regression DID estimates. The first uses the basic regression DID approach to estimate the overall ATT with no conditioning on intermediate covariates or the mediator (green circles). The second (orange triangles) shows how those regression DID estimates change if we include $\mathbf{Z}_i$ and M_{i2} in the regression. Finally, we show the doubly robust ACDE (DR-ACDE) estimates of $\widehat{\tau}_m$ (purple squares) and $\widehat{\gamma}_m$ (pink lines). The DID results just conditioning on baseline covariates replicate the main results of the original study: the perspective-taking intervention increased support for nondiscrimination laws. The magnitude of the DID estimate (0.304) is very similar to the cross-sectional estimate of the effects from the original study (0.36). However, once we add intermediate covariates and the mediator into our DID analysis, the effect attenuates by almost 40% (0.188). Such a change might lead an analyst to conclude that feelings about transgender people mediate the effect of the intervention. Of course, this ignores the potential for posttreatment bias in conditioning on the intermediate confounders.

Our approach to estimating controlled direct effects show a different and more nuanced set of results. For both of the marginal ACDEs, we see that the direct effect estimates are similar in magnitude to the overall ATT. These effects also have much larger standard errors due in part to the estimation of the nuisance functions. The uncertainty in these results makes it difficult to compare to the overall DID estimates.

Figure 3 also shows significant treatment effect heterogeneity by baseline feelings about transgender people. In particular, the conditional ATT for those feeling neutral toward transgender people is much larger than for the other two groups. This group also shows a stark difference between the "controlling for $\mathbf{Z}_i$ and M_{i2}" approach and the DR ACDE estimation strategy. The estimated ACDEs are statistically significant and more than 50% larger in magnitude than the conditional ATT in this case, whereas the DID + mediator approach is statistically insignficant and 50% smaller. Thus, these two different approaches would lead to significantly different conclusions about the role that subjective feelings play in the effect of the intervention. For this neutral group, at least, we can say that there continues to be a direct effect of the intervention for fixed subjective views, and there is no strong evidence that subjective views are a large part of the mechanism for this effect. The extent of statistical uncertainty makes it difficult to make any general conclusions for the other baseline levels of the mediator since all confidence intervals cross 0.

These results are substantively important for the study of political behavior since they show that perspective-taking conversations can have political effects even when subjective feelings are unchanged. This result points to the ability of political campaigns to persuade citizens about legal discrimination without necessarily altering their personal feelings about the group. Showing a disconnection between policy views and emotional orientations is vital since several studies in political science have shown that people form subjective feelings toward outgroups in childhood and rarely change them durably (Sears and Funk, 1999). These results are a positive sign for the health of democracies and how they can increase tolerant public policies without necessarily increasing interpersonal tolerance. However, we acknowledge that our effects are substantial among those with neutral feelings toward transgender people at baseline, which may point to a significant limitation of these effects.

4.2 Additional results

Sensitivity analysis We now present the results of the sensitivity analysis described in Section 3.2. Figure 4 shows bounds on the estimate ACDE-PC as we allow for increasing violations of mediator parallel trends. Essentially, this is allowing the group that stays neutral before and after treatment to have a different trend in $\Delta Y_i(0, m)$ conditional on the baseline covariates. From this, we can see that the group of respondents who stay neutral can have a trend up to 0.22 points different from the overall population and the confidence interval would still exclude zero. That amount of unmeasured confounding would represent roughly 0.165 standard deviations of the observed trend in the "remained neutral" group. As a point of comparison, we can look at what covariates would produce that amount of parallel trends violation if they were omitted (shown as X marks along the zero line). Most of the covariates $\mathbf{X}_i$ would be below this threshold and only a handful of strong confounders (an indicator for a Black respondent, thermometer scores for African American, the social dominance orientation, and prior exposure to trans people) could produce violations this big.

Heterogeneous effects Different subpopulations may have a stronger or weaker ACDE of the intervention than others, so we explored heterogeneous treatment effects based on race and gender identification in Table SM.4. Generally speaking, these effects are quite noisy due to the small subgroup sizes, but most of the variation across groups appears swamped by sampling noise. We do find that for the baseline neutral group, the large ACDE-BC appears driven by white respondents ($\widehat{\tau}_m = 1.34$, $S.E. = 0.55$) versus non-white respondents ($\widehat{\tau}_m = 0.50$, $S.E. = 0.33$). We also observe large differential effects by gender for this baseline group, with women having a higher direct effect ($\widehat{\tau}_m = 0.98$, $S.E. = 0.34$) than non-women ($\widehat{\tau}_m = 0.55$, $S.E. = 0.45$), though this difference is smaller than the difference in the overall effects of the intervention for these two groups.

Different measurement timing choices Given the multiwave nature of the study, we can also investigate how the effect varies as we vary when the outcome and mediator were measured. Given

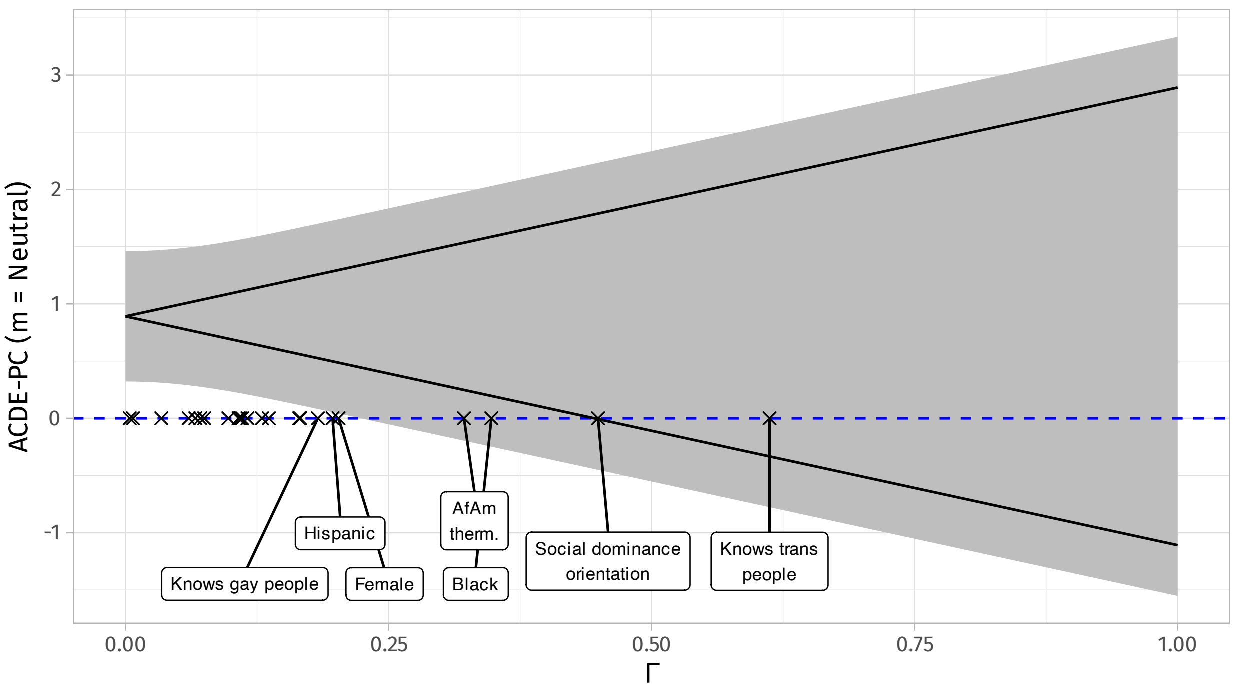

Figure 4: ACDE-PC estimates for m = Neutral group as we weaken the parallel trends assumption. The X marks indicate the estimated severity of parallel trends violation would occur from omitting each of the covariates in $\mathbf{X}_i$, with some of the covariates with larger values labeled.

that we need one posttreatment wave to measure the intermediate covariates, we can use the feeling thermometer mediator at either wave 2 or wave 3, and the outcome can be measured at either wave 3 or 4. In the main results above, we focused on using the mediator at wave 2 and the outcome at wave 3, the latter of which was the first wave that produced a significant ATE in the original study. Table SM.6 shows that the ACDE-BC for the baseline neutral group are all similar in direction and magnitude for the different wave combinations. This result is consistent with a persistent direct effect of the treatment even during the fourth wave of the study, which was three months after treatment. Thus, for some respondents, we can create lasting changes in views on discrimination even without altering their prejudicial views about a group.

Alternative definition of the mediator Our main results used a three-category classification of the feeling thermometer, which allowed us to handle the "clumping" of the feeling thermometer without creating strata too small for inference. We also investigated a five-category classification that grouped respondents into feeling thermometers of (i) exactly 0, (ii) between 1 and 49, (iii) exactly 50, (iv) between 51 and 99, and (v) exactly 100. This alternative measure incorporates the clustering we see at 0 and 100. In Table SM.8, we show these results, which are broadly consistent with our three-category measure results, though with considerably larger uncertainty due to the mediator categories being much smaller. One difference from our main effects is a very large ACDE-BC for the warmest group ($m = 100$), possibly due to ceiling effects on the feeling thermometer score. For respondents with the maximum feeling thermometer at baseline, any positive effect on the outcome must be a direct effect since their subjective feelings cannot be raised any further. However, we must be cautious in interpreting this effect since there are only 35 respondents in this baseline category.

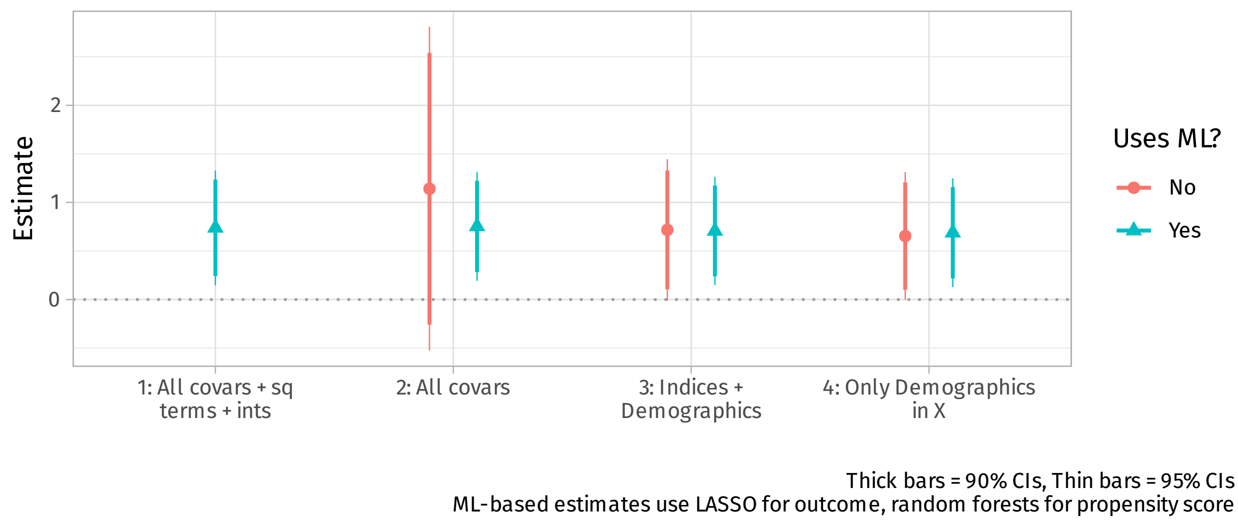

Figure 5: ACDE estimates for m = Neutral across different covariate specifications for the adaptive/ML (blue triangles) and standard (red circles) estimation of the nuisance functions.

Advantages of flexible estimation Finally, we also investigate how using adaptive estimation techniques for the nuisance functions impacts the stability of our estimates across different specification choices. In particular, we varied the choice of variables to pass to either a standard set of models (OLS for the outcome regression and a multinomial logistic regression for the propensity scores) or the adaptive estimators described above. The sets of variables are (a) only demographics in the baseline covariates, (b) demographics and LGBT opinion indices in the baseline covariates, (c) the complete set of baseline covariates, and (d) the complete set of covariates plus squared terms for all continuous variables and all first-order interactions. In this last specification, the number of covariates is far larger than the number of units, so we only used the adaptive design for this specification. Figure 5 presents the results, which show that the adaptive design has a massive impact on the stability of estimates and their uncertainty across these specifications. The increase in uncertainty when adding additional controls is overwhelming for the standard estimators but has almost no impact on the adaptive approach. Thus, the combination of cross-fitting and adaptive nuisance estimation perhaps provides a path toward much less model-dependent estimates and fewer opportunities for intentional or unintentional p-hacking.

5 Conclusion

This paper seeks to estimate the direct effect of a perspective-taking intervention on support for anti-discrimination policies for fixed values of prejudicial feelings about transgender people. To do so, we leverage a multiwave experimental study that allows us to introduce a novel identification strategy for controlled direct effects under a difference-in-difference design. Our key identifying assumptions allow for the mediator to be related to the baseline levels of the potential outcomes, which is far weaker than the selection-on-observables assumption traditionally used to identify the controlled direct effects. Our assumptions require so-called parallel trends assumptions, meaning that the mediator must be unrelated to the *changes* in the potential outcomes over time. This approach highlights how access to baseline measures of the outcome can allow researchers to weaken critical assumptions in evaluating causal mechanisms. We have also built on recent work on doubly robust estimators to propose a doubly robust, semiparametrically efficient estimator for our proposed quantities. These estimators allow researchers to take full advantage of adaptive machine learning algorithms for estimating nuisance functions like propensity scores and outcome regressions. Finally, we also developed a sensitivity analysis procedure for the key parallel trends assumption that allows us to gauge how much unmeasured confounding in the trends it would take to overturn our results.

The empirical results support the presence of a controlled direct effect for respondents who felt neutrally about transgender people at baseline. These estimates may indicate that persuasion about public policy does not necessarily require persuasion about subjective feelings, at least for less polarized individuals. In addition, we find that these effects persisted until the end of the three-month follow-up period and that they are robust to moderate violations of parallel trends. Finally, our results show that proper adjustment for intermediate covariates can lead to different substantive conclusions.

There are several avenues for future substantive and methodological research in this area. Substantively, our results indicate that researchers may benefit from exploring interventions that directly increase support for legal tolerance rather than reducing personal tolerance. Researchers might also benefit from exploring the role of citizens with middling or neutral views about minority groups.

Methodologically, we have focused here on a situation with effectively two time periods and two causal variables (a treatment and a mediator). However, it should be possible to generalize this approach to handle treatment history of arbitrary length. This extension might allow for the identification of and inference of causal effects in marginal structural models with weaker assumptions on confounding between the outcome and the treatment history. In addition, in situations with more pretreatment measurements, it may be possible to use those past measurements to measure and correct for deviations from the parallel trends assumptions. Finally, one could develop a sensitivity analysis approach to the effect homogeneity assumptions to understand how much our conclusions depend on this restriction (see, for example, Brookhart and Schneeweiss, 2007, for such an analysis for instrumental variables). These are the critical identification assumptions for our approach, and attention to them is crucially important.

References

Abadie, Alberto. 2005. "Semiparametric difference-in-differences estimators." *The Review of Economic Studies* 72(1):1–19.

Adida, Claire L., Adeline Lo and Melina R. Platas. 2018. "Perspective taking can promote short-term inclusionary behavior toward Syrian refugees." *Proceedings of the National Academy of Sciences* 115(38):9521–9526.

Allport, Gordon W. 1954. *The Nature of Prejuidice*. Basic Books.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Avin, Chen, Ilya Shpitser and Judea Pearl. 2005. Identifiability of Path-specific Effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05 San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 357–363.

Belloni, Alexandre and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19(2):521–547.

Bickel, P.J., C.A. Klaassen, Y. Ritov and J.A. Wellner. 1998. *Efficient and Adaptive Estimation for Semi-parametric Models*. New York: Springer-Verlag.

Blackwell, Matthew. 2014. "A Selection Bias Approach to Sensitivity Analysis for Causal Effects." *Political Analysis* 22(2):162–169.

Blackwell, Matthew and Anton Strezhnev. 2022. "Telescope matching for reducing model dependence in the estimation of the effects of time-varying treatments: An application to negative advertising." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 185(1):377–399.

Bradic, Jelena, Weijie Ji and Yuqian Zhang. 2021. "High-dimensional Inference for Dynamic Treatment Effects." *arXiv:2110.04924 [cs, econ, math, stat]* .

Broockman, David and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

Brookhart, M. Alan and Sebastian Schneeweiss. 2007. "Preference-Based Instrumental Variable Methods for the Estimation of Treatment Effects: Assessing Validity and Interpreting Results." *The International Journal of Biostatistics* 3(1).
URL: *https://doi.org/10.2202/1557-4679.1072*

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21(1):C1–C68.

Deuchert, Eva, Martin Huber and Mark Schelker. 2019. "Direct and Indirect Effects Based on Difference-in-Differences With an Application to Political Preferences Following the Vietnam Draft Lottery." *Journal of Business & Economic Statistics* 37(4):710–720.

Ding, Peng and Fan Li. 2019. "A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment." *Political Analysis* 27(4):605–615.

Dovidio, John F., Marleen ten Vergert, Tracie L. Stewart, Samuel L. Gaertner, James D. Johnson, Victoria M. Esses, Blake M. Riek and Adam R. Pearson. 2004. "Perspective and Prejudice: Antecedents and Mediating Mechanisms." 30(12):1537–1549.

Frölich, Markus and Martin Huber. 2017. "Direct and Indirect Treatment Effects–Causal Chains and Mediation Analysis with Instrumental Variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79(5):1645–1666.

Goetgeluk, Sylvie, Sijn Vansteelandt and Els Goetghebeur. 2008. "Estimation of controlled direct effects." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70(5):1049–1066.

Heckman, James J., Hidehiko Ichimura and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64(4):605–654.

Holm, Anders and Richard Breen. 2024. "Causal Mediation in Panel Data – Estimation Based on Difference in Differences.".

Huber, Martin, Mark Schelker and Anthony Strittmatter. 2022. "Direct and Indirect Effects based on Changes-in-Changes." *Journal of Business & Economic Statistics* 40(1):432–443.

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

Kang, Joseph D. Y. and Joseph L. Schafer. 2007. “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.” *Statistical Science* 22(4):523–539.

Kennedy, Edward H., Sivaraman Balakrishnan and Max G’Sell. 2020. “Sharp instruments for classifying compliers and generalizing causal effects.” *The Annals of Statistics* 48(4).

Kinder, Donald R. and David O. Sears. 1981. “Prejudice and Politics: Symbolic Racism Versus Racial Threats to the Good Life.” 40(3):414–431.

Kinder, Donald R. and Lynn M. Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. University of Chicago Press.

Krysan, Maria. 2000. “Prejudice, Politics, and Public Opinion: Understanding the Sources of Racial Policy Attitudes.” 26(1):135–168.

Lipsitch, Marc, Eric Tchetgen Tchetgen and Ted Cohen. 2010. “Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies.” *Epidemiology* 21(3):383–388.

Murphy, S A, M J van der Laan, J M Robins and Conduct Problems Prevention Research Group. 2001. “Marginal Mean Models for Dynamic Regimes.” *Journal of the American Statistical Association* 96(456):1410–1423.

Nelson, Thomas E. 1999. “Group Affect and Attribution in Social Policy Opinion.” 61(2):331–362.

Newey, Whitney K. 1990. “Semiparametric efficiency bounds.” *Journal of Applied Econometrics* 5(2):99–135.

Orellana, Liliana, Andrea Rotnitzky and James M. Robins. 2010. “Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content.” *International Journal of Biostatistics* 6(2):–.

Paluck, Elizabeth Levy. 2009. "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda." *Journal of Personality and Social Psychology* 96(3):574–587.

Paluck, Elizabeth Levy and Donald P. Green. 2009. "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice." *Annual Review of Psychology* 60(1):339–367.

Paluck, Elizabeth Levy, Seth A. Green and Donald P. Green. 2019. "The Contact Hypothesis Re-Evaluated." 3(2):129–158.

Petersen, Maya L., Sandra E. Sinisi and Mark J. van der Laan. 2006. "Estimation of direct causal effects." *Epidemiology* 17(3):276–284.

Pettigrew, Thomas F. and Linda R. Tropp. 2006. "A Meta-Analytic Test of Intergroup Contact Theory." 90(5):751–783.

Rabinowitz, Joshua L., David O. Sears, Jim Sidanius and Jon A. Krosnick. 2009. "Why Do White Americans Oppose Race-Targeted Policies? Clarifying the Impact of Symbolic Racism." 30(5):805–828.

Robins, James M. 1986. "A new approach to causal inference in mortality studies with sustained exposure periods-Application to control of the healthy worker survivor effect." *Mathematical Modelling* 7(9-12):1393–1512.
URL: *http://biosun1.harvard.edu/ robins/new-approach.pdf*

Robins, James M. 1999. Testing and Estimation of Direct Effects by Reparameterizing Directed Acyclic Graphs with Structural Nested Models. In *Computation, Causation, and Discovery*. AAAI Press.
URL: *https://doi.org/10.7551/mitpress/2006.003.0017*

Robins, James M. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, ed. P. J. Green, N. L. Hjort and S. Richardson. Oxford University Press pp. 70–81.

Robins, James M., Andrea Rotnitzky and Daniel O. Scharfstein. 1999. Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, ed. Willard Miller, M. Elizabeth Halloran and Donald Berry. Vol. 116 New York, NY: Springer New York pp. 1–94.

Sant'Anna, Pedro H.C. and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of Econometrics* 219(1):101–122.

Sears, David O. and Carolyn L. Funk. 1999. "Evidence of the Long-Term Persistence of Adults' Political Predispositions." *The Journal of Politics* 61(1):1–28.

Shahn, Zach, Oliver Dukes, David Richardson, Eric Tchetgen Tchetgen and James Robins. 2022. "Structural Nested Mean Models Under Parallel Trends Assumptions.".
URL: *https://arxiv.org/abs/2204.10291*

Sidanius, Jim, Felicia Pratto and Lawrence Bobo. 1996. "Racism, Conservatism, Affirmative Action, and Intellectual Sophistication: A Matter of Principled Conservatism or Group Dominance?" 70(3):476–490.

Simonovits, Gábor, Gábor Kézdi and Péter Kardos. 2018. "Seeing the World Through the Other's Eye: An Online Intervention Reducing Ethnic Prejudice." 112(1):186–193.

Sniderman, Paul M., Thomas Piazza, Philip E. Tetlock and Ann Kendrick. 1991. "The New Racism." 35(2):423–447.

Sofer, Tamar, David B. Richardson, Elena Colicino, Joel Schwartz and Eric J. Tchetgen Tchetgen. 2016. "On Negative Outcome Control of Unobserved Confounding as a Generalization of Difference-in-Differences." *Statistical Science* 31(3).

van der Laan, Mark J. and Susan Gruber. 2012. "Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions." *The International Journal of Biostatistics* 8(1).

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford: Oxford University Press.

Zhou, Xiang. 2022. "Semiparametric Estimation for Causal Mediation Analysis with Multiple Causally Ordered Mediators." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(3):794–821.

Zhou, Xiang and Geoffrey T. Wodtke. 2019. "A Regression-with-Residuals Method for Estimating Controlled Direct Effects." 27(3):360–369.

Zhou, Xiang and Teppei Yamamoto. 2023. "Tracing Causal Paths from Experimental and Observational Data." *The Journal of Politics* 85(1):250–265.
URL: *https://doi.org/10.1086/720310*

Supplemental Materials (to appear online)

A Bounding with a lagged dependent variable approach

In this section we contrast the targets of inference under the difference-in-differences framework and the sequential ignorability with lagged dependent variable framework. For simplicity, we assume a binary mediator and that $M_{i1} = 0$ throughout and suppress any such conditioning statement. Let $F_{Y_1}(y \mid d, m, \mathbf{x}, \mathbf{z})$ be the cumulative density function of Y_{i1} given $D_i = d$, $M_{i2} = m$, $\mathbf{X}_i = \mathbf{x}$, and $\mathbf{Z} = \mathbf{z}$, $G_{Y_1}(y \mid d, \mathbf{x}, \mathbf{z})$ be the same distribution function without conditioning on M_{i2}, and let $\overline{\mu}(d, m, \mathbf{x}, \mathbf{z}, y) = \mathbb{E}[Y_{i2} \mid D_i = d, M_{i2} = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}, Y_{i1} = y]$.

Next, we describe the targets of inference for both the DID and LDV approaches. These are quantities that, under each set of assumptions, identify the ACDE but remain valid observational quantities even when those assumptions do not hold. Our bracketing result will order these quantities and so is valid regardless of whether or not the identification assumptions actually hold. First, we write the quantity that, under parallel trends, would identify $\mathbb{E}[Y_{i2}(0, 0)]$:

$$\widetilde{\mu}_{0,\text{DID}} = \mathbb{E}[Y_{i1} \mid D_i = 0, M_{i2} = 0] + \int_{\mathbf{x},\mathbf{z}} \mathbb{E}[\Delta Y_i \mid D_i = 0, M_{i2} = 0, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}] dP(\mathbf{x}, \mathbf{z} \mid D_i = 0),$$

with $\widetilde{\mu}_{1,\text{DID}}$ being defined similarly. Under Assumption 1 and 2, $\widetilde{\tau}_{\text{DID}} = \widetilde{\mu}_{1,\text{DID}} - \widetilde{\mu}_{0,\text{DID}}$ would identify the ACDE, τ. Under a lagged dependent variable, the g-computational formula gives the following identification formula for $\mathbb{E}[Y_{i2}(0, 0)]$:

$$\widetilde{\mu}_{0,\text{LDV}} = \int_{\mathbf{x},\mathbf{z},y} \mathbb{E}[Y_{i2} \mid D_i = 0, M_{i2} = 0, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}, Y_{i1} = y] dG_{Y_1}(y \mid \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0),$$

with $\widetilde{\mu}_{1,\text{LDV}}$ defined similarly. If LDV sequential ignorability holds, $Y_{i2}(d, m) \perp\!\!\!\perp M_{i2} \mid D_i = d, \mathbf{X}_i, \mathbf{Z}_i, Y_{i1}$, then $\widetilde{\tau}_{\text{LDV}} = \widetilde{\mu}_{1,\text{LDV}} - \widetilde{\mu}_{0,\text{LDV}}$ would identify the ACDE.

Theorem 3. *The difference between $\widetilde{\mu}_{0,DID}$ and $\widetilde{\mu}_{0,LDV}$ is*

$$\begin{aligned} \widetilde{\tau}_{DID} - \widetilde{\tau}_{LDV} = &\int_{\mathbf{x},\mathbf{z},y} \Delta_1(y) \left(dF_{Y_1}(y \mid 1, 0, \mathbf{x}, \mathbf{z}) - dG_{Y_1}(y \mid 1, \mathbf{x}, \mathbf{z})\right) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0) \\ &- \int_{\mathbf{x},\mathbf{z},y} \Delta_0(y) \left(dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) - dG_{Y_1}(y \mid 0, \mathbf{x}, \mathbf{z})\right) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0), \end{aligned} \tag{11}$$

where $\Delta_d(y) = \overline{\mu}(d, 0, \mathbf{x}, \mathbf{z}, y) - y$.

Proof. Using iterated expectations, we can write $\widetilde{\mu}_{0,\mathrm{DID}}$ as

$$\begin{aligned}\widetilde{\mu}_{0,\mathrm{DID}} &= \int_{\mathbf{x},\mathbf{z},y} y dG_{Y_1}(y \mid \mathbf{x},\mathbf{z}) dP(\mathbf{x},\mathbf{z} \mid D_i = 0) \\ &+ \int_{\mathbf{x},\mathbf{z},y} \mathbb{E}[\Delta Y_i \mid D_i = 0, M_{i2} = 0, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}, Y_{i1} = y] dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) dP(\mathbf{x},\mathbf{z} \mid D_i = 0) \\ &= \int_{\mathbf{x},\mathbf{z},y} y dG_{Y_1}(y \mid \mathbf{x},\mathbf{z}) dP(\mathbf{x},\mathbf{z} \mid D_i = 0) \\ &+ \int_{\mathbf{x},\mathbf{z},y} \Delta_0(y) dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) dP(\mathbf{x},\mathbf{z} \mid D_i = 0)\end{aligned}$$

Combining this with the definition of $\widetilde{\mu}_{0,\mathrm{LDV}}$, we obtain

$$\begin{aligned}\widetilde{\mu}_{0,\mathrm{DID}} - \widetilde{\mu}_{0,\mathrm{LDV}} = &\int_{\mathbf{x},\mathbf{z},y} \Delta(y) dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) dP(\mathbf{x},\mathbf{z} \mid D_i = 1) \\ &- \int_{\mathbf{x},\mathbf{z},y} \Delta_0(y) dG_{Y_1}(y \mid \mathbf{x},\mathbf{z}) dP(\mathbf{x},\mathbf{z} \mid D_i = 0).\end{aligned}$$

Applying the same logic to $\widetilde{\mu}_{1,\mathrm{DID}} - \widetilde{\mu}_{1,\mathrm{LDV}}$ yields the result. □

Our ACDE-PC estimand, on the other hand, has a more specific relationship with the sequential ignorability approach. In fact, because the identification assumptions for that estimand are simply parallel trends for a four-category outcome, we can apply the results of Ding and Li (2019) to obtain a bracketing result between the DID estimand and the LDV estimand. Let $\widetilde{\gamma}_{\mathrm{DID}}$ and $\widetilde{\gamma}_{\mathrm{LDV}}$ be the targets of inference for these two settings, identified in a similar manner to the two above. Following Ding and Li (2019), we first invoke conditions on the data generating process:

Condition 1 (Stationarity)**.** $\partial\overline{\mu}(d, m, \mathbf{x}, \mathbf{z}, y)/\partial y < 1$ *for all* y.

Condition 2 (Stochastic Monotonicity)**.** *Either (a)* $F_{Y_1}(y \mid d, 1, \mathbf{x}, \mathbf{z}) \geq F_{Y_1}(y \mid d, 0, \mathbf{x}, \mathbf{z})$ *for all* y*; or (b)* $F_{Y_1}(y \mid d, 0, \mathbf{x}, \mathbf{z}) \geq F_{Y_1}(y \mid d, 1, \mathbf{x}, \mathbf{z})$.

Condition 1 is a limit on the growth of the time series of the outcome and with a linear model, it would require that the coefficient on the lagged dependent variable be less than one. This is a commonly invoked assumption with panel and time-series data. Condition 2 characterizes the relationship between the lagged dependent variable and the mediator, with Condition 2(a) meaning that the group with $M_{i2} = 1$ has higher baseline outcomes across the entire distribution compared to the

$M_{i2} = 0$ group and vice versa for Condition 2(b). We say Condition 1 and 2 are conditions rather than assumptions because they are both empirically testable (Ding and Li, 2019).

Ding and Li (2019) have shown that under Conditions 1 and 2(a) we have $\widetilde{\gamma}_{\text{DID}} \geq \widetilde{\gamma}_{\text{LDV}}$, and under Conditions 1 and 2(b), we have $\widetilde{\gamma}_{\text{DID}} \leq \widetilde{\gamma}_{\text{LDV}}$. Thus, if one of these two sets of conditions holds and one of the two sets of identifying assumptions holds, then the two estimands will bracket the true value of the ACDE-PC.

B Proofs

B.1 Identification

Proof of Proposition 1. Here, we first prove the IPW identification result for τ_m. The proof for γ_m is very similar and so we omit it. Below we combine $\mathbf{X}_i$ and $\mathbf{Z}_i$ into a single vector $\mathbf{X}_i$ since their role in the proof is the same. We begin with the first term of τ_m. By randomization and the law of total probability we have:

$$
\begin{aligned}
\mathbb{E}\{Y_{i2}(1,m) \mid M_{i1} = m\} &= \mathbb{E}\{Y_{i2}(1,m) \mid D_i = 1, M_{i1} = m\} \\
&= \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(1,m) \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m) \\
&= \mathbb{E}(Y_{i1}(0,m) \mid D_i = 1, M_{i1} = m) \\
&\quad + \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(1,m) - Y_{i1}(0,m) \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m)
\end{aligned}
$$

The first term is identified and, using Assumption 2 we can write the second term as:

$$
\int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i(1,m) \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m)
$$

Let $\overline{\pi}_{dm}(k) = \mathbb{P}(M_{i2} = m \mid D_i = d, M_{i1} = k)$. By consistency and then Bayes' rule, this becomes,

$$
\begin{aligned}
&\int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = 0) \\
&= \int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} \frac{\overline{\pi}_{1m}(m)}{\pi_{1m}(m, \mathbf{x})} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m, M_{i2} = m)
\end{aligned}
$$

Once again applying the law of total probability and then using the definition of conditional probability, we can simplify this to:

$$
\mathbb{E}\left\{ \frac{\Delta Y_i}{\pi_{1m}(m, \mathbf{X}_i)} \,\middle|\, D_i = 1, M_{i1} = m, M_{i2} = m \right\} (\overline{\pi}_{1m}(m)) = \mathbb{E}\left\{ \frac{W_{i1m} D_i W_{i2m} \Delta Y_i}{\rho_m \delta \pi_{1m}(m, \mathbf{X}_i)} \right\}
$$

Thus, we can write the first term in the τ_m (using randomization on the first term):

$$\mathbb{E}\{Y_{i2}(1,m) \mid M_{i1} = m\} = \mathbb{E}(Y_{i1}(0,m) \mid M_{i1} = m) + \mathbb{E}\left\{\frac{W_{i1m}D_iW_{i2m}}{\rho_m\delta\pi_{1m}(m,\mathbf{X}_i)}\Delta Y_i\right\}$$

We now turn to the second term of the τ_m. Again using the law of total probability and Assumption 2, we have:

$$\begin{aligned}
&\mathbb{E}\{Y_{i2}(0,m) \mid M_{i1} = m\} = \mathbb{E}\{Y_{i2}(0,m) \mid D_i = 0, M_{i1} = m\} \\
&= \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(0,m) \mid D_i = 0, M_{i1} = m, \mathbf{X}_i = \mathbf{x}\} dP(\mathbf{x} \mid D_i = 0, M_{i1} = 0) \\
&= \mathbb{E}(Y_{i1}(0,m) \mid M_{i1} = m) \\
&\qquad + \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(0,m) - Y_{i1}(0,m) \mid D_i = 0, M_{i1} = m, \mathbf{X}_i = \mathbf{x}\} dP(\mathbf{x} \mid D_i = 0, M_{i1} = m)
\end{aligned}$$

Once again, using the law of total probability and Assumption 2, this term becomes:

$$\begin{aligned}
&\int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(0,m) - Y_{i1}(0,m) \mid D_i = 0, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} dP(\mathbf{x} \mid D_i = 0, M_{i1} = m) \\
&\int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i = 0, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} dP(\mathbf{x} \mid D_i = 0, M_{i1} = m) \\
&= \int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i = 0, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} \frac{\overline{\pi}_{0m}(m)}{\pi_{0m}(m,\mathbf{x})} dP(\mathbf{x} \mid D_i = 0, M_{i1} = m, M_{i2} = m)
\end{aligned}$$

Finally, using the law of total probability and the definition of conditional expectation, we can write this term as:

$$\begin{aligned}
&\mathbb{E}\left\{\frac{\Delta Y_i}{\rho_m(1-\delta)\pi_{0m}(m,\mathbf{X}_i)} \,\middle|\, D_i = 0, M_{i1} = m, M_{i2} = m\right\} \overline{\pi}_{0m}(m) \\
&\qquad = \mathbb{E}\left\{\frac{W_{i1m}(1-D_i)W_{i2m}}{\rho_m(1-\delta)\pi_{0m}(m,\mathbf{X}_i)}\Delta Y_i\right\}
\end{aligned}$$

Combining this with the results on the first term gives the desired result for τ_m.

For the regression identification formulas, note that under our assumptions we have

$$\begin{aligned}
\mathbb{E}[\Delta Y_i(1,m) \mid M_{i0} = m] &= \mathbb{E}[\Delta Y_i(1,m) \mid M_{i0} = m, D_i = 1] \\
&= \mathbb{E}\left[\mathbb{E}[\Delta Y_i(1,m) \mid M_{i0} = m, D_i = 1, \mathbf{X}_i, \mathbf{Z}_i] \mid M_{i0} = m, D_i = 1\right] \\
&= \mathbb{E}\left[\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i) \mid M_{i0} = m, D_i = 1\right] \\
&= \mathbb{E}\left[\frac{W_{i1m}D_i}{\rho_m\delta}\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)\right].
\end{aligned}$$

The first equality holds by randomization, the second by iterated expectations, the third by the definition of μ_{dm}, and the fourth by the definition of conditional expectation. A similar result holds for $\mathbb{E}[\Delta Y_i(0,m) \mid M_{i0} = m]$ which obtains the identification.

$\square$

B.2 Multiple robustness

Proof of Theorem 1. We write $\psi_m(\mathbf{O}_i;\boldsymbol{\eta}_\tau) = \psi_{1m}(\mathbf{O}_i;\boldsymbol{\eta}_\tau) - \psi_{0m}(\mathbf{O}_i;\boldsymbol{\eta}_\tau)$, where

$$\psi_{1m}(\mathbf{O}_i;\boldsymbol{\eta}_\tau) = \left(\frac{W_{i1m}D_iW_{i2m}}{\rho_m\delta\pi_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i)}\right)(\Delta Y_i - \mu_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i)) + \frac{W_{i1m}D_i}{\rho_m\delta}\mu_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i)$$
$$\psi_{0m}(\mathbf{O}_i;\boldsymbol{\eta}_\tau) = \left(\frac{W_{i1m}(1-D_i)W_{i2m}}{\rho_m(1-\delta)\pi_{0m}(m,\mathbf{X}_i,\mathbf{Z}_i)}\right)(\Delta Y_i - \mu_{0m}(m,\mathbf{X}_i,\mathbf{Z}_i)) + \frac{W_{i1m}(1-D_i)}{\rho_m(1-\delta)}\mu_{0m}(m,\mathbf{X}_i,\mathbf{Z}_i).$$

We demonstrate the double robustness result on the first expression $\widehat{\psi}_{i,m,1}$ with the corresponding result for $\widehat{\psi}_{i,m,0}$ following similarly. The goal is to show that $\mathbb{P}_n\{\psi_{1m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}_\tau)\} \xrightarrow{p} \mathbb{E}[\Delta Y_i(1,m) \mid M_{i1} = m]$ under the cases described in the Theorem. We first consider the case where the propensity score model is correctly specified, so that $\widehat{\boldsymbol{\eta}}_\tau \xrightarrow{p} (\pi_{dm}, \mu^*_{dm}, \rho_m, \delta)$, where μ^*_{1m} and ν^*_{1m} are functions that do not necessarily correspond to μ_{1m} and ν_{1m}. Note that $\widehat{\delta} \xrightarrow{p} \delta$ and $\widehat{\rho}_m \xrightarrow{p} \rho_m$ by the LLN. Then by Slutsky's Theorem, we can write $\mathbb{P}_n\{\psi_{1m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}_\tau)\}$ as

$$\begin{aligned}
&\mathbb{E}\left\{\left(\frac{W_{i1m}D_iW_{i2m}}{\rho_m\delta\pi_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i)}\right)(\Delta Y_i - \mu^*_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i))\right\} + \mathbb{E}\left\{\frac{W_{i1m}D_i}{\rho_m\delta}\mu^*_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i)\right\} + o_p(1)\\
&= \mathbb{E}\left\{\left(\frac{W_{i1m}D_i}{\rho_m\delta}\right)(\mu_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i) - \mu^*_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i))\right\} + \mathbb{E}\left\{\frac{W_{i1m}D_i}{\rho_m\delta}\mu^*_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i)\right\} + o_p(1)\\
&= \mathbb{E}\left\{\frac{W_{i1m}}{\rho_m}\mu_{1m}(m,\mathbf{X}_i,\mathbf{Z}_i)\right\} + o_p(1) = \mathbb{E}[\Delta Y_i(1,m) \mid M_{i1} = m] + o_p(1)
\end{aligned}$$

The first equality follows from iterated expectations and the definition of π_{dm}, the second by randomization of D_i and the last by the fact that

$$\mu_{dm}(k,\mathbf{X}_i,\mathbf{Z}_i) = \mathbb{E}[\Delta Y_i(d,m) \mid M_{i1} = k,\mathbf{X}_i,\mathbf{Z}_i],$$

and the definition of conditional expectation. This, combined with the equivalent result for ψ_{0m}, establishes consistency when the propensity score model is correct.

Now we turn to the setting where the outcome regressions are correctly specified so that $\widehat{\boldsymbol{\eta}}_\tau \xrightarrow{p} (\pi^*_{dm}, \mu_{dm}, \rho_m, \delta)$ With these, we can write $\mathbb{P}_n\{\psi_{1m}(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_\tau)\}$ as

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\frac{W_{i1m}D_iW_{i2m}}{\rho_m\delta\pi^*_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)}\right)(\Delta Y_i - \mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i))\right\} + \mathbb{E}\left\{\frac{W_{i1m}D_i}{\rho_m\delta}\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)\right\} + o_p(1) \\
&= \mathbb{E}\left\{\left(\frac{W_{i1m}D_iW_{i2m}}{\rho_m\delta\pi^*_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)}\right)(\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i) - \mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i))\right\} + \mathbb{E}\left\{\frac{W_{i1m}D_i}{\rho_m\delta}\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)\right\} + o_p(1) \\
&= \mathbb{E}\left\{\frac{W_{i1m}}{\rho_m}\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)\right\} + o_p(1) = \mathbb{E}[\Delta Y_i(1, m) \mid M_{i1} = m] + o_p(1).
\end{aligned}
$$

This, combined with the equivalent result for ψ_{0m}, establishes consistency when the outcome regressions are correct. The result for γ_m also follows similarly.

□

B.3 Efficient influence function

Here we show that the influence functions for our doubly robust estimators are (uncentered) versions of the *efficient influence functions* (EIFs) for our target parameters. EIFs are important to nonparametric and semiparametric estimators because the variance of the efficient influence function serves as a lower bound for the mean squared error of any estimator across any distribution consistent with the identification assumptions. This is a form of "minimax" lower bound: no estimator can achieve a lower worst-case mean square error than this bound. If our estimators have that same influence function, then we hope that these estimators will obtain this bound, at least asymptotically. We now show that once we center the influence functions for our identification results, we obtain the EIFs and the semiparametric efficiency bounds.

We now define the centered EIFs for our estimands. Let $\nu_{dm} = \mathbb{E}[Y_{i2}(d, m) \mid M_{i0} = m]$. For ACDE-BC, define the following:

$$
\begin{aligned}
\widetilde{\psi}_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) &= \psi_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) - \frac{W_{i1m}D_i}{\rho_m\delta}\nu_{1m} \\
\widetilde{\psi}_{0m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) &= \psi_{0m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) - \frac{W_{i1m}(1 - D_i)}{\rho_m(1 - \delta)}\nu_{0m} \\
\widetilde{\psi}_{m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) &= \widetilde{\psi}_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) - \widetilde{\psi}_{0m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau)
\end{aligned}
$$

For the ACDE-PC, define the centered version of the influence function as:

$$\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) = \phi_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) - \frac{W_{i1m} D_i W_{i2m}}{\lambda_{dm}} \gamma_m.$$

Theorem 4. *(a) Under Assumptions 1, 2, 3, 5, and suitable regularity conditions, the efficient influence function for τ_m is $\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau)$, and the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau)^2]$. (b) Under Assumptions 1, 4, 5, and suitable regularity conditions, the efficient influence function for γ_m is $\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma)$, and the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma)^2]$.*

The regularity conditions here involve technical requirements to ensure pathwise differentiability of the efficient influence function. See, for example, Bickel et al. (1998, Chapter 3) for more details on these conditions.

Proof of Theorem 4. Define the collection of potential outcomes in each period as $\mathbf{Y}_{i2}(\bullet) = \{Y_{i2}(0, m), Y_{i2}(1, m)\}_{m \in \mathcal{M}}$ and $\mathbf{Y}_{i1}(\bullet) = \{Y_{i1}(0, m)\}_{m \in \mathcal{M}}$ with representative values $\mathbf{y}_2(\bullet)$ and $\mathbf{y}_1(\bullet)$, respectively. Then the full data is given by

$$\mathbf{H}_i = (\mathbf{Y}_{i2}(\bullet), \mathbf{Y}_{i1}(\bullet), M_{i2}, \mathbf{Z}_i, D_i, \mathbf{X}_i, M_{i1}),$$

and let $\mathbf{h}$ be a possible value of $\mathbf{H}$. Then the density of $\mathbf{H}$ for some sigma-finite measure is

$$\begin{aligned}
\overline{q}(\mathbf{h}) = \prod_{m_2 \in \mathcal{M}} \prod_{m_1 \in \mathcal{M}} & \overline{f}(\mathbf{y}_2(\cdot), \mathbf{y}_1(\cdot) \mid m_2, D_i = 1, m_1, \mathbf{z}, \mathbf{x})^{w_{m_2} d w_{m_1}} \\
& \times \overline{f}(\mathbf{y}_2(\cdot), \mathbf{y}_1(\cdot) \mid m_2, D_i = 0, m_1, \mathbf{z}, \mathbf{x})^{w_{m_2}(1-d) w_{m_1}} \\
& \times \pi_{1m_2}(m_1, \mathbf{z}, \mathbf{x})^{w_{m_2} d w_{m_1}} \pi_{0m_2}(m_1, \mathbf{z}, \mathbf{x})^{w_{m_2}(1-d) w_{m_1}} \qquad , \\
& \times f(\mathbf{z} \mid D_i = 1, m_1, \mathbf{x})^{d w_{m_1}} f(\mathbf{z} \mid D_i = 0, m_1, \mathbf{x})^{(1-d) w_{m_1}} \\
& \times f(\mathbf{x} \mid m_1)^{w_{m_1}} \delta^d (1 - \delta)^{(1-d)} \rho_{m_1}^{w_{m_1}}
\end{aligned}$$

where w_{m_1} is 1 when $M_{i1} = m_1$ and 0 otherwise, with w_{m_2} defined similarly. In addition to the propensity scores that have already been defined, this density contains the following:

- $\overline{f}(\mathbf{y}_2(\cdot), \mathbf{y}_1(\cdot) \mid m_2, D_i = d, m_1, \mathbf{z}, \mathbf{x})$ is the density of the potential outcomes conditional on $M_{i2} = m_2$, $D_i = d$, $M_{i1} = m_1$, $\mathbf{Z}_i = \mathbf{z}$, and $\mathbf{X}_i = \mathbf{x}$, where $m_1, m_2 \in \mathcal{M}$, $d \in \{0, 1\}$, $\mathbf{z} \in \mathbb{R}^{k_z}$, and $\mathbf{x} \in \mathbb{R}^{k_x}$.

- $f(\mathbf{z} \mid D_i = d, m_1, \mathbf{x})$ is the density of $\mathbf{Z}_i$ conditional on $D_i = d, \mathbf{X}_i = \mathbf{x}$, and $M_{i1} = m_1$.

- $f(\mathbf{x} \mid m_1)$ is the density of $\mathbf{X}_i$ conditional on $M_{i1} = m_1$.

We now the turn to the density of the observed data, $\mathbf{O}_i = (Y_{i2}, Y_{i1}, M_{i2}, \mathbf{Z}_i, D_i, \mathbf{X}_i, M_{i1})$. We write the density of the observed outcomes as

$$f(y_2, y_1 \mid m_2, 1, m_1, \mathbf{z}, \mathbf{x}),$$

which marginalizes the $\overline{f}(\cdot)$ over the potential outcomes where $D_i \neq 1$, $M_{i2} \neq m_2$, or $M_{i1} \neq m_1$. Consider a possible value of the observed data

$$\mathbf{o} = (y_2, y_1, j_2, d, j_1, z, x)'$$

The density of the observed data $\mathbf{O}_i$ can be written as

$$\begin{aligned} q(\mathbf{o};\theta) = \prod_{m_2 \in \mathcal{M}} \prod_{m_1 \in \mathcal{M}} & \left[f(y_2, y_1 \mid m_2, 1, m_1, \mathbf{z}, \mathbf{x}) \pi_{1m_2}(m_1, \mathbf{z}, \mathbf{x}) \right]^{d\mathbf{1}(m_2 = j_2, m_1 = j_1)} \\ & \times \left[f(y_2, y_1 \mid m_2, 0, m_1, \mathbf{z}, \mathbf{x}) \pi_{0m_2}(m_1, \mathbf{z}, \mathbf{x}) \right]^{(1-d)\mathbf{1}(m_2 = j_2, m_1 = j_1)} \\ & \times \left[f(\mathbf{z} \mid D_i = 1, m_1, \mathbf{x})^d f(\mathbf{z} \mid D_i = 0, m_1, \mathbf{x})^{(1-d)} \right]^{\mathbf{1}(m_1 = j_1)} \\ & \times f(\mathbf{x} \mid m_1)^{\mathbf{1}(m_1 = j_1)} \delta^d (1-\delta)^{(1-d)} \rho_{m_1}^{\mathbf{1}(m_1 = j_1)}. \end{aligned}$$

We consider a regular parametric submodel for the joint distribution of $\mathbf{O}_i$, with log likelihood

$$\begin{aligned} & \log q(\mathbf{o};\theta) = \\ & \sum_{m_2 \in \mathcal{M}} \sum_{m_1 \in \mathcal{M}} \Big[d\mathbf{1}(m_2 = j_2, m_1 = j_1) \left(\log f(y_2, y_1 \mid m_2, 1, m_1, \mathbf{z}, \mathbf{x}; \theta) + \log \pi_{1m_2}(m_1, \mathbf{z}, \mathbf{x}; \theta) \right) \\ & \qquad + (1-d)\mathbf{1}(m_2 = j_2, m_1 = j_1) \left(\log f(y_2, y_1 \mid m_2, 0, m_1, \mathbf{z}, \mathbf{x}; \theta) + \log \pi_{0m_2}(m_1, \mathbf{z}, \mathbf{x}; \theta) \right) \Big] \\ & + \sum_{m_1 \in \mathcal{M}} \mathbf{1}(m_1 = j_1) \left(d \log f(\mathbf{z} \mid 1, m_1, \mathbf{x}; \theta) + (1-d) \log f(\mathbf{z} \mid 0, m_1, \mathbf{x}; \theta) + \log f(x \mid m_1; \theta) \right) \end{aligned}$$

where, $q(\cdot;\theta_0) = q(\cdot)$ so that θ_0 is the true value of the parameters. This parametric submodel yields the following score:

$$S(\mathbf{o};\theta) = S_y(y_2, y_1, j_2, s, j_1, \mathbf{z}, \mathbf{x}; \theta) + S_m(j_2, s, j_1, \mathbf{z}, \mathbf{x}; \theta) + S_z(\mathbf{z}, j_1, s, \mathbf{x}; \theta) + S_x(\mathbf{x}, j_1; \theta)$$

where,

$$
\begin{aligned}
S_y(y_2, y_1, j_2, s, j_1, \mathbf{z}, \mathbf{x}; \theta) &= \sum_{m_1 \in \mathcal{M}} \sum_{d \in \{0,1\}} \sum_{m_2 \in \mathcal{M}} \mathbf{1}\left(m_1 = j_1, d = s, m_2 = j_2\right) \frac{d}{d\theta} \log f(y_2, y_1 \mid m_2, d, m_1, \mathbf{z}, \mathbf{x}; \theta) \\
S_m(j_2, s, j_1, \mathbf{z}, \mathbf{x}; \theta) &= \sum_{m_1 \in \mathcal{M}} \sum_{d \in \{0,1\}} \sum_{m_2 \in \mathcal{M}} \mathbf{1}\left(m_1 = j_1, d = s, m_2 = j_2\right) \frac{\dot{\pi}_{dm_2}(m_1, \mathbf{z}, \mathbf{x}; \theta)}{\pi_{dm_2}(m_1, \mathbf{z}, \mathbf{x}; \theta)} \\
S_z(\mathbf{z}, s, j_1, \mathbf{x}; \theta) &= \sum_{m_1 \in \mathcal{M}} \sum_{d \in \{0,1\}} \mathbf{1}\left(m_1 = j_1, d = s\right) \frac{d}{d\theta} \log f(\mathbf{z} \mid d, m_1, \mathbf{x}; \theta) \\
S_x(\mathbf{x}, j_1; \theta) &= \sum_{m_1 \in \mathcal{M}} \mathbf{1}(m_1 = j_1) \frac{d}{d\theta} \log f(\mathbf{x} \mid m_1; \theta)
\end{aligned}
$$

Let $L_0^2(F_W)$ be the usual Hilbert space of zero-mean, square-integrable functions with respect to the distribution F_W. The tangent space of the model is $\mathcal{H} = \mathcal{H}_y + \mathcal{H}_m + \mathcal{H}_z + \mathcal{H}_x$, where

$$
\begin{aligned}
\mathcal{H}_y &= \{S_y(Y_{i2}, Y_{i1}, M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) : S_y(Y_{i2}, Y_{i1}, M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) \in L_0^2(F_{Y_2, Y_1 | M_2, D, M_1, \mathbf{Z}, \mathbf{X}})\} \\
\mathcal{H}_m &= \{S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) : S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) \in L_0^2(F_{M_2 | D, M_1, \mathbf{Z}, \mathbf{X}})\} \\
\mathcal{H}_z &= \{S_z(\mathbf{Z}_i, D_i, M_{i1}, \mathbf{X}_i) : S_z(\mathbf{Z}_i, D_i, M_{i1}, \mathbf{X}_i) \in L_0^2(F_{\mathbf{Z} | D, M_1, \mathbf{X}})\} \\
\mathcal{H}_x &= \{S_x(\mathbf{X}_i, M_{i1}) : S_x(\mathbf{X}_i, M_{i1}) \in L_0^2(F_{\mathbf{X} | M_1})\},
\end{aligned}
$$

The further restrictions on the tangent space are that we have $\mathbb{E}[S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) \mid D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i] = \sum_{m_2 \in \mathcal{M}} \dot{\pi}_{D_i, m_2}(M_{i1}, \mathbf{Z}_i, \mathbf{X}_i)$ and

$$
\mathbb{E}[S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i)^2 \mid D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i] = \sum_{m_2 \in \mathcal{M}} \dot{\pi}_{D_i m_2}(M_{i1}, \mathbf{Z}_i, \mathbf{X}_i)^2 / \pi_{D_i m_2}(M_{i1}, \mathbf{Z}_i, \mathbf{X}_i).
$$

We can write the ACDE as a function of the regular parametric submodel as

$$
\begin{aligned}
\tau_m(\theta) = &\int_{\mathbf{x}} \int_{\mathbf{z}} \int_{y_1, y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x}; \theta) f(\mathbf{z} \mid 1, m, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta) dy_1 dy_2 d\mathbf{z} d\mathbf{x} \\
&- \int_{\mathbf{x}} \int_{\mathbf{z}} \int_{y_1, y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 0, m, \mathbf{z}, \mathbf{x}; \theta) f(\mathbf{z} \mid 0, m, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta) dy_1 dy_2 d\mathbf{z} d\mathbf{x},
\end{aligned}
$$

where $\tau_m(\theta_0) = \tau_m$.

Our proposed influence function will be the efficient influence function if it is in the tangent space $\mathcal{H}$ and meets the following condition:

$$
\frac{\partial \tau_m(\theta_0)}{\partial \theta} = \mathbb{E}\left[\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau) S(\mathbf{O}_i; \theta_0)\right].
$$

We can derive the pathwise derivative as

$$
\begin{aligned}
\frac{\partial \tau_m(\theta_0)}{\partial \theta} = & \int_{\mathbf{x}}\int_{\mathbf{z}}\int_{y_1,y_2} \big[(y_2 - y_1)S(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x}) f(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x}) f(\mathbf{z} \mid 1, m, \mathbf{x}) \\
& \qquad \times f(\mathbf{x} \mid m) dy_2 dy_2 d\mathbf{z} d\mathbf{x}\big] \\
& + \int_{\mathbf{x}}\int_{\mathbf{z}}\int_{y_1,y_2} \big[(y_2 - y_1)S(y_2, y_1 \mid m, 0, m, \mathbf{z}, \mathbf{x}) f(y_2, y_1 \mid m, 0, m, \mathbf{z}, \mathbf{x}) f(\mathbf{z} \mid 0, m, \mathbf{x}) \\
& \qquad \times f(\mathbf{x} \mid m) dy_2 dy_2 d\mathbf{z} d\mathbf{x}\big] \\
& + \int_{\mathbf{x}}\int_{\mathbf{z}} (\mu_{1m}(m, \mathbf{z}, \mathbf{x})) S(\mathbf{z} \mid 1, m, \mathbf{x}) f(\mathbf{z} \mid 1, m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{z} d\mathbf{x} \\
& + \int_{\mathbf{x}}\int_{\mathbf{z}} (\mu_{0m}(m, \mathbf{z}, \mathbf{x})) S(\mathbf{z} \mid 0, m, \mathbf{x}) f(\mathbf{z} \mid 0, m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{z} d\mathbf{x} \\
& + \int_{\mathbf{x}}\int_{\mathbf{z}} (\mu_{1m}(m, \mathbf{z}, \mathbf{x})) S(\mathbf{x} \mid m) f(\mathbf{z} \mid 1, m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{z} d\mathbf{x} \\
& + \int_{\mathbf{x}}\int_{\mathbf{z}} (\mu_{0m}(m, \mathbf{z}, \mathbf{x})) S(\mathbf{x} \mid m) f(\mathbf{z} \mid 0, m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{z} d\mathbf{x},
\end{aligned}
\tag{12}
$$

Upon inspection, $\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau)$ satisfies the condition and is in $\mathcal{H}$. For example, the first line of (12) can be rewritten

$$\mathbb{E}\left[\frac{W_{i1m} D_i W_{i2m}}{\rho_m \delta \pi_{1m}(\mathbf{X}_i, \mathbf{Z}_i)} (\Delta Y_i - \mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)) S_y(Y_{i2}, Y_{i1}, M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i)\right];$$

the third line can be written as

$$\mathbb{E}\left[\frac{W_{i1m} D_i}{\rho_m \delta} (\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i) - \nu_{1m}) S_z(\mathbf{Z}_i, D_i, M_{i1}, \mathbf{X}_i)\right];$$

and the fifth term as

$$\mathbb{E}\left[\frac{W_{i1m} D_i}{\rho_m \delta} (\mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i) - \nu_{1m}) S_x(\mathbf{X}_i, M_{i1})\right].$$

Using the orthogonality of the score functions, we can combine these to show that they are equal to $\mathbb{E}[\widetilde{\psi}_{1m}(\mathbf{O}_i; \boldsymbol{\eta}_\tau) S(\mathbf{O}_i; \theta_0)]$. Combining these steps with similar derivations for the other lines in (12)

Thus, by Theorem 3.1 of Newey (1990), $\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau)$ is the efficient influence function for τ_m and the latter is a pathwise differentiable parameter. This also implies that the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\tau)^2]$.

For our other estimand, note that

$$
\begin{aligned}
\gamma_m = \mathbb{E}\,[& \mathbb{E}\,[\Delta Y_i \mid M_{i1} = m, D_i = 1, M_{i2} = m, X_i] \\
& - \mathbb{E}\,[\Delta Y_i \mid M_{i1} = m, D_i = 1, M_{i2} = m, X_i] \mid M_{i1} = m, D_i = 1, M_{i2} = m]
\end{aligned}
$$

Thus, under the regular parametric submodel, we can write this as

$$
\begin{aligned}
\gamma_m(\theta) = & \frac{\int_{\mathbf{x}} \int_{y_1,y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 1, m, \mathbf{x}; \theta) \pi_{1m}(m, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta) dy_1 dy_2 d\mathbf{x}}{\int_{\mathbf{x}} \pi_{1m}(m, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta)} \\
& - \frac{\int_{\mathbf{x}} \int_{y_1,y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 0, m, \mathbf{x}; \theta) \pi_{1m}(m, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta) dy_1 dy_2 d\mathbf{x}}{\int_{x} \pi_{1m}(m, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta)}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\frac{\partial \gamma_m(\theta_0)}{\partial \theta} = & \frac{\int_{\mathbf{x}} \int_{y_1,y_2} (y_2 - y_1) S(y_2, y_1 \mid m, 1, m, \mathbf{x}) f(y_2, y_1 \mid m, 1, m, \mathbf{x}) \pi_{1m}(m, \mathbf{x}) f(\mathbf{x} \mid m) dy_1 dy_2 d\mathbf{x}}{\mathbb{E}[W_{i1m} D_i W_{i2m}]/\delta \rho_m} \\
& - \frac{\int_{\mathbf{x}} \int_{y_1,y_2} (y_2 - y_1) S(y_2, y_1 \mid m, 0, m, \mathbf{x}) f(y_2, y_1 \mid m, 0, m, \mathbf{x}) \pi_{1m}(m, \mathbf{x}) f(\mathbf{x} \mid m) dy_1 dy_2 d\mathbf{x}}{\mathbb{E}[W_{i1m} D_i W_{i2m}]/\delta \rho_m} \\
& + \frac{\int_{\mathbf{x}} \left(\mu_{1m}(m, \mathbf{x}) - \mu_{0m}(m, \mathbf{x}) - \gamma_m\right) \dot{\pi}_{1m}(m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{x}}{\mathbb{E}[W_{i1m} D_i W_{i2m}]/\delta \rho_m} \\
& + \frac{\int_{\mathbf{x}} \left(\mu_{1m}(m, \mathbf{x}) - \mu_{0m}(m, \mathbf{x}) - \gamma_m\right) \pi_{1m}(m, \mathbf{x}) S(\mathbf{x} \mid m) f(\mathbf{x} \mid m) d\mathbf{x}}{\mathbb{E}[W_{i1m} D_i W_{i2m}]/\delta \rho_m}
\end{aligned}
$$

To verify that it is in $\mathcal{H}$, we can rewrite $\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma)$ as

$$
\begin{aligned}
\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) = & \left(\frac{W_{i1m} D_i W_{i2m}}{\lambda_{1m}}\right) \left(\Delta Y_i - \mu_{1m}(m, \mathbf{X}_i)\right) \\
& - \left(\frac{W_{i1m}(1 - D_i) W_{i2m}}{\lambda_{1m}}\right) \left(\frac{\pi_{1m}(m, \mathbf{X}_i)\delta}{\pi_{0m}(m, \mathbf{X}_i)(1 - \delta)}\right) \left(\Delta Y_i - \mu_{i,0m}\right) \\
& + \frac{W_{i1m} D_i}{\lambda_{1m}} \left(W_{i2m} - \pi_{1m}(m, \mathbf{X}_i)\right) \left(\mu_{1m}(m, \mathbf{X}_i) - \mu_{0m}(m, \mathbf{X}_i) - \gamma_m\right) \\
& + \frac{W_{i1m} D_i}{\lambda_{1m}} \pi_{1m}(m, \mathbf{X}_i) \left(\mu_{1m}(m, \mathbf{X}_i) - \mu_{0m}(m, \mathbf{X}_i) - \gamma_m\right).
\end{aligned}
$$

From there, it is straightforward to verify that

$$
\frac{\partial \gamma_m(\theta_0)}{\partial \theta} = \mathbb{E}\left[\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma) S(\mathbf{O}_i; \theta_0)\right].
$$

Thus it is the efficient influence function for γ_m and the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_\gamma)^2]$.

$\square$

B.4 Asymptotic distribution of the cross-fitting estimator

We now derive the asymptotic distribution of the cross-fitting estimator. To do so, we define the statistical functional of interest as a function of the underlying probability distribution:

$$\tau_m(P) = \int_{\mathbf{x},\mathbf{z}} \int_{y_2,y_1} (y_2 - y_1) dP(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x}) dP(z, x \mid 1, m) \\ - \int_{\mathbf{x},\mathbf{z}} \int_{y_2,y_1} (y_2 - y_1) dP(y_2, y_1 \mid m, 0, m, \mathbf{z}, \mathbf{x}) dP(z, x \mid 0, m)$$

We denote $\mathbb{P}$ as the true distribution of the data so that

$$\tau_m(\mathbb{P}) = \mathbb{E}\left[\frac{W_{i1m} D_i}{\rho_m \delta} \mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i) - \frac{W_{i1m}(1 - D_i)}{\rho_m(1 - \delta)} \mu_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)\right] = \tau_m$$

as the true value of the parameter. Let $\widehat{\mathbb{P}}_n$ be any distribution on $\mathbf{O}$ such that the marginal distribution of $(\mathbf{X}_i, \mathbf{Z}_i, D_i, M_{i0})$ obtain their empirical distributions, but the nuisance functions are equal to their estimated value from some potentially data-adaptive procedure $(\widehat{\mu}_{dm}, \widehat{\pi}_{dm})$. Under this distribution, we have

$$\tau_m(\widehat{\mathbb{P}}_n) = \mathbb{P}_n\left[\frac{W_{i1m} D_i}{\widehat{\rho}_m \widehat{\delta}} \widehat{\mu}_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i) - \frac{W_{i1m}(1 - D_i)}{\widehat{\rho}_m(1 - \widehat{\delta})} \widehat{\mu}_{1m}(m, \mathbf{X}_i, \mathbf{Z}_i)\right]$$

as the plugin estimator based on outcome regression.

The following lemma is from the Supplemental Materials for Kennedy, Balakrishnan and G'Sell (2020) and follows from an application of Chebyshev's inequality. For any random function $\widehat{g}(\mathbf{O}_i)$, let $\mathbb{P}(\widehat{g}_i) = \mathbb{P}(\widehat{g}(\mathbf{O}_i)) = \int \widehat{g}(\mathbf{o}) d\mathbb{P}(\mathbf{o})$, which is equivalent to the expectation of $\widehat{g}$ according to the distribution $\mathbb{P}$.

Lemma SM.1. *Let $\widehat{f}(\mathbf{o})$ be a function estimated from a sample $\mathbf{O}_{-b} = \{\mathbf{O}_i : B_i \neq b\}$ and let $\mathbb{P}_n^b$ be the empirical measure over $\mathbf{O}_b = \{\mathbf{O}_i : B_i = b\}$, which is independent of $\mathbf{O}_{-b}$. Then,*

$$(\mathbb{P}_n^b - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\left\|\widehat{f} - f\right\|}{\sqrt{n}}\right)$$

Here we describe the regularity conditions that are required to prove Theorem 2.

Assumption 6 (Regularity conditions)**.** *We assume that (a) $\mathbb{P}[\epsilon_1 \leq \widehat{\pi}_{1m} \leq 1 - \epsilon_1] = 1$, $\mathbb{P}[\epsilon_d \leq \widehat{\pi}_d \leq 1 - \epsilon_d] = 1$, and $\mathbb{P}[\epsilon_2 \leq \widehat{\pi}_{i,2m} \leq 1 - \epsilon_2] = 1$ for some values of $\epsilon_1, \epsilon_d, \epsilon_2 > 0$; (b) $\|Y_{it}\|_q \leq C_y$, $\left\|\mu_{i,dm}\right\|_q \leq C_\mu$, and $\left\|\nu_{i,dm}\right\|_q \leq C_\nu$ for some fixed strictly positive constants C_y, C_μ, C_ν and $q > 2$.*

Proof of Theorem 2. We focus on the result for $\widehat{\tau}_m$ since the derivation for $\widehat{\gamma}_m$ follows similarly. To ease notation and without loss of generality, let $\rho_m = \widehat{\rho}_m = 1$ so there is only one baseline mediator value possible. Let $\widehat{\mathbb{P}}_{-b}$ be the empirical measure like $\widehat{\mathbb{P}}_n$ but with the nuisance functions estimated without fold b. For this proof, we write the EIFs as functions of a measure rather than the nuisance terms, so that $\widetilde{\psi}_m(\mathbf{O}_i;\mathbb{P})$ is the EIF under the true data generating distribution and $\widetilde{\psi}_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b})$ is the EIF when the nuisance functions are estimated without fold b. Define the estimated effect from fold b as

$$\widehat{\tau}_{m,b} = \mathbb{P}_n^b\left\{\psi_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b})\right\} = \tau_m(\widehat{\mathbb{P}}_{-k}) + \mathbb{P}_n^b\left\{\widetilde{\psi}_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b})\right\}$$

We can write the estimation error in fold b as

$$\begin{aligned}
\widehat{\tau}_{m,b} - \tau_m &= \tau_m(\widehat{\mathbb{P}}_{-k}) + \mathbb{P}_n^b\left\{\widetilde{\psi}_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b})\right\} - \tau_m(\mathbb{P}) \\
&= (\mathbb{P}_n^b - \mathbb{P})\left\{\widetilde{\psi}_m(\mathbf{O}_i;\mathbb{P})\right\} + (\mathbb{P}_n^b - \mathbb{P})\left\{\widetilde{\psi}_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b}) - \widetilde{\psi}_m(\mathbf{O}_i;\mathbb{P})\right\} + R_2(\widehat{\mathbb{P}}_{-b},\mathbb{P}), \\
&= S_b^* + T_{1b} + T_{2b},
\end{aligned}$$

where

$$R_2(\overline{P},P) = \tau_m(\overline{P}) - \tau_m(P) + \int \widetilde{\psi}_m(\mathbf{o};\overline{P})dP(\mathbf{o}).$$

For ease of exposition, We assume that we have equal-sized fold, though all results go through as long as the number of folks is finite. In this case, the overall estimate is $\widehat{\tau}_m = \frac{1}{B}\sum_{b=1}^B \widehat{\tau}_{m,b}$, so we can write the overall estimation error as

$$\widehat{\tau}_m - \tau_m = S^* + \frac{1}{B}\sum_{b=1}^B T_{1b} + \frac{1}{B}\sum_{b=1}^B T_{2b} = S^* + T_1 + T_2,$$

where $S^* = (\mathbb{P}_n - \mathbb{P})\,\widetilde{\psi}_{im}$.

We take each term in turn. First, S^* is the average of n iid mean-zero random variables with finite variance, so can employ the central limit theorem to establish that it will converge in distribution to $N(0,\mathbb{V}[\widetilde{\psi}_{im}])$. Note that $\mathbb{V}[\widetilde{\psi}_{im}] = \mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i;\mathbb{P})^2]$.

For T_1, we first note that $\widetilde{\psi}_m(\mathbf{O}_i;P) = \psi_m(\mathbf{O}_i;P) + \tau_m(P)$ and $(\mathbb{P}_n^b - \mathbb{P})(\tau_m(\widehat{\mathbb{P}}_{-b}) - \tau_m(\mathbb{P})) = 0$ because $\tau_m(\widehat{\mathbb{P}}_{-b})$ and $\tau_m(\mathbb{P})$ are constant with respect to those measures. We then write this empirical process term as

$$T_{1b} = (\mathbb{P}_n^b - \mathbb{P})\left\{\widetilde{\psi}_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b}) - \widetilde{\psi}_m(\mathbf{O}_i;\mathbb{P})\right\} = (\mathbb{P}_n^b - \mathbb{P})\left\{\psi_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b}) - \psi_m(\mathbf{O}_i;\mathbb{P})\right\},$$

By Lemma SM.1, to show that T_{2b} is $o_p(1/\sqrt{n})$, we have to show that $\left\|\psi_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b})-\psi_m(\mathbf{O}_i;\mathbb{P})\right\| = o_p(1)$. If this holds for all T_{1k}, then it also will hold for T_1. Then, omitting arguments for cleaner exposition, we have

$$
\begin{aligned}
&\|\psi_m(\mathbf{O}_i;\widehat{\mathbb{P}}_{-b})-\psi_m(\mathbf{O}_i;\mathbb{P})\| \\
&\quad= \left\| \frac{D_iW_{i2m}}{\widehat{\delta}\widehat{\pi}_{1m}}\left(\Delta Y_i-\widehat{\mu}_{1m}\right) - \frac{D_iW_{i1m}}{\delta\pi_{1m}}\left(\Delta Y_i-\mu_{i1m}\right) \right. \\
&\qquad + \frac{(1-D_i)W_{i2m}}{(1-\widehat{\delta})\widehat{\pi}_{0m}}\left(\Delta Y_i-\widehat{\mu}_{0m}\right) - \frac{(1-D_i)W_{i1m}}{(1-\delta)\pi_{1m}}\left(\Delta Y_i-\mu_{0m}\right) \\
&\qquad \left. + \frac{D_i}{\widehat{\delta}}\widehat{\mu}_{1m} - \frac{D_i}{\delta}\mu_{1m} - \frac{1-D_i}{1-\widehat{\delta}}\widehat{\mu}_{0m} + \frac{1-D_i}{1-\delta}\mu_{0m} \right\| \\
&\quad= \left\| \frac{D_iW_{i2m}}{\delta\pi_{1m}}(\mu_{1m}-\widehat{\mu}_{1m}) + \frac{D_iW_{i2m}}{\delta\widehat{\pi}_{1m}}\left(\Delta Y_i-\widehat{\mu}_{1m}\right)\left(\pi_{1m}-\widehat{\pi}_{1m}\right) + \frac{D_iW_{i1m}}{\delta\widehat{\delta}\widehat{\pi}_{1m}}\left(\Delta Y_i-\widehat{\mu}_{1m}\right)\left(\delta-\widehat{\delta}\right) \right. \\
&\qquad - \frac{(1-D_i)W_{i2m}}{(1-\delta)\pi_{0m}}(\mu_{0m}-\widehat{\mu}_{0m}) - \frac{(1-D_i)W_{i2m}}{(1-\delta)\widehat{\pi}_{0m}}\left(\Delta Y_i-\widehat{\mu}_{0m}\right)\left(\pi_{0m}-\widehat{\pi}_{0m}\right) \\
&\qquad - \frac{(1-D_i)W_{i1m}}{(1-\delta)(1-\widehat{\delta})\widehat{\pi}_{0m}}\left(\Delta Y_i-\widehat{\mu}_{0m}\right)\left(\widehat{\delta}-\delta\right) + \frac{D_i}{\delta}\left(\widehat{\mu}_{1m}-\mu_{1m}\right) - \frac{1-D_i}{1-\delta}\left(\widehat{\mu}_{0m}-\mu_{0m}\right) \\
&\qquad \left. + \frac{D_i}{\delta\widehat{\delta}}\widehat{\mu}_{1m}\left(\delta-\widehat{\delta}\right) - \frac{1-D_i}{(1-\delta)(1-\widehat{\delta})}\widehat{\mu}_{0m}\left(\delta-\widehat{\delta}\right) \right\|, \\
&\quad\lesssim \left\|\widehat{\delta}-\delta\right\| + \max_d \|\widehat{\pi}_{dm}-\pi_{dm}\| + \max_d \|\widehat{\mu}_{dm}-\mu_{dm}\| = o_{\mathbb{P}}(1)
\end{aligned}
$$

where the hats are estimated from non-b folds and recall that we write $a \lesssim b$ if $a \leq Cb$ for some positive constant $C > 0$. The second equality is just rearranging. The last inequality follows teh triangle inequality, the fact that the propensity scores (and their estimates) are bounded away from zero (per Assumption 6), and combination of the bounded moment conditions from Assumption 6 and Hölder's inequality. Here we have also used the fact that the estimated and true propensity scores are bounded away from zero. By Lemma SM.1, T_{1b} and thus T_1 must be $o_{\mathbb{P}}(1/\sqrt{N})$.

For T_2, we must show that $T_{2b} = R_2(\widehat{\mathbb{P}}_{-b},\mathbb{P}) = o_{\mathbb{P}}(1/\sqrt{n})$. First, note that $\tau_m(\mathbb{P}) = \nu_{1m}-\nu_{0m}$ and $\tau_m(\widehat{\mathbb{P}}_{-b}) = \widehat{\nu}_{1m}-\widehat{\nu}_{0m}$ where $\widehat{\nu}_{1m} = \mathbb{P}_{n,-b}\{(D_i/\widehat{\delta})\widehat{\mu}_{1m}\}$ is the average of the estimated outcome regression over the empirical distribution of the covariates conditional on $D_i = 1$ (and $M_{i0} = m$

though this is suppressed in the notation). We can write the remainder term as

$$
\begin{aligned}
R_2(\widehat{\mathbb{P}}_{-b}, \mathbb{P}) = \tau_m(\widehat{\mathbb{P}}_{-b}) - \tau_m(\mathbb{P}) \\
&+ \int \Bigg[\frac{D_i W_{i2m}}{\widehat{\delta}\widehat{\pi}_{1m}} (\Delta Y_i - \widehat{\mu}_{1m}) - \frac{(1-D_i)W_{i2m}}{(1-\widehat{\delta})\widehat{\pi}_{0m}} (\Delta Y_i - \widehat{\mu}_{0m}) \\
&\quad + \frac{D_i}{\widehat{\delta}}\widehat{\mu}_{1m} - \frac{1-D_i}{1-\widehat{\delta}}\widehat{\mu}_{0m} - \frac{D_i}{\widehat{\delta}}\widehat{\nu}_{1m} + \frac{1-D_i}{1-\widehat{\delta}}\widehat{\nu}_{0m} \Bigg] d\mathbb{P}.
\end{aligned}
$$

Again, the nuisance functions are estimated on the non-b fold, so they are constant with respect to the data and measure $\mathbb{P}$. By iterated expectations and the parallel trends assumptions, we have

$$
\begin{aligned}
\int \frac{D_i W_{i2m}}{\widehat{\delta}\widehat{\pi}_{1m}}(\Delta Y_i - \widehat{\mu}_{1m}) d\mathbb{P} &= \int \frac{\delta\pi_{1m}}{\widehat{\delta}\widehat{\pi}_{1m}}(\mu_{1m} - \widehat{\mu}_{1m}) d\mathbb{P} \\
\int \frac{D_i}{\widehat{\delta}}\widehat{\mu}_{1m} d\mathbb{P} &= \int \frac{\delta}{\widehat{\delta}}\widehat{\mu}_{1m} d\mathbb{P} \\
\int \frac{D_i}{\widehat{\delta}}\mu_{1m} d\mathbb{P} &= \int \frac{\delta}{\widehat{\delta}}\nu_{1m} d\mathbb{P}
\end{aligned}
$$

Using these fact, we can show

$$
\begin{aligned}
R_2(\widehat{\mathbb{P}}_{-b}, \mathbb{P}) &= (\widehat{\nu}_{1m} - \widehat{\nu}_{0m}) - (\nu_{1m} - \nu_{0m}) \\
&+ \int \Bigg[\frac{\delta\pi_{1m}}{\widehat{\delta}\widehat{\pi}_{1m}} (\mu_{1m} - \widehat{\mu}_{1m}) - \frac{(1-\delta)\pi_{0m}}{(1-\widehat{\delta})\widehat{\pi}_{0m}} (\mu_{0m} - \widehat{\mu}_{0m}) + \frac{D_i}{\widehat{\delta}}\widehat{\mu}_{1m} - \frac{1-D_i}{1-\widehat{\delta}}\widehat{\mu}_{0m} - \tau_m(\mathbb{P}) \Bigg] d\mathbb{P}. \\
&= (\widehat{\nu}_{1m} - \widehat{\nu}_{0m}) - (\nu_{1m} - \nu_{0m}) \\
&+ \int \Bigg[\frac{\delta\pi_{1m}}{\widehat{\delta}\widehat{\pi}_{1m}} (\mu_{1m} - \widehat{\mu}_{1m}) - \frac{(1-\delta)\pi_{0m}}{(1-\widehat{\delta})\widehat{\pi}_{0m}} (\mu_{0m} - \widehat{\mu}_{0m}) + \frac{D_i}{\widehat{\delta}}(\widehat{\mu}_{1m} - \mu_{1m}) - \frac{1-D_i}{1-\widehat{\delta}}(\widehat{\mu}_{0m} - \mu_{0m}) \\
&\quad + \frac{D_i}{\widehat{\delta}}\mu_{1m} - \frac{1-D_i}{1-\widehat{\delta}}\mu_{0m} - \frac{D_i}{\widehat{\delta}}\widehat{\nu}_{1m} + \frac{1-D_i}{1-\widehat{\delta}}\widehat{\nu}_{0m} \Bigg] d\mathbb{P}. \\
&= (\widehat{\nu}_{1m} - \widehat{\nu}_{0m}) - (\nu_{1m} - \nu_{0m}) \\
&+ \int \Bigg[\frac{\delta\pi_{1m}}{\widehat{\delta}\widehat{\pi}_{1m}} (\mu_{1m} - \widehat{\mu}_{1m}) - \frac{(1-\delta)\pi_{0m}}{(1-\widehat{\delta})\widehat{\pi}_{0m}} (\mu_{0m} - \widehat{\mu}_{0m}) + \frac{\delta}{\widehat{\delta}}(\widehat{\mu}_{1m} - \mu_{1m}) - \frac{1-\delta}{1-\widehat{\delta}}(\widehat{\mu}_{0m} - \mu_{0m}) \\
&\quad - \frac{D_i}{\widehat{\delta}}(\widehat{\nu}_{1m} - \nu_{1m}) + \frac{1-D_i}{1-\widehat{\delta}}(\widehat{\nu}_{0m} - \nu_{0m}) \Bigg] d\mathbb{P}. \\
&= \int \Bigg[\frac{\delta\pi_{1m}}{\widehat{\delta}\widehat{\pi}_{1m}}(\widehat{\mu}_{1m} - \mu_{1m})(\widehat{\pi}_{1m} - \pi_{1m}) - \frac{(1-\delta)\pi_{0m}}{(1-\widehat{\delta})\widehat{\pi}_{0m}}(\widehat{\mu}_{0m} - \mu_{0m})(\widehat{\pi}_{0m} - \pi_{0m}) \\
&\quad + \frac{D_i}{\widehat{\delta}}(\widehat{\nu}_{1m} - \nu_{1m})(\widehat{\delta} - \delta) - \frac{1-D_i}{1-\widehat{\delta}}(\widehat{\nu}_{0m} - \nu_{0m})(\widehat{\delta} - \delta) \Bigg] d\mathbb{P}.
\end{aligned}
$$

Recall that all estimated nuisance functions are bounded, which allows us to bound the remainder term as

$$\begin{aligned}|R_2(\widehat{\mathbb{P}}_{-b}, \mathbb{P})| \lesssim \int \Big[& |\widehat{\mu}_{1m} - \mu_{1m}||\widehat{\pi}_{1m} - \pi_{1m}| + |\widehat{\mu}_{0m} - \mu_{0m}||\widehat{\pi}_{0m} - \pi_{0m}| \\ & + |\widehat{\nu}_{1m} - \nu_{1m}||\widehat{\delta} - \delta| + |\widehat{\nu}_{0m} - \nu_{0m}||\widehat{\delta} - \delta| \Big] d\mathbb{P}. \\ & \leq \max_d \|\widehat{\mu}_{dm} - \mu_{dm}\| \, \|\widehat{\pi}_{dm} - \pi_{dm}\| + \max_d \|\widehat{\nu}_{dm} - \nu_{dm}\| \left\|\widehat{\delta} - \delta\right\|\end{aligned}$$

The second inequality follows from Cauchy-Schwarz. By assumption, the first term is $o_{\mathbb{P}}(1/\sqrt{n})$. Note that $\widehat{\nu}_{dm}$ is a sample average of $\widehat{\mu}_{dm}$ so it is consistent and thus has $\max_d \|\widehat{\nu}_{dm} - \nu_{dm}\| = o_{\mathbb{P}}(1)$. Given that $\left\|\widehat{\delta} - \delta\right\| = O_{\mathbb{P}}(1/\sqrt{n})$, then second term is $o_{\mathbb{P}}(1/\sqrt{n})$. This implies that T_{2b} and thus T_2 is also $o_{\mathbb{P}}(1/\sqrt{n})$, and we have,

$$\sqrt{n}(\widehat{\tau}_m - \tau_m) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\psi}_{im} + o_{\mathbb{P}}(1),$$

and combined with the CLT results about (I), the desired result obtains.

□

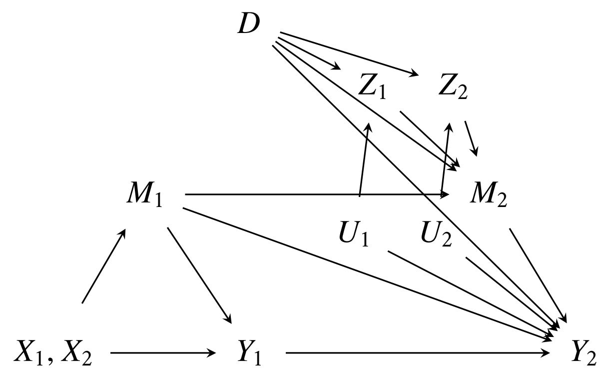

Figure SM.6: Directed acyclic graph showing the simulation setup.

C Simulation Results

We now evaluate the finite-sample performance of our estimator with a simulation experiment. Specifically, we are interested in how the doubly robust estimator compares to both traditional difference-in-differences and non-doubly robust CDE methods as well as how different machine learning techniques in the doubly robust approach can handle misspecification. We evaluate the performance of our estimator against two alternative approaches for computing direct effects—traditional regression DID controlling for baseline covariates $\mathbf{X}_i$ and the mediator, and the same specification also controlling for intermediate covariates $\mathbf{Z}_i$—and against plug-in estimators based on the IPW and outcome regression approaches from Proposition 1 in the text. As the results show, our method performs well against these alternatives even when the working models are misspecified, particularly at larger sample sizes.

The DGP follows the DAG in Figure SM.6. Treatment has independent probability $p_d = 0.5$, and we generate two observed baseline variables, $\mathbf{X}_i = (X_{i1}, X_{i2})' \sim \mathcal{N}_2(0, \sigma_x^2 \mathbf{I}_2)$, where $\sigma_x^2 = 0.01$, and two unobserved independent baseline variables $U_{i1}, U_{i2} \sim \mathcal{N}(0, 0.01)$. We draw the baseline mediator as $M_{i1} = \mathbb{I}(X_{i1}+X_{i2}+\varepsilon_{im1} \geq 0)$. The baseline outcome follows $Y_{i1} = 1+0.4M_{i1}+\mathbf{X}_i'\boldsymbol{\beta}+\varepsilon_{iy1}$, where $\boldsymbol{\beta} = (0.5, 0.5)'$, and then we generate the intermediate confounders with heterogeneous treatment effects, $Z_{ij} = \delta_{ij} D_i + 5U_{ij} + \varepsilon_{izj}$, where $\delta_{ij} \sim \mathcal{N}(0.25, 0.0025)$ for $j \in \{1, 2\}$. The posttreatment mediator follows $M_{i2} = \mathbb{I}(-1 + 1.5D_i + 0.4M_{i1} + \mathbf{Z}_i'\boldsymbol{\gamma} + \varepsilon_{im2} \geq 0)$, where $\boldsymbol{\gamma} = (0.75, 0.75)'$. The second-period outcome is

$$Y_{i2} = Y_{i1} + 0.4M_{i1} + 0.2D_i + 0.3M_{i2} + 5U_{i1} + 5U_{i2} + \varepsilon_{iy2},$$

where $(\varepsilon_{im1}, \varepsilon_{iy1}, \varepsilon_{iz1}, \varepsilon_{iz2}, \varepsilon_{im2}, \varepsilon_{iy2}) \sim \mathcal{N}_6(0, \Sigma_\varepsilon)$ and Σ_ε is a diagonal matrix with $\text{diag}(\Sigma_\varepsilon) = (0.01, 0.01, 0.04, 0.04, 1, 0.01)'$. In order to test how these methods perform when the relevant models are misspecified, we also construct transformations of the covariates $X_{i1}, X_{i2}, Z_{i1}, Z_{i2}$ as follows, employing a similar setup to Kang and Schafer (2007):

$$X_{i1}^* = (\exp(X_{i1}/2) - 1)^2, \quad X_{i2}^* = X_{i2}/(1 + \exp(X_{i1})) + 10,$$

$$Z_{i1}^* = (X_{i1}Z_{i1}/25 + 0.6)^3, \quad Z_{i2}^* = (X_{i2} + Z_{i2} + 20)^2.$$

For each simulated dataset, we construct seven estimates for the marginalized ACDE-BC, τ. First, we simply regress $\Delta Y_i = Y_{i2} - Y_{i1}$ on D_i, controlling for M_{i1}, M_{i2}, X_{i1} and X_{i2} ("DID + Mediator"). Second, we add the intermediate covariates to this specification ("DID + Mediator + Covariates"). Third, we use our doubly robust ACDE estimator with the same outcome regression as the DID + Mediator + Covariates estimator and three different propensity score estimators for M_{i2}: logistic regression ("DR ACDE (Logit)"), the Lasso ("DR ACDE (Lasso)"), and random forests ("DR ACDE (RF)"). For the Lasso approach, we include all squared terms and two-way interactions of the covariates in both the propensity score and outcome regression models. Finally, we create separate IPW ("Inverse Propensity Weighting") and outcome regression ("Outcome Regressions") estimators using the propensity score and outcome regression models from the DR ACDE logit approach and plugging our estimates into sample versions of equations 2 and 3 in Proposition 1 respectively (estimating standard errors using a bootstrap with 500 samples).

We ran 1000 replications of this DGP and computed the average bias, the root mean square error (RMSE), and the coverage of nominal 95% confidence intervals for sample sizes of 250, 500, and 1000 for four scenarios: using the "correctly specified" covariates $(X_{i1}, X_{i2}, Z_{i1}, Z_{i2})$, using the "incorrectly specified" transformed versions $(X_{i1}^*, X_{i2}^*, Z_{i1}^*, Z_{i2}^*)$, and using the misspecified versions either in any propensity score models or in any outcome regression models, but not both. For each iteration of the Monte Carlo simulation, we also calculated the true values of τ_m and τ.

Figure SM.7 presents the results of this simulation. Under both correctly and incorrectly specified models, we can see that the DID estimators exhibit large biases at all sample sizes and have correspondingly high RMSEs and low coverage. As expected, this performance comes from confounding bias when excluding the intermediate covariates and posttreatment bias when including these covariates. Under this DGP, these biases can be made larger or smaller by manipulating the strength of the relationships on those paths.

When the relevant models are correctly specified, the IPW and outcome regression methods have very little bias and relatively accurate coverage (especially the later, which also has very low RMSE in this situation). When they are misspecified, however, the performance of these approaches becomes much worse, although still not as bad as the "naive" DID estimators.

Relative to these methods, the performance of our DR ACDE estimator is more consistent across the correct and incorrect specifications, although it varies based on the estimation engine employed. When using the correctly specified variables, all of the multiply robust methods are similar in having low bias, low RMSE, and close-to-correct coverage, particularly at higher sample sizes. This largely continues to be the case when either the propensity score model or the outcome regressions are based on the misspecified variables, as would be expected given double robustness, although the bias is slightly greater for the random forest. Conversely, when both models are incorrectly specified, all three DR approaches have higher bias, but the increase is more muted for the random forest than the others. The Lasso, which uses a more extensive set of basis functions, and the random forest, which allows data-driven estimation of interactive relationships between variables, both outperform the logit in this situation, showing that flexible approaches can reduct bias even when hampered by misspecified covariates.

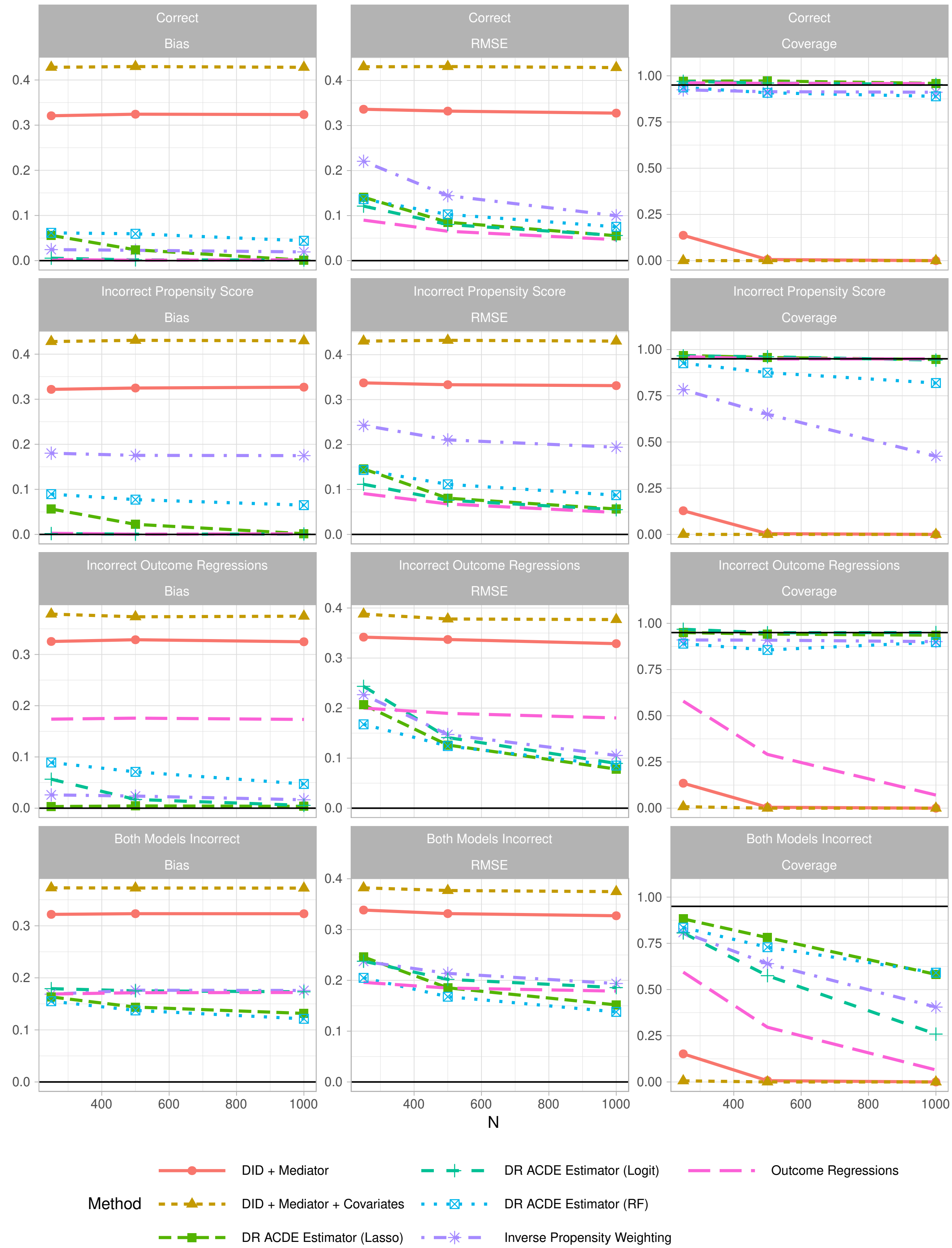

Figure SM.7: Performance of our doubly robust estimator compared with difference-in-differences controlling for the mediator and baseline covariates; difference-in-differences controlling for the mediator, baseline covariates, and intermediate covariates; inverse propensity weighting; and outcome regressions.

D Additional Tables for Empirical Application

Table SM.3: List of covariates

Pre-treatment covariates	Post-treatment covariates
Trans law support	Obama feeling thermometer (Δ)
Registered Democrat	Trans tolerance (Δ)
Political ideology	Gender norms (Δ)
Religiousity	Trans law support (Δ)
Knows trans people	
Female	
Hispanic	
Af.-Am.	
Age	
Survey in Spanish	
Transgender tolerance	
Gender norms	
Obama feeling thermometer	

Table SM.4: ACDE-BC estimates by subgroup

Subgroup	Baseline Mediator	ACDE-BC (s.e.)	n (ACDE)	ATT (s.e.)	n (ATT)
All	Marginal	0.401 (0.261)	369	0.296 (0.137)	369
All	m = Cool	-0.132 (0.284)	72	-0.198 (0.328)	94
All	m = Neutral	0.747 (0.260)	138	0.540 (0.219)	168
All	m = Warm	0.296 (0.261)	159	0.254 (0.170)	107
Non-white	Marginal	0.328 (0.323)	261	0.203 (0.156)	261
Non-white	m = Cool	0.127 (0.331)	53	-0.206 (0.348)	74
Non-white	m = Neutral	0.500 (0.326)	97	0.348 (0.255)	118
Non-white	m = Warm	0.226 (0.321)	111	0.366 (0.214)	69
White	Marginal	0.616 (0.562)	108	0.474 (0.304)	108
White	m = Cool	n.a.	19	n.a.	20
White	m = Neutral	1.326 (0.553)	41	0.866 (0.474)	50
White	m = Warm	0.371 (0.540)	48	0.207 (0.323)	38
Woman	Marginal	0.475 (0.345)	210	0.349 (0.186)	210
Woman	m = Cool	0.070 (0.376)	40	0.032 (0.525)	49
Woman	m = Neutral	0.977 (0.344)	76	0.795 (0.305)	98
Woman	m = Warm	0.031 (0.342)	94	0.094 (0.223)	63
Non-woman	Marginal	0.194 (0.446)	159	0.136 (0.215)	159
Non-woman	m = Cool	-0.943 (0.514)	32	-0.094 (0.546)	45
Non-woman	m = Neutral	0.545 (0.449)	62	0.098 (0.342)	70
Non-woman	m = Warm	0.738 (0.446)	65	0.471 (0.276)	44

Table SM.6: ACDE-BC estimates mediator and outcome timing

Outcome	Mediator	Estimate	Std. Error	n
t = 3	t = 2	0.747	0.260	138
t = 4	t = 2	0.644	0.246	130
t = 4	t = 3	0.527	0.230	142

Table SM.8: ACDE-BC estimates with alternative discretization of the mediator

Baseline Mediator	ACDE-BC (s.e.)	n (ACDE)	ATT (s.e.)	n (ATT)
Marginal	0.457 (0.635)	369	0.292 (0.137)	369
m = 0	0.024 (0.930)	23	0.258 (0.551)	37
0 < m < 50	-0.847 (0.412)	49	-0.331 (0.484)	57
m = 50	0.752 (0.627)	138	0.540 (0.219)	168
50 < m < 100	0.287 (0.838)	124	0.242 (0.188)	77
m = 100	2.180 (0.600)	35	-0.429 (0.434)	30