# Selection and parallel trends*

Dalia Ghanem† Pedro H. C. Sant'Anna‡ Kaspar Wüthrich§

First draft on arXiv: March 17, 2022. This draft: February 1, 2026

**Abstract**

We study the role of selection into treatment in difference-in-differences (DiD) designs. We derive necessary and sufficient conditions for parallel trends assumptions under general classes of selection mechanisms. These conditions characterize the empirical content of parallel trends and clarify the trade-offs between assumptions about selection into treatment and restrictions on the time series properties of the potential outcomes required for DiD methods. We use the necessary and sufficient conditions to provide a selection-based decomposition of the bias of DiD and provide easy-to-implement strategies for benchmarking its components. We also provide templates for justifying DiD in applications with and without covariates. Reanalyses of the causal effect of NSW training programs and the effect of the Medicaid expansion demonstrate the usefulness of our selection-based approach to benchmarking the bias of DiD.

**Keywords:** causal inference, conditional parallel trends, covariates, difference-in-differences, selection mechanism, time-invariant and time-varying unobservables, treatment effects

**JEL Codes:** C21, C23

*We are grateful to Isaiah Andrews, Manuel Arellano, Dmitry Arkhangelsky, Stéphane Bonhomme, Irene Botosaru, Christoph Breunig, Federico Bugni, Brantly Callaway, Ivan Canay, Clément de Chaisemartin, Gordon Dahl, Aureo de Paula, Graham Elliott, Joachim Freyberger, Bulat Gafarov, Bryan Graham, Lena Janys, Stefan Hoderlein, Christian Hansen, Keisuke Hirano, Peter Hull, Guido Imbens, Désiré Kédagni, Pat Kline, Nikolay Kudrin, Matt Masten, Eric Mbakop, David McKenzie, Eduardo Morales, Mikkel Plagborg-Møller, Vitor Possebom, Niklas Potrafke, Demian Pouzo, Jonathan Roth, Aleksey Tetenov, Andres Santos, Yuya Sasaki, Vira Semenova, Xiaoxia Shi, Valentin Verdier, Chris Walters, and many seminar and conference participants for comments. We used Grammarly for grammar checking, Github Co-Pilot for coding assistance, and `refine.ink` for a final proofreading check on January 18, 2026. The usual disclaimer applies.

†Department of Agricultural & Resource Economics, University of California, Davis. One Shields Ave, Davis CA 95616; dghanem@ucdavis.edu

‡Department of Economics, Emory University, 1602 Fishburne Dr, Atlanta, GA 30322; pedro.santanna@emory.edu

§Department of Economics, University of Michigan, 238 Lorch Hall, 611 Tappan Ave., Ann Arbor, MI 48109-1220, CESifo; kasparwu@umich.edu

> *. . . while the new papers [in the DiD literature] clarify very well the statistical assumptions needed for estimation, effective use of these methods also requires being able to understand what the threats to these assumptions are in different contexts, and to make a plausible rhetorical argument as to why we should think the assumptions hold.*
> — David McKenzie, *World Bank Development Impact Blog* (McKenzie, 2022)

# 1 Introduction

This paper provides a new perspective on difference-in-differences (DiD) identification through the lens of how units select into treatment. Parallel trends, the identifying assumption underlying DiD, requires the change in the expected (untreated) potential outcome over time to be the same in the treatment and control group. It is thus inherently a joint restriction on selection into treatment and the time series properties of the untreated potential outcome. Whether and to what extent there is a trade-off between these two types of restrictions is not well-understood, however. In particular, there is no general framework that characterizes the implications of parallel trends under selection-based restrictions. Such a framework is crucial for researchers to exploit contextual and economic information about how units select into treatment, which is available in many applications. It would enable researchers — in the words of David McKenzie (McKenzie, 2022) — "to understand what the threats to these [parallel trends] assumptions are in different contexts, and to make a plausible rhetorical argument as to why we should think the assumptions hold." By providing researchers with a selection-based framework for exploiting contextual and economic information to assess the plausibility of parallel trends, we complement the existing statistical and visual plausibility checks in DiD analyses.

Our goal is to provide a nonparametric selection-based characterization of the parallel trends assumption. To achieve this goal, we start by providing a general necessary and sufficient condition for parallel trends to hold for all selection mechanisms in a class defined by *the unobservables that determine selection*, which we denote by $\omega_i$.[1] This general result allows us to provide necessary and sufficient conditions for many empirically relevant classes of selection mechanisms.

The focus on classes of selection mechanisms defined by $\omega_i$ is motivated by the reduced-form nature of DiD methods and the myriad of empirical contexts analyzed using these methods. While contextual and economic information about selection may allow researchers to posit what determines selection into treatment, it is generally difficult to specify a particular functional form for the selection mechanism. For instance, if we consider settings where

[1] From a theoretical perspective, this necessary and sufficient condition is the condition for nonparametric identification using DiD for a given class of selection mechanisms.

selection into treatment is at the aggregate level, such as at the county or state level as in the Medicaid expansion (e.g., Miller et al., 2021; Baker et al., 2026), specifying a selection mechanism might be intractable or prohibitively complex.[2] The advantage of our necessary and sufficient conditions is that researchers can use them to assess and justify parallel trends without needing to defend a particular choice of selection mechanism.[3]

Before considering various restricted classes of selection mechanisms, we take a step back and ask: What are the necessary and sufficient conditions for parallel trends if we do not impose any restrictions on selection? The first corollary of our general necessary and sufficient condition answers this question with a "no-free-lunch" result. When researchers are not willing to restrict the selection mechanism at all, $\omega_i$ can include time-invariant and time-varying determinants of the untreated potential outcomes in the pre- and post-treatment period, $Y_{i1}(0)$ and $Y_{i2}(0)$. We show that parallel trends holds for all selection mechanisms in this class if and only if the untreated potential outcome is constant across time up to deterministic mean shifts, $Y_{i2}(0) - Y_{i1}(0) = E[Y_{i2}(0) - Y_{i1}(0)]$. This result shows that if one is not willing to restrict selection into treatment, then one needs to essentially rule out time-varying unobservables.

This "no-free-lunch" result motivates considering restricted classes of selection mechanisms, such as selection on treatment effects ("Roy-style selection"), selection on pre-treatment unobservables ("imperfect foresight"), selection on lagged outcomes, and selection on time-invariant unobservables ("selection on fixed effects"). The necessary and sufficient conditions for these classes characterize the trade-offs between restrictions on selection and restrictions on the time series properties of the untreated potential outcome and its unobservable determinants. The more restrictions researchers are willing to impose on how units select into treatment, the weaker the time series restrictions necessary and sufficient for parallel trends to hold. Conversely, in settings where the available economic and contextual knowledge is not sufficient to justify strong restrictions on selection, researchers need to impose restrictive assumptions on the time series properties of the potential outcomes to justify parallel trends.

Consider, for example, a setting with Roy-style selection where the units select into treatment if their treatment effect exceed the costs. If the units know their treatment effect and costs, then the necessary and sufficient condition for parallel trends requires the difference between the potential outcomes over time, $Y_{i2}(0) - Y_{i1}(0)$, to be mean independent of those

[2]This echoes the point made by Abadie (2021, p.404) about the difficulty of specifying selection mechanisms in synthetic control analyses of comparative case studies.

[3]If researchers have sufficient contextual and economic information to fully specify the selection mechanism, then we recommend they rely on identification strategies that directly exploit such information rather than DiD.

determinants of selection. Alternatively, suppose that the units select into treatment based on pre-treatment unobservables. In this case, the necessary and sufficient condition for parallel trends is a martingale-type condition on $Y_{it}(0) - E[Y_{it}(0)]$. Finally, in the context of selection on time-invariant unobservables, parallel trends is equivalent to a time homogeneity condition on the expectation of $Y_{it}(0) - E[Y_{it}(0)]$ conditional on the time-invariant unobservables. The advantage of our general necessary and sufficient condition is that researchers can use it to derive equivalent conditions for parallel trends for other empirically relevant settings.

The necessary and sufficient conditions we provide can serve as theory-based templates allowing researchers to assess and justify parallel trends in empirical applications based on contextual and economic information about how units select into treatment. To illustrate, we specialize our general results to the standard two-way fixed effects model for $Y_{it}(0)$. The resulting conditions explicitly allow for selection on time-invariant and time-varying unobservables, thus formalizing what "quasi-random" assignment means in the context of DiD analyses.

An appealing feature of our selection-based approach is that our necessary and sufficient conditions for parallel trends can help researchers better understand the bias of DiD. We provide a selection-based approach for benchmarking the bias of DiD when the validity of these necessary and sufficient conditions is questionable. To illustrate, we consider the leading case where selection on pre-treatment unobservables is a concern. We decompose the bias of DiD into two terms. The first component captures the bias resulting from selection on post-treatment unobservables. The second component captures the bias due to deviations from the martingale condition, which is necessary and sufficient for parallel trends under imperfect foresight. Under a linear relaxation of the martingale condition, the second bias component is equal to the product of the martingale deviation and the pre-treatment difference between the treatment and control group. From a theoretical perspective, our analysis allows us to better understand the bias of DiD and the extent to which pre-treatment information is useful for assessing the magnitude of this bias. For practitioners, we provide simple approaches to benchmark and sign both bias components, allowing them to assess the robustness of DiD in empirical applications.

We illustrate the usefulness of the selection-based approach for benchmarking the bias components of DiD based on two empirical applications. First, we consider the estimation of the causal effect of the NSW training programs. This application is well-suited for our purposes because there is an experimental benchmark allowing us to estimate the bias of DiD. Without covariates, the bias of DiD relative to the experimental benchmark is large and significant. The proposed benchmarking strategy yields the same sign of the bias as the

experimental estimate. The decomposition further demonstrates that the bias of DiD is very sensitive to violations of the martingale property because there is a large pre-treatment difference between the treatment and control group. Incorporating covariates into the analysis reduces the estimated bias relative to the experimental benchmark and also renders DiD more robust by reducing the pre-treatment difference between the treatment and control group. Second, we revisit the DiD analysis of the causal effect of the Medicaid expansion. The proposed benchmarking strategy suggests that the sign of the bias depends on the degree of selection on post-treatment unobservables. As in the NSW application, DiD without covariates is sensitive to violations of the martingale condition. Controlling for covariates almost fully removes pre-treatment differences between the treated and control states, rendering DiD robust to violations of the martingale condition.

### 1.1 Related literature

This paper contributes to several branches of the literature on causal inference using panel data. First, we contribute to the classical literature on canonical DiD setups by providing a novel selection-based perspective on the parallel trends assumptions underlying this literature.[4]

Our second contribution is to the more recent literature on DiD methods. See, e.g., de Chaisemartin and D'Haultfœuille (2023) and Roth et al. (2023) for surveys. Within this strand of the literature, our paper is most closely related to Roth and Sant'Anna (2023), Arkhangelsky et al. (2021), and Arkhangelsky and Imbens (2022), though our focus greatly differs from theirs. Roth and Sant'Anna (2023) discuss necessary and sufficient conditions under which parallel trends holds for all (monotonic) transformations of the untreated potential outcome. We, on the other hand, take the outcome model (and thus the specific transformation) as given and study the connection between parallel trends and selection into treatment. Arkhangelsky et al. (2021) and Arkhangelsky and Imbens (2022) propose doubly robust estimation methods that leverage restrictions on outcome models and/or selection models with unconfoundedness-type restrictions; see also Athey et al. (2021). Our results complement theirs as we maintain the parallel trends assumption and discuss the types of restrictions on selection compatible with it. Moreover, our analysis shows that parallel trends is compatible with various types of selection on unobservables, unlike standard unconfoundedness assumptions (e.g., Imbens, 2004; Imbens and Wooldridge, 2009).

Our third contribution is to the literature on sensitivity analysis, partial identification,

[4]See, e.g., Ashenfelter (1978), Ashenfelter and Card (1985), Heckman and Robb (1985), Card (1990), Card and Krueger (1994), Meyer et al. (1995), and Angrist and Krueger (1999) for early developments in the DiD literature, and Lechner (2010, Section 2) for a historical perspective.

and robust inference under violations of parallel trends. Existing work has proposed different ways to use pre-treatment information to bound the ATT (e.g., Manski and Pepper, 2018; Rambachan and Roth, 2023; Ban and Kédagni, 2023). We complement this work by providing a selection-based decomposition of the bias of DiD and simple empirical strategies for benchmarking its components. Our approach uses pre-treatment periods to learn about the deviations from the selection-based necessary and sufficient conditions for parallel trends, whereas existing approaches rely on more "reduced-form" quantities, such as pre-treatment parallel trends violations. Our selection-based approach also differs from the analysis by Marx et al. (2024). They derive partial identification results under monotone treatment selection assumptions on the untreated potential outcome, which they motivate using an economic model of learning with binary outcomes. By contrast, we characterize the bias of DiD in terms of deviations from the necessary and sufficient conditions for parallel trends.

Our fourth contribution is to the literature imposing explicit selection and/or outcome models to develop and compare different methods for estimating treatment effects, including DiD.[5] We contribute to this literature by providing general necessary and sufficient conditions for parallel trends, which are derived for general selection and outcome models that nest models considered in this literature. Our conditions thus clarify trade-offs between assumptions on selection and time-varying unobservables that are relevant for those models. Within this strand of the literature, our paper is most closely related to contemporaneous work by Marx et al. (2024), though our focus markedly differs from theirs. Marx et al. (2024) analyze parallel trends through the lens of various examples of dynamic choice models. In doing so, they focus on explicit models of selection, and many of their examples are for binary outcomes. By contrast, we provide general necessary and sufficient conditions for parallel trends without imposing restrictions on the nature of the outcome variables or explicit models of dynamic choice. The general classes of selection mechanisms we consider nest, and can be motivated by, dynamic choice models. Unlike Marx et al. (2024), we also incorporate covariates into our analysis and study settings with multiple groups and periods.

Finally, a byproduct of our analysis is an explicit connection between DiD and the literature on nonseparable panel models. In Appendix E, we show that our sufficient conditions for parallel trends imply combinations of identifying assumptions in this literature (e.g. Altonji and Matzkin, 2005; Bester and Hansen, 2009; Hoderlein and White, 2012; Chernozhukov et al., 2013).

[5] See, e.g., Ashenfelter and Card (1985), Heckman and Robb (1985), Card and Hyslop (2005), Chabé-Ferret (2015), Blundell and Dias (2009), de Chaisemartin and D'Haultfœuille (2018), Verdier (2020), de Chaisemartin and D'Haultfœuille (2022), Marx et al. (2024), Chabé-Ferret (2025). Most papers in this literature examine sharp DiD designs, as we do, de Chaisemartin and D'Haultfœuille (2018) and Marx et al. (2024) also consider fuzzy DiD designs.

### 1.2 Notation

For a random vector $W_{it}$, where $i = 1, \ldots, n$ and $t = 1, 2$, we denote its time series by $W_i \equiv (W_{i1}, W_{i2})$.[6] We use $F_W$ to denote the distribution of the random vector $W$ and $\mathcal{W}$ to denote its support. Let $f(z, w)$ be a function defined on $\mathcal{Z} \times \mathcal{W}$. We say that $f(z, w)$ is a trivial function of $w$ if $f(z, w) = f(z, w') = h(z)$ for all $z \in \mathcal{Z}$, $w \neq w'$, and $(w, w') \in \mathcal{W}^2$. We say that $f(z, w)$ is a symmetric function in $z$ and $w$ if $f(z, w) = f(w, z)$ for all $(z, w) \in \mathcal{Z} \times \mathcal{W}$. We use the notation $\overset{d}{=}$ to denote equality of distribution. For random variables, $X_i$, $Z_i$, and $W_i$, $Z_i|W_i, X_i \overset{d}{=} Z_i|X_i, W_i$ denotes that $F_{Z_i|W_i,X_i}(z|w, x) = F_{Z_i|X_i,W_i}(z|w, x)$ for $(z, w, x) \in \mathcal{Z} \times \mathcal{W} \times \mathcal{X}$.

## 2 Setup, selection mechanism, and examples

In the main text, we consider the classical DiD setup with two groups and two periods, where the selection decision is made at the same level as the unit of observation. We extend our results to settings with disaggregate data (e.g., on individuals) where the selection decision is made at a more aggregate level (e.g., at the state level) in Appendix A as well as to DiD designs with multiple groups and multiple periods in Appendix B.

Let $D_{it}$ and $Y_{it}$ denote the treatment status and outcome for unit $i \in \{1, \ldots, n\}$ in period $t \in \{1, 2\}$. Here the index $i$ refers to the unit making the decision to select into treatment. This could be an individual or a more aggregate administrative unit, such as a county or state. The treatment group ($G_i = 1$) selects the treatment path $D_i = (0, 1)$; the control group ($G_i = 0$) selects $D_i = (0, 0)$. The potential outcomes with and without the treatment are $Y_{it}(1)$ and $Y_{it}(0)$, respectively.[7] We abstract from covariates for now to focus on the issues arising from selection on time-invariant and time-varying unobservables. We discuss the additional implications of including covariates in Appendix D.

We consider the standard parallel trends assumption. Throughout the paper, we assume that all relevant moments exist and $\{Y_{i1}(0), Y_{i2}(0), G_i\}$ is i.i.d. across $i$.

**Assumption PT.** *The (unconditional) parallel trends assumption holds:*

$$E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] = E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0].$$

Under Assumption PT, the average treatment effect on the treated group in period $t = 2$,

[6] We define all vectors in this paper as row vectors.

[7] To focus attention on the role of the parallel trends assumption, we assume that there are no anticipatory effects. This is a standard assumption in the DiD literature. See, for example, Roth et al. (2023) for a discussion.

$\text{ATT} \equiv E[Y_{i2}(1) - Y_{i2}(0)|G_i = 1]$, is identified from the "difference-in-differences" as follows:

$$\text{ATT} = E[Y_{i2} - Y_{i1}|G_i = 1] - E[Y_{i2} - Y_{i1}|G_i = 0] \equiv \text{DiD}\,.$$

We work with a general nonseparable model for $Y_{it}(0)$,

$$Y_{it}(0) = \xi_t(\alpha_i, \varepsilon_{it}), \quad i = 1, \ldots, n, \quad t = 1, 2, \tag{1}$$

where $\alpha_i$, $\varepsilon_{i1}$, and $\varepsilon_{i2}$ are finite-dimensional vector-valued random variables, and $\xi_t(\cdot)$ is an unrestricted time-varying function. The outcome model (1), while not imposing any restrictions on $Y_{it}(0)$, allows us to distinguish between time-invariant and time-varying unobservables. This is necessary to define selection mechanisms that can directly depend on these unobservables. If, instead, we were to work directly with potential outcomes, this would rule out important examples of selection mechanisms, such as selection on fixed effects (e.g., Ashenfelter and Card, 1985).

The determinants of selection into treatment vary widely across DiD applications. In some applications, the unit $i$ making the selection decision is a state or county, whereas in other settings it is an individual economic agent, such as an individual, a household, or a firm. We therefore introduce a general selection mechanism that accommodates many different types of selection. To motivate this general selection mechanism, it is helpful to consider examples of selection mechanisms relevant for our setting.

**Example 2.1** (Selection on untreated outcomes)**.** *Selection on untreated potential outcomes goes back to at least the seminal work of Ashenfelter and Card (1985) in the context of individuals selecting into job training programs. It remains relevant in DiD applications where selection decisions are made by more aggregate units, which are prevalent in contemporary empirical research. For example, in a survey of governors on the reasons for expanding Medicaid, Sommers and Epstein (2013) find that the health outcomes in their states were one of the factors affecting their support for Medicaid expansion.*

*Ashenfelter and Card (1985, p.651) studied the case where individuals select into the training programs if their pre-treatment earnings $Y_{i1}(0)$ fall below a fixed threshold, so that $G_i = 1\{Y_{i1}(0) \leq c\}$.*[8] *In the context of Medicaid expansion, policymakers in a state might expand Medicaid if the health outcomes of their constituents (before the expansion) fall below a certain threshold.*

*More generally, let $\omega_i$ denote the information set available to the units when deciding*

[8] Ashenfelter and Card (1985) consider a more general setting with multiple pre-treatment periods, where selection depends on earnings in the $k$th pre-treatment period.

*whether to select into the treatment and consider the following mechanism,*

$$G_i = 1\left\{E[Y_{i1}(0) + \beta Y_{i2}(0)|\omega_i] \leq E[\kappa_{i2}|\omega_i]\right\}, \tag{2}$$

*where $\beta \in [0,1]$ is a discount factor and $\kappa_{i2}$ is a unit-specific random threshold. This example demonstrates the importance of allowing for additional unobservables in addition to the determinants of $Y_{it}(0)$, $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$.* □

**Example 2.2** (Selection on treatment effects (Roy-style selection))**.** *Let $\tau_{i2}$ denote the treatment effect in period $t = 2$, $\tau_{i2} = Y_{i2}(1) - Y_{i2}(0)$. Suppose that units select into the treatment if the expected gains from treatment given the information set $\omega_i$, $E[\tau_{i2}|\omega_i]$, exceed the expected cost of treatment, $E[\kappa_{i2}|\omega_i]$, $G_i = 1\{E[\tau_{i2}|\omega_i] \geq E[\kappa_{i2}|\omega_i]\}$. In the context of the job training setting, units might decide to participate in the program if the expected gain in their earnings outweighs the expected costs, whereas in the Medicaid setting, policymakers might support a Medicaid expansion if the expected improvements in the health outcomes of their constituents outweigh the expected costs. Indeed, these expected improvements were among the factors mentioned in the survey in Sommers and Epstein (2013).* □

**Example 2.3** (Selection on fixed effects)**.** *DiD methods have traditionally been motivated using two-way fixed effects models. Fixed effects assumptions allow for unrestricted dependence between time-invariant unobservables and the regressors, thereby implicitly allowing for selection on time-invariant unobservables.*[9] *A simple example is $G_i = 1\{\alpha_i \leq c\}$, where $\alpha_i$ is a scalar, which corresponds to the selection mechanism on p.650 in Ashenfelter and Card (1985). In the training program example, this mechanism captures settings where individuals select into the program if the permanent earnings component $\alpha_i$ falls below a threshold $c$, which is "based on potential trainees' discount rates, time horizons, and tastes for training" (Ashenfelter and Card, 1985, p.651). In the context of aggregate treatments, such as Medicaid expansion, policymakers might decide to expand Medicaid based on their political affiliations or state-specific characteristics, which are plausibly time-invariant.* □

In the previous examples, all agents make their selection decision in the same way and based on the same information set $\omega_i$. This is likely an oversimplification in many contexts where DiD is used, especially in aggregate selection settings. In the following example, we consider a setting with heterogeneous units whose selection decisions depend on different unobservables.

[9] See, e.g., Chamberlain (1984); Arellano (2003); Evdokimov (2010); Wooldridge (2010); Hoderlein and White (2012); Chernozhukov et al. (2013).

**Example 2.4** (Selection with heterogeneous units)**.** *For simplicity, we consider a setting with two types of units, noting that the example can be easily generalized to settings with more than two types. Let $\mu_i \in \{0,1\}$ be an indicator for the unit's type. If $\mu_i = 1$, unit $i$ adopts the treatment if the expected benefits outweigh the expected costs given $\omega_i^1$; if $\mu_i = 0$, unit $i$ selects into treatment if the expected discounted sum of untreated outcomes given $\omega_i^0$ falls below a certain threshold, so that*

$$G_i = 1\{E[Y_{i2}(1) - Y_{i2}(0)|\omega_i^1] \geq E[\kappa_{i2}|\omega_i^1]\}^{\mu_i} 1\{E[Y_{i1}(0) + \beta Y_{i2}(0)|\omega_i^0] \leq E[\kappa_{i2}|\omega_i^0]\}^{1-\mu_i}.$$

*Here, $G_i$ depends on the unobserved type $\mu_i$ as well as the information sets used by both types, $\omega_i^0$ and $\omega_i^1$.* □

Motivated by these examples, we consider the following general selection mechanism,

$$G_i = g(\omega_i, \nu_i), \quad i = 1, \ldots, n. \tag{3}$$

We allow $\omega_i$ to be a function or a subvector of $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \mu_i, \eta_{i1}, \eta_{i2})$ and can thereby accommodate the above examples as special cases. Moreover, we can accommodate many other economic models of selection (e.g., Heckman and Robb, 1985; Chabé-Ferret, 2015; Marx et al., 2024). The additional unobservables $(\mu_i, \eta_{i1}, \eta_{i2})$ capture any other determinants of selection that may be correlated with $Y_{it}(0)$, such as individual-specific thresholds in Example 2.1 or treatment effects and costs in Example 2.2.

The scalar unobservable, $\nu_i$, omitted for simplicity in the above examples, captures determinants of selection that are independent of $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \mu_i, \eta_{i1}, \eta_{i2})$ and allows us to accommodate the important special case of random assignment. We impose the following assumption.

**Assumption SEL.** *$P(\nu_i > c) \in (0,1)$ for some $c \in \mathbb{R}$ and $\nu_i \perp\!\!\!\perp (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \mu_i, \eta_{i1}, \eta_{i2})$.*

Note that since $G_i = D_{i2}$, $g(\cdot)$ can be equivalently viewed as the selection mechanism for $D_{i2}$. Let $\mathcal{G}_\omega$ denote the class of all selection mechanisms $g(\cdot)$ mapping from the support of $(\omega_i, \nu_i)$ to $\{0,1\}$. We index the class of selection mechanisms with $\omega_i$ since it could be correlated with the potential outcomes. Since $\omega_i$ can be interpreted as the units' information sets, the categorization of selection mechanisms based on $\omega_i$ is motivated by the usefulness of information sets for analyzing causal inference methods (see, e.g., Section 2 in Heckman and Vytlacil, 2007).

**Remark 2.1** (Parallel trends and functional form)**.** *Throughout this paper, we take the functional form of the outcome as given. We thereby abstract from the issues arising from the sensitivity of DiD to functional form specification (e.g., Roth and Sant'Anna, 2023).* □

# 3 Necessary and sufficient conditions for parallel trends

In this section, we provide necessary and sufficient conditions for parallel trends. Section 3.1 provides a general necessary and sufficient condition for parallel trends. We then apply this general condition in Sections 3.2–3.3 to derive necessary and sufficient conditions for various practically relevant scenarios by specifying what units select on.

## 3.1 A general necessary and sufficient condition for parallel trends

Here, we provide a general result that allows for specifying necessary and sufficient conditions for parallel trends for any class of selection mechanisms $\mathcal{G}_\omega$. Consistent with the nonparametric treatment of selection mechanisms in DiD analyses, we provide necessary and sufficient conditions for Assumption PT to hold for all $g \in \mathcal{G}_\omega$.[10]

Considering necessary and sufficient conditions for Assumption PT for all $g \in \mathcal{G}_\omega$ ensures that the invocation of Assumption PT is robust to the choice of selection mechanism within this class.[11] While robustness to the exact specification of the selection mechanism is important in all DiD applications, it is especially relevant in settings with aggregate selection (e.g., at the county or state level). When multiple entities or actors are involved, or when the decision is based on aggregating heterogeneous individual preferences, the selection mechanisms are often complex and difficult to model. Consider again the Medicaid expansion example. The survey of governors in Sommers and Epstein (2013) shows that there are many different factors affecting the governors' views on expanding Medicaid, ranging from fiscal considerations to the potential health outcomes of their constituents and the potential benefits of the expansion in terms of health insurance coverage and health outcomes.

The following theorem provides a necessary and sufficient condition for any class of selection mechanisms $\mathcal{G}_\omega$. We focus on non-degenerate DiD designs, that is, designs with $P(G_i = 1) \in (0, 1)$.

**Theorem 3.1** (Necessary and sufficient condition for PT to hold for all $g \in \mathcal{G}_\omega$)**.** *Suppose that Assumption SEL holds. Suppose further that either* $P(E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] > 0) < 1$ *or* $P(E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] < 0) < 1$*. Then, Assumption PT holds for all* $g \in \mathcal{G}_\omega$ *satisfying* $P(G_i = 1) \in (0, 1)$ *if and only if* $E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] = 0$ *a.s.*

The "if" direction of Theorem 3.1 follows by the law of iterated expectation. The proof of the "only if" direction is constructive: we provide an explicit "least favorable" selection

[10] These conditions imply that Assumption PT holds for all $g \in \mathcal{G}_\omega$, which is in the spirit of standard notions of nonparametric identification (e.g., Hansen, 2022, Definition 2.10).

[11] In Section 3.4, we further demonstrate that, even in the case where a researcher knows the underlying selection mechanism, the necessary and sufficient condition for all $g \in \mathcal{G}_\omega$ rules out cases where Assumption PT holds due to *peculiar* combinations of parameters of the data-generating process (see Figure 1).

mechanism that yields the necessary condition. For the case where $P(E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] > 0) < 1$, this selection mechanism takes the following form,

$$G_i = 1\{\nu_i > c\}1\{E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] \leq 0\}. \tag{4}$$

Assumption SEL (together with $P(E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] > 0) < 1$) ensures that this selection mechanism is non-degenerate when the necessary and sufficient condition holds.[12]

Stating the necessary and sufficient condition in Theorem 3.1 in terms of what the units select on and their information sets, $\omega_i$, rather than explicit selection mechanisms, has a key advantage: While it may be possible to determine (based on contextual and economic knowledge) what information units select on, there is likely uncertainty about the exact form of the selection mechanism, especially in settings with aggregate selection. The necessary and sufficient condition in Theorem 3.1 therefore relieves the researcher from the need to impose additional structure on the selection mechanism that is not justified by their context.

Next, we apply Theorem 3.1 to various practically relevant classes of selection mechanisms.

### 3.2 What if we impose no restrictions on selection?

Unlike other causal inference methods, DiD does not explicitly restrict selection into treatment. This begs the question: What if researchers are indeed not willing to impose any assumptions on selection so that parallel trends needs to hold for all selection mechanisms? To answer this question, we apply Theorem 3.1 with $\omega_i$ including all unobservables that could enter the selection mechanism in (3).

**Corollary 3.1** (No restrictions on selection)**.** *Under the assumptions of Theorem 3.1 with* $\omega_i = (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \mu_i, \eta_{i1}, \eta_{i2})$*, Assumption PT holds for all* $g \in \mathcal{G}_\omega$ *satisfying* $P(G_i = 1) \in (0, 1)$ *if and only if* $\dot{Y}_{i1}(0) = \dot{Y}_{i2}(0)$ *a.s. The result continues to hold if* $\omega_i = (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$*.*

To interpret the necessary and sufficient condition in Corollary 3.1, it is helpful to rewrite it as

$$Y_{i2}(0) - Y_{i1}(0) = E[Y_{i2}(0) - Y_{i1}(0)].$$

This shows that absent any restrictions on selection, parallel trends implies that the potential outcomes are constant over time, except for common mean shifts. This essentially rules out

[12] Under the necessary and sufficient condition, $1\{E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] \leq 0\} = 1$. In this case, the first component of the selection mechanism $1\{\nu_i > c\}$ plays the role of a tie-breaker, ensuring that mechanism (4) is non-degenerate.

time-varying unobservables. To see this, consider the following standard two-way model

$$Y_{it}(0) = \alpha_i + \lambda_t + \varepsilon_{it}, \quad E[\varepsilon_{it}] = 0, \tag{5}$$

where $\lambda_t$ is a nonstochastic time trend. The necessary and sufficient condition specialized to this separable outcome model is $\varepsilon_{i1} = \varepsilon_{i2}$, implying that $\varepsilon_{it}$ is time-invariant.

Given that the necessary and sufficient condition for the unrestricted class of selection mechanisms is implausible in most applications, we next consider restricted classes of selection mechanisms.

### 3.3 Necessary and sufficient conditions for restricted classes of mechanisms

DiD applications differ substantially in terms of what determines selection into treatment. Corollary 3.1 shows that restrictions on selection are unavoidable in realistic settings. In the following, we consider various restrictions on selection mechanisms that are practically relevant and well-established in the literature. The list of restrictions we consider is not exhaustive. The advantage of Theorem 3.1 is that researchers can specialize the necessary and sufficient condition for the class of selection mechanisms relevant for their application.

#### 3.3.1 Imperfect foresight

Selection on pre-treatment unobservables is likely in many applications. An example is when units make their selection decision based on expected future potential outcomes and costs, while only having access to pre-treatment information (e.g., $\omega_i = (\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$ in Examples 2.1 and 2.2). Another example is when units select into treatment in response to negative (or positive) pre-treatment shocks or their pre-treatment outcome falling below (or above) a specific threshold (e.g., Example 2.1 with $\beta = 0$ and $\omega_i = (\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$) and more broadly when there is feedback (e.g., Bonhomme, 2025; Chamberlain, 2022). Finally, imperfect foresight is relevant when individuals are myopic.

We first consider the case where selection depends on time-invariant and all pre-treatment unobservables, so that $\omega_i = (\alpha_i, \varepsilon_{i1}, \nu_i, \eta_{i1})$. For this case, Theorem 3.1 implies the following corollary.[13]

**Corollary 3.2** (Imperfect foresight: Case 1)**.** *Under the assumptions of Theorem 3.1 with* $\omega_i = (\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$*, Assumption PT holds for all* $g \in \mathcal{G}_\omega$ *satisfying* $P(G_i = 1) \in (0, 1)$ *if and only if* $E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = \dot{Y}_{i1}(0)$ *a.s.*

[13] We are grateful to Eric Mbakop for encouraging us to pursue necessary and sufficient conditions instead of necessary conditions only under imperfect foresight (and selection on fixed effects, discussed below).

The necessary and sufficient condition in Corollary 3.2 is a martingale-type condition on the untreated potential outcomes with respect to the unobservables that determine selection in this case. To build further intuition and to compare this result to Corollary 3.1, note that the necessary and sufficient condition can be equivalently written as

$$Y_{i2}(0) - Y_{i1}(0) = E[Y_{i2}(0) - Y_{i1}(0)] + \zeta_{i2}, \quad E[\zeta_{i2}|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = 0.$$

That is, the condition in Corollary 3.2 allows the untreated potential outcomes to vary over time beyond deterministic mean shifts but requires the stochastic component of the change over time, $\zeta_{i2}$, to be mean-independent of the pre-treatment unobservables $(\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$.

In the two-way model (5), the condition in Corollary 3.2 becomes $E[\varepsilon_{i2}|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = \varepsilon_{i1}$, a martingale-type property that implies that $\varepsilon_{i2} - \varepsilon_{i1} + \zeta_{i2}$, where $\zeta_{i2}$ is an innovation satisfying $E[\zeta_{i2}|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = 0$. This necessary and sufficient condition relates to the consistency of the first-difference estimator under violations of strict exogeneity when the idiosyncratic shocks follow a unit root.[14]

The martingale-type condition in Corollary 3.2 arises because the units select on the determinants of $Y_{i1}(0)$, $(\alpha_i, \varepsilon_{i1})$. As a result, the condition in Theorem 3.1 with $\omega_i = (\alpha_i, \varepsilon_{i1}, \nu_i, \eta_{i1})$, $E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = 0$, simplifies to $E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = \dot{Y}_{i1}(0)$, as stated in Corollary 3.2. If selection is based on pre-treatment unobservables that do not include the determinants of $Y_{i1}(0)$, the martingale condition does not arise, as the next corollary shows.

**Corollary 3.3** (Imperfect foresight: Case 2)**.** *Under the assumptions of Theorem 3.1 with $\omega_i = (\mu_i, \eta_{i1})$, Assumption PT holds for all $g \in \mathcal{G}_\omega$ satisfying $P(G_i = 1) \in (0, 1)$ if and only if $E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\mu_i, \eta_{i1}] = 0$ a.s.*

The necessary and sufficient condition in Corollary 3.3 resembles but is different from the standard definition of a martingale-difference condition on the difference $\Delta\dot{Y}_{i2}(0) = \dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)$ (e.g., Hamilton, 1994, p.189) since the conditioning set does not include lagged values of $\Delta\dot{Y}_{i2}(0)$. It can therefore be consistent with a wider class of time series processes than the condition in Corollary 3.2, which implies that $\dot{Y}_{it}(0)$ is a martingale.

### 3.3.2 Roy-style selection

Here, we consider settings with Roy-style selection, as in Example 2.2. We first consider the case where the units know their treatment effect $\tau_{i2} = Y_{i2}(1) - Y_{i2}(0)$ and costs $\kappa_{i2}$ in period $t = 2$. For this case, Theorem 3.1 implies the following corollary.

[14] We thank Stéphane Bonhomme for pointing out this connection.

**Corollary 3.4** (Roy-style selection)**.** *Under the conditions of Theorem 3.1 with* $\omega_i = (\tau_{i2}, \kappa_{i2})$*, Assumption PT holds for all* $g \in \mathcal{G}_\omega$ *satisfying* $P(G_i = 1) \in (0, 1)$ *if and only if* $E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\tau_{i2}, \kappa_{i2}] = 0$ *a.s.*

Rewriting the necessary and sufficient condition as $E[\dot{Y}_{i2}(0)|\tau_{i2}, \kappa_{i2}] = E[\dot{Y}_{i1}(0)|\tau_{i2}, \kappa_{i2}]$ demonstrates that it requires the conditional expectation of the demeaned untreated potential outcome given the treatment effects and costs to be equal across time. The condition would hold immediately if $(\tau_{i2}, \kappa_{i2})$ were independent of the untreated potential outcomes. However, this is an arguably unrealistic restriction in many applications.

In the two-way model (5), the necessary and sufficient condition in Corollary 3.4 simplifies to $E[\varepsilon_{i2}|\tau_{i2}, \kappa_{i2}] = E[\varepsilon_{i1}|\tau_{i2}, \kappa_{i2}]$. The condition would be clearly violated if $\tau_{i2}$ is a monotonic transformation of $\varepsilon_{i2}$. The condition is more plausible if instead $\tau_{i2}$ and $\kappa_{i2}$ were determined by time-invariant factors. This discussion demonstrates that under Roy-style selection, parallel trends implies restrictions on treatment effect heterogeneity.

In some applications, assuming that the units know their treatment effects and costs in $t = 2$ might not be plausible. Suppose instead that selection is based on expected treatment effects and costs conditional on all the available pre-treatment information, $E[\tau_{i2}|\omega_i]$ and $E[\kappa_{i2}|\omega_i]$, respectively, where $\omega_i = (\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$. For this case, the result in Corollary 3.2 applies, and the necessary and sufficient condition for parallel trends is the martingale-type condition, $E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = \dot{Y}_{i1}(0)$. If, instead, the expectations are conditional on the time-invariant unobservables $(\alpha_i, \mu_i)$, then the result in Corollary 3.5 below applies. More generally, Theorem 3.1 allows for considering many other variants of Roy-style selection.

### 3.3.3 Selection on fixed effects

Here, we consider the classical case of selection on fixed effects. Selection on fixed effects is plausible, for example, if the units' information sets only contain the time-invariant unobservables (in addition to $\nu_i$), so that $\omega_i = (\alpha_i, \mu_i)$, or if selection is directly based on fixed effects, as in Example 2.3.

The following corollary provides the necessary and sufficient condition under selection on fixed effects.

**Corollary 3.5** (Selection on fixed effects)**.** *Under the conditions of Theorem 3.1 with* $\omega_i = (\alpha_i, \mu_i)$*, Assumption PT holds for all* $g \in \mathcal{G}_\omega$ *satisfying* $P(G_i = 1) \in (0, 1)$ *if and only if* $E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\alpha_i, \mu_i] = 0$ *a.s.*

Corollary 3.5 shows that the necessary and sufficient condition for parallel trends under selection on fixed effects is a time-homogeneity restriction on the conditional mean of the

untreated potential outcome. To interpret this necessary condition, note that it implies that

$$Y_{i2}(0) - Y_{i1}(0) = E[Y_{i2}(0) - Y_{i1}(0)] + \zeta_{i2}, \quad E[\zeta_{i2}|\alpha_i, \mu_i] = 0.$$

This shows that Corollary 3.5 implies a weaker mean-independence condition on the stochastic trend component $\zeta_{i2}$ than, for example, Corollary 3.2, thus highlighting a trade-off between restrictions on selection and the evolution of the untreated potential outcomes over time.

In the context of the two-way model (5), the necessary condition in Corollary 3.5 simplifies to $E[\varepsilon_{i1}|\alpha_i, \mu_i] = E[\varepsilon_{i2}|\alpha_i, \mu_i]$, a time-homogeneity assumption on the conditional mean of the idiosyncratic shocks.[15] While it is not surprising that the condition $E[\varepsilon_{i1}|\alpha_i, \mu_i] = E[\varepsilon_{i2}|\alpha_i, \mu_i]$ is sufficient for parallel trends if $G_i = g(\alpha_i, \mu_i, \nu_i)$, Corollary 3.5 demonstrates that this condition is in fact *necessary* for parallel trends under model (5).

### 3.3.4 Selection on lagged outcomes (unconfoundedness)

A popular alternative to the parallel trends assumption is to assume that selection is based on lagged dependent variables (e.g., Angrist and Pischke, 2009; Ding and Li, 2019), $Y_{i2}(0) \perp\!\!\!\perp G_i \mid Y_{i1}(0)$, which we will refer to as *unconfoundedness*. Here we apply Theorem 3.1 to characterize the parallel trends assumption under unconfoundedness and shed light on the connection between these popular assumptions. Setting $\omega_i = Y_{i1}(0)$ in Theorem 3.1, so that $G_i$ satisfies unconfoundedness, provides such a characterization.

**Corollary 3.6** (Unconfoundedness)**.** *Under the assumptions of Theorem 3.1 with* $\omega_i = Y_{i1}(0)$*, Assumption PT holds for all* $g \in \mathcal{G}_\omega$ *satisfying* $P(G_i = 1) \in (0, 1)$ *if and only if* $E[\dot{Y}_{i2}(0)|Y_{i1}(0)] = \dot{Y}_{i1}(0)$ *a.s.*

Researchers imposing unconfoundedness typically do not impose additional explicit assumptions on the exact form of the selection mechanism. This provides an additional motivation for focusing on nonparametric conditions and deriving necessary and sufficient conditions for Assumption PT holding for all $g \in \mathcal{G}_\omega$.

Corollary 3.6 shows that parallel trends is equivalent to the demeaned potential outcomes $\dot{Y}_{it}(0)$ satisfying a martingale property under unconfoundedness. Written in terms of original outcomes $Y_{it}(0)$, the condition becomes $E[Y_{i2}(0)|Y_{i1}(0)] = Y_{i1}(0) + E[Y_{i2}(0) - Y_{i1}(0)]$, which holds, for example, if $Y_{it}(0)$ is a random walk with drift.

It is interesting to relate the result in Corollary 3.6 to results in the existing literature.

[15] The time-homogeneity condition in Corollary 3.5 relates to the strict exogeneity condition in fixed effects models. Suppose that $G_i = g(\alpha_i)$, then the strict exogeneity assumption $E[\varepsilon_{it}|G_i, \alpha_i] = 0$ implies that $E[\varepsilon_{it}|\alpha_i] = 0$, which in turn implies the time homogeneity condition in Corollary 3.5 with $\omega_i = \alpha_i$.

**Remark 3.1** (Connection to Ding and Li (2019))**.** *Here, we connect the analysis to Ding and Li (2019) (see also Angrist and Pischke, 2009). Ding and Li (2019) assume that*

$$E[Y_{i2}|Y_{i1}, G_i] = \theta_1 + \theta_2 Y_{i1} + \tau G_i. \tag{6}$$

*Proposition 1 in Ding and Li (2019), written in terms of population coefficients, implies that the estimand under* (6) *is*

$$E[Y_{i2}|G_i = 1] - E[Y_{i2}|G_i = 1] - \theta_2(E[Y_{i1}|G_i = 1] - E[Y_{i1}|G_i = 0]) \tag{7}$$

*This estimand is equivalent to* DiD *if and only if* $\theta_2 = 1$.[16]

*To relate the result in Ding and Li (2019) to Corollary 3.6, note that under unconfoundedness,* (6) *implies that* $E[Y_{i2}(0)|Y_{i1}(0)] = \theta_1 + \theta_2 Y_{i1}(0)$. *Corollary 3.6 implies that* $\theta_2 = 1$ *is necessary and sufficient for Assumption PT to hold. Since* DiD = ATT *under Assumption PT, the result in Ding and Li (2019, Proposition 1) is consistent with the necessary and sufficient condition in Corollary 3.6 under the linearity assumption* (6). □

### 3.3.5 Other selection mechanisms and trade-offs

The previous subsections illustrate the implications of Theorem 3.1 for various empirically relevant classes of selection mechanisms. Importantly, the result in Theorem 3.1 is very general and allows us to characterize the empirical content of parallel trends for many other relevant classes of selection mechanisms, including many of the existing selection models discussed in the literature and reviewed in Section 1.1. Applying Theorem 3.1 only requires specifying $\omega_i$, that is, what information the units select on. In practice, the specification of $\omega_i$ should be guided by contextual and economic knowledge.

Varying $\omega_i$ allows us to characterize trade-offs between restrictions on selection, encoded in $\omega_i$, and restrictions on the time series properties of the untreated potential outcomes. The richer the information that the units select on, the more restrictive the time series restriction required for parallel trends to hold. The time series restrictions are particularly strong if selection is based on the unobservable determinants of $Y_{it}(0)$, that is, if $\omega_i$ includes or depends on $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$.

The trade-offs between assumptions on selection and the time series properties of the untreated potential outcomes are particularly easy to see under explicit models for the untreated potential outcomes. To illustrate, suppose that $Y_{it}(0)$ is given by model (5). Then,

[16] Note that main bracketing result, Theorem 1 in Ding and Li (2019), does not apply in this case since Condition 1 (stationarity) is violated.

our necessary and sufficient conditions imply that parallel trends holds, for example, in the following scenarios:[17]

(a) Imperfect foresight (case 1): (i) $\omega_i = (\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$ and (ii) $E[\varepsilon_{i2}|\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = \varepsilon_{i1}$

(b) Imperfect foresight (case 2): (i) $\omega_i = (\mu_i, \eta_{i1})$ and (ii) $E[\varepsilon_{i2} - \varepsilon_{i1}|\mu_i, \eta_{i1}] = 0$

(c) Roy-style selection: (i) $\omega_i = (\tau_{i2}, \kappa_{i2})$ and (ii) $E[\varepsilon_{i2}|\tau_{i2}, \kappa_{i2}] = E[\varepsilon_{i1}|\tau_{i2}, \kappa_{i2}]$

(d) Selection on fixed effects: (i) $\omega_i = (\alpha_i, \nu_i)$ and (ii) $E[\varepsilon_{i2}|\alpha_i, \nu_i] = E[\varepsilon_{i1}|\alpha_i, \nu_i]$

The conditions (a)–(d) provide practitioners with explicit theory-based templates for assessing and justifying parallel trends assumptions and can be used in conjunction with the selection mechanisms in Examples 2.1, 2.2, 2.3, and 2.4, or other selection mechanisms in the literature. These conditions allow researchers to provide, in the words of McKenzie (2022), "plausible rhetorical arguments as to why we should think the [parallel trends] assumptions hold."

### 3.4 The relevance of parallel trends for all $g \in \mathcal{G}_\omega$ for empirical practice

Theorem 3.1 and Corollaries 3.1–3.5 present necessary and sufficient conditions for parallel trends to hold for all $g \in \mathcal{G}_\omega$ because we are interested in nonparametric conditions, consistent with the nonparametric (model-agnostic) treatment of selection mechanisms in the DiD analyses. Here we elaborate on the practical relevance of focusing on *parallel trends for all* $g \in \mathcal{G}_\omega$. This is a crucial question, as it might not be obvious why practitioners should consider *parallel trends for all* $g \in \mathcal{G}_\omega$, when it is clearly stronger than *parallel trends for a specific* $g \in \mathcal{G}_\omega$. To keep our discussion concrete, we focus on the case of imperfect foresight where the units select on pre-treatment unobservables, so that $\omega_i = (\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$, but the arguments we make are relevant for any class of selection mechanisms.

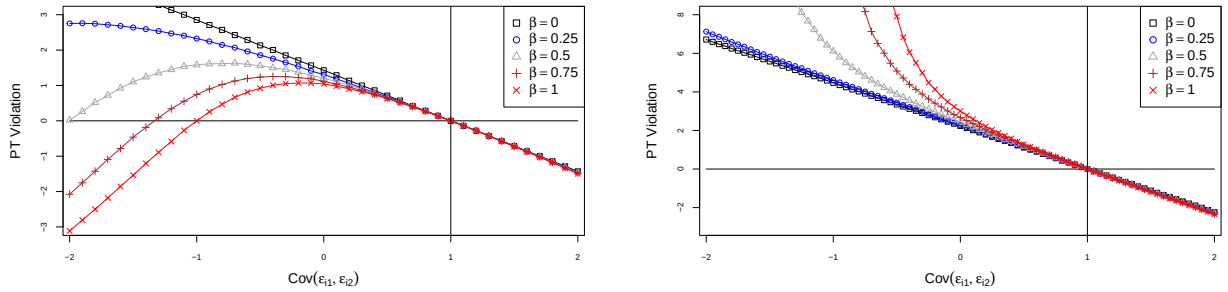
First, in applications where contextual or economic knowledge suggests that units select on pre-treatment unobservables, there is likely uncertainty about the exact form of the selection mechanism. For instance, practitioners might not want to take a stance on (i) how expectations are formed (e.g., subjective expectations may be different from conditional expectations) or (ii) whether selection is based on the discounted sum of expected untreated outcomes (Example 2.1), on expected gains (Example 2.2), or other quantities motivated by economic models of selection. As discussed above, the uncertainty about the exact form of the selection mechanism can be particularly pronounced in settings with aggregate selection.

[17] Theorem 3.1 provides a general framework for deriving sufficient conditions, depending on what unit select on. However, in some applications, researchers might be interested in imposing other types of assumptions on selection. In Appendix C, we provide a sufficient condition based on symmetry of the selection mechanism.

Second, even if one is certain about the exact parametric form of the units' selection mechanism, the exact distribution of unobservables, and how expectations are formed, one would typically not want parallel trends to depend on specific parameter choices. To illustrate, consider Example 2.1 with $\omega_i = (\alpha_i, \varepsilon_{i1})$. Figure 1 plots the parallel trends violation, $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0]$, for different discount factors $\beta$ against $Cov(\varepsilon_{i1}, \varepsilon_{i2})$ for two different outcome models: (a) a separable model with autocorrelated shocks, (b) an autoregressive model with a drift. For both models, regardless of $\beta$, the parallel trends violation is exactly zero when $Cov(\varepsilon_{i1}, \varepsilon_{i2}) = 1$, which corresponds to the martingale condition (Panels (a) and (b) in Figure 1). For the separable model, parallel trends additionally holds for very specific combinations of $\beta$ and $Cov(\varepsilon_{i1}, \varepsilon_{i2})$ (Panel (a) in Figure 1). These additional instances of parallel trends, however, require a researcher to not only be willing to choose a parametric distribution, but also to rely on very particular combinations of the discount factor $\beta$ and $Cov(\varepsilon_{i1}, \varepsilon_{i2})$ (in addition to a specific selection and outcome model). By focusing on parallel trends for all $g \in \mathcal{G}_\omega$, we rule out these additional cases and focus on "robust" instances of parallel trends.

Figure 1: Numerical Illustration: Example 2.1 with $\omega_i = (\alpha_i, \varepsilon_{i1})$

(a) Separable model (b) Autoregressive model




*Notes:* For both (a) and (b), $G_i = 1\{Y_{i1}(0) + \beta E[Y_{i2}(0)|\alpha_i, \varepsilon_{i1}] \leq c\alpha_i\}$ for $i = 1, \ldots, n$. The parallel trends violation (PT violation), $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0]$, is computed numerically using $n = 10^6$. For (a), the untreated potential outcomes are generated as follows: $Y_{it}(0) = \alpha_i + \lambda_t + \varepsilon_{it}$ for $t = 1, 2$, where $\lambda_1 = 0$, $\alpha_i \sim N(0, 1)$, $\alpha_i \perp\!\!\!\perp (\varepsilon_{i1}, \varepsilon_{i2})$, and $(\varepsilon_{i1}, \varepsilon_{i2})$ are jointly normal. We normalize $Var(\varepsilon_{i1}) = 1$ such that $E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_{i1}] = \alpha_i + Cov(\varepsilon_{i1}, \varepsilon_{i2})\varepsilon_{i1}$, where $Cov(\varepsilon_{i1}, \varepsilon_{i2}) = \rho_\varepsilon \sigma_2$, $\sigma_2^2 \equiv Var(\varepsilon_{i2})$ and $\rho_\varepsilon \equiv Corr(\varepsilon_{i1}, \varepsilon_{i2})$. The plot in (a) presents the PT violations for a grid of values of $Cov(\varepsilon_{i1}, \varepsilon_{i2})$ with $\lambda_2 = \sigma_2 = 2$, $c = 0.5$. For (b), $Y_{i1}(0) = \alpha_i + \varepsilon_{i1}$, $Y_{i2}(0) = \rho_2 \alpha_i + \lambda_2 + \varepsilon_{i2}$, where $\alpha_i \perp\!\!\!\perp (\varepsilon_{i1}, \varepsilon_{i2})$, $\varepsilon_{i2} = \rho_2 \varepsilon_{i1} + \zeta_{i2}$, and $(\varepsilon_{i1}, \zeta_{i2})$ are jointly normal. In this model, $E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_{i1}] = \rho_2 \dot{Y}_{i1}(0)$. We normalize $Var(\varepsilon_{i1}) = 1$ such that $Cov(\varepsilon_{i1}, \varepsilon_{i2}) = \rho_2$ and use $\sigma_2^2$ to denote $Var(\zeta_{i2})$. The plot in (b) presents the PT violation for a grid of values of $Cov(\varepsilon_{i1}, \varepsilon_{i2})$ with $\lambda_2 = \sigma_2 = 2$, $c = 0.25$.

### 3.5 Extensions

Here, we summarize three main extensions. We refer to the corresponding appendices for details.

**Disaggregate data and aggregate decisions**. In many DiD applications, the selection decisions are made at the aggregate level (e.g., at the county or state level), while outcome data are available at the disaggregate level (e.g., at the individual or firm level). In Appendix A, we consider a sharp DiD design with $n_s$ individuals, indexed by $i = 1, \ldots, n_s$, belonging to aggregate unit $s$. Let $Y_{st}(0)$ denote the untreated potential outcome for aggregate unit $s$ in period $t$ (e.g., the average of the disaggregate outcomes $Y_{ist}(0)$, $Y_{st}(0) = n_s^{-1} \sum_{i=1}^{n_s} Y_{ist}(0)$) and $G_s$ the selection decision of unit $s$. The necessary and sufficient conditions in this section directly apply to this setting by replacing $i$ with $s$ and interpreting the unobservables and potential outcomes as aggregate quantities. That said, being explicit about the aggregation can help "microfound" restrictions on selection, as we discuss in Appendix A.

**Multiple periods and groups**. In Appendix B, we extend our results to DiD designs with multiple periods and multiple groups.[18] Specifically, we consider a staggered adoption setting with $T$ periods, where no units are treated at $t = 1$ and some units remain untreated at $t = T$. Appendix B demonstrates that the necessary and sufficient condition in Theorem 3.1 extends naturally to this setting, and based on it, it is straightforward to extend our theoretical results to DiD settings with multiple periods and staggered adoption.

**Covariates**. In many applications, parallel trends may only be plausible conditional on covariates (e.g., Heckman et al., 1997; Abadie, 2005; Sant'Anna and Zhao, 2020a; Callaway and Sant'Anna, 2021). Therefore, we study the role of covariates through the lens of selection into treatment in Appendix D. We explicitly allow for a vector of both time-invariant and time-varying covariates, $X_{it}$, assuming that $X_{it}$ is not affected by the treatment. In Appendix D.1, we show that the necessary and sufficient conditions for conditional parallel trends imply separability requirements on how the covariates can enter the outcome model. Appendix D.2 provides selection-based templates for justifying conditional parallel trends assumptions for separable models. In Appendix D.3, we propose a weaker conditional parallel trends assumption that accommodates a rich class of nonseparable models and provide sufficient conditions for this assumption.

[18] Our setup and notation build on Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Roth et al. (2023).

# 4 Selection-based bias decomposition with an application to imperfect foresight

The necessary and sufficient conditions in Section 3 demonstrate that if we allow for selection on time-varying shocks and in particular on the determinants of the untreated potential outcomes, parallel trends implies strong restrictions on the time series properties of these outcomes. Here we analyze the bias of DiD when these necessary and sufficient conditions are violated.

The bias analysis accommodates, but does not require, data on additional pre-treatment periods. Suppose that there is one additional pre-treatment period, $t = 0$, in which no units are treated, so that $Y_{i0} = Y_{i0}(0)$ for $i = 1, \ldots, n$. We allow selection to also depend on the shocks in period $t = 0$, that is, we allow $\omega_i$ to be a function of $(\varepsilon_{i0}, \eta_{i0})$.

## 4.1 Bias decomposition

The following lemma provides a decomposition of the bias of DiD.

**Lemma 4.1** (Bias decomposition)**.** *Suppose that $P(G_i = 1) \in (0, 1)$. Then, for a given $\omega_i$, the bias of DiD can be decomposed as follows,*

$$\text{DiD} - \text{ATT} = \Delta^{\text{sel}}_{\text{post}} + \Delta^{\text{dev}}_{\text{post}},$$

*where*

$$\begin{aligned}
\Delta^{\text{sel}}_{\text{post}} &= \frac{E[(G_i - E[G_i|\omega_i])(\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)}, \\
\Delta^{\text{dev}}_{\text{post}} &= \frac{E[E[G_i|\omega_i]E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i]]}{P(G_i = 1)P(G_i = 0)}.
\end{aligned}$$

The decomposition in Lemma 4.1 shows that the bias of DiD is equal to the sum of two components. The component $\Delta^{\text{sel}}_{\text{post}}$ captures the parallel trends violation due to selection on unobservables not contained in $\omega_i$. The component $\Delta^{\text{dev}}_{\text{post}}$ captures the parallel trends violation due to deviations from the necessary and sufficient condition for Assumption PT when selection is based on $\omega_i$.

In the following, we apply the bias decomposition in Lemma 4.1 to settings where selection based on pre-treatment unobservables is likely. Other classes of selection mechanisms could also be considered.

### 4.2 Application to imperfect foresight

Suppose that researchers deem selection on pre-treatment unobservables likely. In this case, there are two potential sources of bias: selection on post-treatment unobservables and violations of the martingale condition. In the following, we characterize these two bias terms and provide stratgies for benchmarking them using pre-treatment data.

To simplify the notation, define $\varepsilon_i^t \equiv (\varepsilon_{i0}, \ldots, \varepsilon_{it})$ and $\eta_i^t \equiv (\eta_{i0}, \ldots, \eta_{it})$ for $t > 0$. Suppose that $\omega_i = (\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1)$. In this case, the necessary and sufficient condition in Corollary 3.2 generalizes to

$$E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1] = \dot{Y}_{i1}(0). \tag{8}$$

To aid interpretation and assess the magnitude of the bias components in Lemma 4.1, we characterize them under the following linear relaxation of the martingale condition (8).[19]

**Assumption REL.** *The following relaxation of the martingale condition holds:*[20]

$$E[\dot{Y}_{it}(0)|\alpha_i, \varepsilon_i^{t-1}, \mu_i, \eta_i^{t-1}] = \rho_t \dot{Y}_{i(t-1)}(0), \quad i = 1, \ldots, n, \quad t = 1, 2$$

Assumption REL imposes an AR(1) model with time-varying coefficients on $\dot{Y}_{it}(0)$,

$$\dot{Y}_{it}(0) = \rho_t \dot{Y}_{i(t-1)}(0) + \zeta_{it}, \quad E[\zeta_{it}|\alpha_i, \varepsilon_i^{t-1}, \mu_i, \eta_i^{t-1}] = 0. \tag{9}$$

Note that Assumption REL is imposed on the demeaned potential outcomes and thus allows for mean shifts in $Y_{it}(0)$. If $\rho_2 = 1$, Assumption REL reduces to the martingale assumption, $E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1] = \dot{Y}_{i1}(0)$. As a result, deviations from this martingale property under Assumption REL are fully characterized by the deviation of $\rho_2$ from 1, $(\rho_2 - 1)$.

The following proposition characterizes the bias components under Assumption REL.

**Proposition 4.1** (Bias characterization under linear martingale relaxation)**.** *Suppose that* $P(G_i = 1) \in (0, 1)$, $\omega_i = (\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1)$, *and Assumption REL holds. Then,*

$$\begin{aligned} \Delta_{\text{post}}^{\text{sel}} &= E[\zeta_{i2}|G_i = 1] - E[\zeta_{i2}|G_i = 0], \\ \Delta_{\text{post}}^{\text{dev}} &= (\rho_2 - 1)(E[Y_{i1}|G_i = 1] - E[Y_{i1}|G_i = 0]). \end{aligned}$$

[19] We focus on linear relaxations for convenience. Extensions to nonparametric relaxations of the form $E[\dot{Y}_{it}(0)|\alpha_i, \varepsilon_{i0}, \ldots, \varepsilon_{i(t-1)}] = \sigma_t \rho(\dot{Y}_{i(t-1)}(0))$, where $\rho(\cdot)$ is an arbitrary nonparametric function and $\sigma_1$ is normalized to one, are straightforward.

[20] Assumption REL yields a linear autoregressive model. This class of models has been studied extensively in the time series literature under restrictions on the heterogeneity of the coefficient (e.g., Nicholls and Quinn, 1982; Regis et al., 2022).

Proposition 4.1 shows that $\Delta^{\text{sel}}_{\text{post}}$ is equal to the mean difference in $\zeta_{i2}$ between both groups. Recall that $\zeta_{i2}$ is the difference between $\dot{Y}_{i2}(0)$ and its conditional expectation given pre-treatment unobservables $E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon^1_i, \mu_i, \eta^1_i]$. If selection depends on post-treatment unobservables including $\varepsilon_{i2}$, then $\zeta_{i2}$ is correlated with selection $G_i$, so that $E[\zeta_{i2}|G_i = 1]$ is not equal to $E[\zeta_{i2}|G_i = 0]$.

Proposition 4.1 further shows that $\Delta^{\text{dev}}_{\text{post}}$ is equal to the product of the martingale deviation, $(\rho_2 - 1)$, and the observed pre-treatment difference, $E[Y_{i1}|G_i = 1] - E[Y_{i1}|G_i = 0]$. This shows that the sensitivity of DiD with respect to violations of the martingale assumption depends on the pre-treatment group difference. It underscores that only if the martingale property holds exactly can we ignore pre-treatment differences (and the selection on unobservables they are indicative of). This discussion motivates using covariate adjustment for reducing the pre-treatment difference and thus the potential bias of DiD due to violations of the martingale assumption. We illustrate this point in Section 5 and defer the formal analysis of the DiD bias decomposition with covariates to Appendix F. If the treatment is randomly assigned, then $E[Y_{i1}|G_i = 1] - E[Y_{i1}|G_i = 0] = 0$ and $\Delta^{\text{dev}}_{\text{post}} = 0$.

An important takeaway from Proposition 4.1 is that the bias of DiD is an affine function of the martingale deviation $(\rho_2 - 1)$, where the slope is the pre-treatment difference, $E[Y_{i1}|G_i = 1] - E[Y_{i1}|G_i = 0]$, and the intercept is the post-treatment difference, $E[\zeta_{i2}|G_i = 1] - E[\zeta_{i2}|G_i = 0]$. The only component that is directly observable from the data is the pre-treatment difference. We next demonstrate how we can benchmark $\rho_2$ and $E[\zeta_{i2}|G_i = 1] - E[\zeta_{i2}|G_i = 0]$ using pre-treatment data.

### 4.3 Benchmarking DiD bias components

Here, we provide benchmarks for $\Delta^{\text{sel}}_{\text{post}}$ and $\Delta^{\text{dev}}_{\text{post}}$ based on pre-treatment data that allow practitioners to assess and sign the bias of DiD in applications.

The unobservable component in $\Delta^{\text{dev}}_{\text{post}}$ is $\rho_2$ for which there are two natural benchmarks. First, under Assumption REL, $\rho_1$ can be identified from the pre-treatment data by noting that $E[\dot{Y}_{i1}(0)|\dot{Y}_{i0}(0)] = E[E[\dot{Y}_{i1}(0)|\alpha_i, \varepsilon_{i0}, \mu_i, \eta_{i0}]|\dot{Y}_{i0}(0)] = \rho_1 \dot{Y}_{i0}(0)$, such that $\rho_1$ is identified as the coefficient of a population regression of $\dot{Y}_{i1}$ on $\dot{Y}_{i0}$. This is a useful benchmark because $\rho_2 = \rho_1$ under time-homogeneity of $\rho_t$ in Assumption REL. Second, we can consider the persistence in the control group, $E[\tilde{Y}_{i2}(0)|\tilde{Y}_{i1}(0), G_i = 0] = \rho^0_2 \tilde{Y}_{i1}(0)$, where $\tilde{Y}_{it}(0) \equiv Y_{it}(0) - E[Y_{it}(0)|G_i = 0]$, and use $\rho^0_2$ to inform $\rho_2$. This is a useful benchmark because $\rho_2 = \rho^0_2$ under unconfoundedness.[21]

[21] Specifically, the unconfoundedness assumption $Y_{i2}(0) \perp\!\!\!\perp G_i|Y_{i1}(0)$ implies that $\rho_2 = \rho^0_2$. This follows because $Y_{i2}(0) \perp\!\!\!\perp G_i|Y_{i1}(0)$ implies that $E[\dot{Y}_{i2}(0)|\dot{Y}_{i1}(0), G_i = 0] = E[\dot{Y}_{i2}(0)|\dot{Y}_{i1}(0)]$. The proposed benchmarking strategy allows researchers to consider a range of values for $\rho_2$, including $\rho^0_2$, the value of $\rho_2$ identified under unconfoundedness.

As for $\Delta^{\text{sel}}_{\text{post}}$, it is helpful to consider its observable pre-treatment analogue,

$$\Delta^{\text{sel}}_{\text{pre}} \equiv E[\zeta_{i1}|G_i = 1] - E[\zeta_{i1}|G_i = 0], \quad \text{where } \zeta_{i1} = \dot{Y}_{i1} - \rho_1 \dot{Y}_{i0}.$$

To relate $\Delta^{\text{sel}}_{\text{post}}$ and $\Delta^{\text{sel}}_{\text{pre}}$, note that (assuming $\Delta^{\text{sel}}_{\text{pre}} \neq 0$)

$$\Delta^{\text{sel}}_{\text{post}} = \frac{\rho_{G,\zeta_2}\sigma_{\zeta_2}}{\rho_{G,\zeta_1}\sigma_{\zeta_1}} \Delta^{\text{sel}}_{\text{pre}},$$

where $\rho_{G,\zeta_t} \equiv Corr(G_i, \zeta_{it})$ and $\sigma^2_{\zeta_t} \equiv Var(\zeta_{it})$.

In the case where $\sigma_{\zeta_1} = \sigma_{\zeta_2}$, the relative magnitude of $\Delta^{\text{sel}}_{\text{post}}$ to $\Delta^{\text{sel}}_{\text{pre}}$ is simply the ratio of the (scale-free) correlation coefficients, $\rho_{G,\zeta_2}/\rho_{G,\zeta_1}$. This ratio measures the relative degree of selection on pre- vs. post-treatment unobservables. For example, if contextual knowledge suggests that there is "more selection" on pre-treatment than on post-treatment unobservables, then $|\Delta^{\text{sel}}_{\text{post}}| \leq |\Delta^{\text{sel}}_{\text{pre}}|$, assuming $\text{sgn}(\rho_{G,\zeta_1}) = \text{sgn}(\rho_{G,\zeta_2})$. The edge case where $\Delta^{\text{sel}}_{\text{post}} = \Delta^{\text{sel}}_{\text{pre}}$ captures settings where the extent of selection on post- and pre-treatment unobservables is the same.

In practice, we recommend that researchers determine the robustness of DiD results based on the benchmarks we discuss above. We illustrate this approach in two empirical applications in Section 5.

## 5 Empirical illustration of bias decomposition

Here we illustrate the bias decomposition in two applications. In the first application, we revisit the NSW training program, where we have access to an experimental estimate of the ATT and thus an estimate of the bias of DiD. This bias estimate allows us to directly evaluate the bias decomposition and benchmarking strategy using a LaLonde (1986)-style exercise. In the second application, we consider the Medicaid expansion to demonstrate the usefulness of the bias decomposition to DiD applications with aggregate selection.

Both applications demonstrate that pre-treatment differences in mean outcomes between the treatment and control groups matter. While we can ignore such differences under parallel trends, our analysis underscores the central role they play once we entertain the possibility of violations of parallel trends. Both applications further highlight that covariates are crucial for reducing the pre-treatment difference in mean outcomes and thereby rendering DiD less sensitive to martingale violations.

### 5.1 NSW training program

#### 5.1.1 Setup and DiD analysis

The evaluation of job training programs is one of the classical applications of DiD in economics. Here, we revisit the analysis of the causal effect of the NSW training programs on post-treatment earnings (e.g., LaLonde, 1986). We use the same dataset as Sant'Anna and Zhao (2020a) and consider the "Dehejia and Wahba (1999, 2002) sample."[22] This sample combines the experimental treatment group (185 individuals) with an observational control group (15,992 individuals).

The outcome of interest is earnings. We observe individual-level data on earnings for two pre-treatment periods, 1974 and 1975, and one post-treatment period, 1978. We also have access to a set of baseline covariates: age, years of education, and indicators for high school dropouts, married individuals, Black and Hispanic individuals.

The unconditional DiD estimate using 1975 as the pre-treatment period ($t = 1$) and 1978 as the post-treatment period ($t = 2$) is equal to $\widehat{\text{DiD}} = 3{,}621$ (s.e. 610). A comparison to the experimental benchmark, which is 1,794 (s.e. 671), shows that the unconditional DiD substantially overestimates the returns to the training program. The estimated bias relative to the experimental benchmark is statistically and economically significant at 1,827, comparable in magnitude to the experimental benchmark.

With covariates, the regression-adjusted DiD estimate under conditional parallel trends is equal to $E_n[\widehat{\text{DiD}}(X_i)|G_i = 1] = 2{,}436$ (s.e. 653), where $E_n$ denotes the sample average and $\widehat{\text{DiD}}(X_i)$ is the conditional DiD estimate obtained using the regression-adjusted DiD estimator. This shows that adjusting for differences in baseline covariates reduces the bias of DiD to 642, about a third of the bias of the unconditional DiD relative to the experimental benchmark.

It is standard to report the results from pre-trends tests when there are additional pre-treatment periods. Based on the pre-treatment data from 1974 and 1975, the unconditional and regression-adjusted DiD estimates are 198 (s.e. 280) and 335 (s.e. 309), respectively.

Despite the non-rejections of the pre-trends tests, the sensitivity of the DiD estimates to parallel trends violations remains a major concern for three reasons. First, pre-tests are, by construction, not direct tests of parallel trends assumptions. Second, these tests can be substantially underpowered (e.g., Roth, 2022). Finally, building on the necessary and sufficient conditions in Section 3, Ghanem et al. (2026) show that pre-trends can be uninformative under imperfect foresight. These issues are particularly evident in this application, where the pre-test does not reject, despite unconditional DiD being significantly biased relative to the

[22] The data are from the `DRDID` `R`-package (Sant'Anna and Zhao, 2020b).

experimental benchmark. Next, we demonstrate how our selection-based bias decomposition can help us better understand the difference between the DiD and the experimental estimate in this application.

### 5.1.2 Decomposing the bias of DiD

We start by illustrating the bias decomposition without covariates. Replacing the population expectations by sample averages, we obtain

$$\begin{aligned}\widehat{\Delta}_{\text{post}} = \widehat{\Delta}_{\text{post}}\left(\Delta^{\text{sel}}_{\text{post}}, \rho_2\right) &= \Delta^{\text{sel}}_{\text{post}} + (\rho_2 - 1)(E_n[Y_{i1}|G_i = 1] - E_n[Y_{i1}|G_i = 0]), \\ &= \Delta^{\text{sel}}_{\text{post}} + (\rho_2 - 1)(-12{,}119).\end{aligned}$$

There is a substantial pre-treatment difference: average earnings in 1975 are much lower in the treatment than in the control group.

Figure 2a displays $\widehat{\Delta}_{\text{post}}$ as a function of $\rho_2$ together with the bias estimate based on the experimental benchmark. Suppose first that $\Delta^{\text{sel}}_{\text{post}} = 0$. In this case, the bias of DiD equals $\widehat{\Delta}_{\text{post}} = (\rho_2 - 1)(-12{,}119)$, depicted by the blue line. It is solely driven by violations of the martingale property (i.e., differences between $\rho_2$ and 1). Alternatively, consider the edge case where $\Delta^{\text{sel}}_{\text{post}} = \Delta^{\text{sel}}_{\text{pre}}$. This corresponds to the case with equal sign and strength of selection on $\zeta_{i1}$ and $\zeta_{i2}$, as discussed in Section 4.3. The sample analogue of $\Delta^{\text{sel}}_{\text{pre}}$ equals $-2,049$, resulting in the following bias estimate (red line in Figure 2a),

$$\widehat{\Delta}_{\text{post}} = -2,049 + (\rho_2 - 1)(-12{,}119).$$

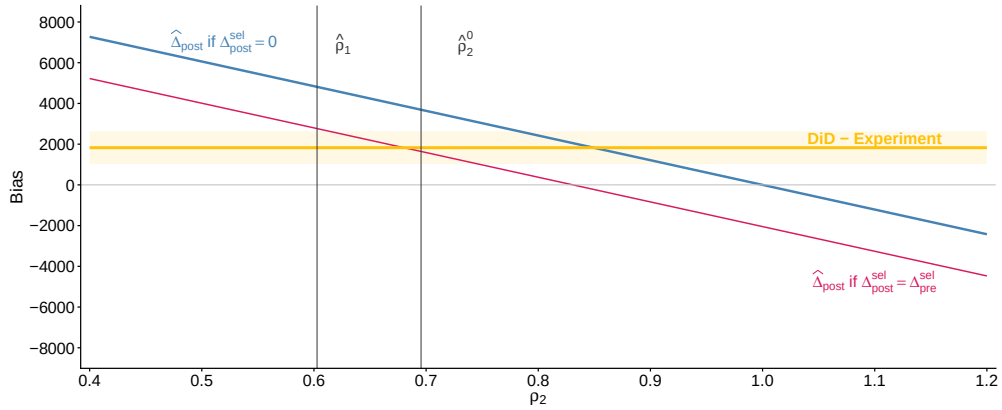
As we discussed in Section 4.3, there are two natural benchmarks for $\rho_2$: $\rho_1$, the pre-treatment counterpart of $\rho_2$ and $\rho_2^0$, its control group counterpart. The corresponding estimates are $\widehat{\rho}_1 = 0.603$ and $\widehat{\rho}_2^0 = 0.695$ and are depicted in Figure 2a.[23] Both benchmark values would suggest that the unconditional DiD is upwardly biased, consistent with the experimental bias estimate.

The analysis without covariates demonstrates that the bias of DiD is very sensitive to deviations from the martingale property. The lack of robustness is driven by the treatment and control groups being very different before the treatment. This discussion suggests that we may reduce the pre-treatment difference and improve the robustness of DiD by adjusting

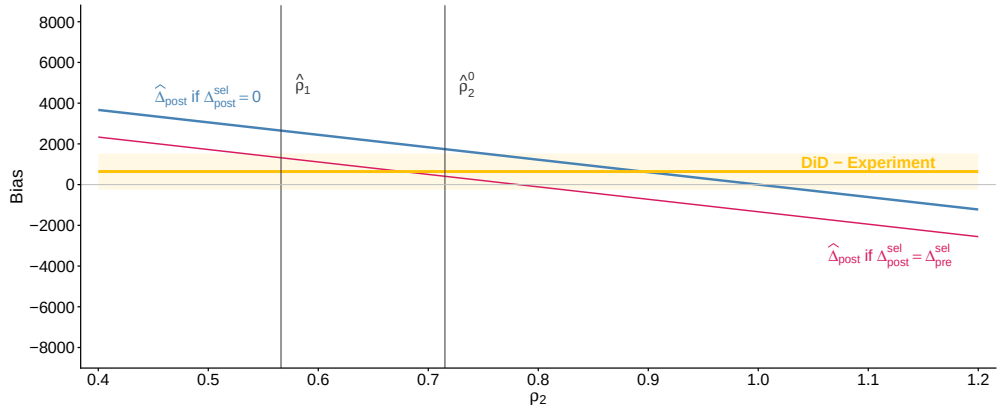
[23] Recall that the post-treatment earnings are measured in 1978, so that $\rho_2$ measures the persistence over three years. To account for the difference in periodicity when estimating $\rho_1$, we proceed in two steps. First, we regress $\dot{Y}_{i1975}$ on $\dot{Y}_{i1974}$ to obtain an estimate of the yearly persistence in the pre-treatment period, $\tilde{\rho}_1 = 0.845$. Second, we adjust for the difference in periodicity by computing $\widehat{\rho}_1$ as $\widehat{\rho}_1 = (\widehat{\tilde{\rho}}_1)^3 = 0.603$. This is justified under a linear AR(1) model for the demeaned outcomes in the pre-treatment period.

Figure 2: Benchmarking the Bias of DiD: Application to NSW Training Program

(a) Without covariates




(b) With covariates




*Notes:* Figure 2a displays the results from the bias decomposition without covariates. Figure 2b shows the results from the bias decomposition with regression adjustment using age, years of education, indicators for high school dropouts, married individuals, Black and Hispanic individuals, age squared, age cubed (divided by 1,000), and years of schooling squared. The shaded areas depict 95% confidence intervals for the difference between the unconditional DiD (regression-adjusted DiD in Figure 2b) and the experimental estimates using Bayesian-bootstrapped standard errors clustered at the individual level, based on 10,000 bootstrap draws. Data: Sant'Anna and Zhao (2020b).

for differences in baseline covariates.

We therefore incorporate covariates into our analysis in Figure 2b. In Appendix F, we show that under a linear relaxation of the conditional martingale property, the unconditional bias of DiD with covariates can be decomposed as

$$\Delta_{\text{post}} \quad = \quad \Delta^{\text{sel}}_{\text{post}} + (\rho_2 - 1)(E[Y_{i1}|G_i = 1] - E[E[Y_{i1}|G_i = 0, X_i]|G_i = 1]),$$

where $\Delta^{\text{sel}}_{\text{post}} \equiv E[\Delta^{\text{sel}}_{\text{post}}(X_i)|G_i = 1]$.

Analogous to the unconditional bias decomposition, consider first the case where $\Delta^{\text{sel}}_{\text{post}} =$

0. Using the regression-adjusted estimator for the pre-treatment difference described in Appendix F.2, we obtain the following bias estimate (blue line in Figure 2b), $\widehat{\Delta}_{\text{post}} = (\rho_2 - 1)(-6,113)$. Adjusting for differences in baseline covariates reduces the magnitude of the pre-treatment difference by approximately 50%. As a result, incorporating covariates makes the bias of DiD less sensitive to violations of the martingale property.

Alternatively, consider the case where $\Delta^{\text{sel}}_{\text{pre}} = \Delta^{\text{sel}}_{\text{post}}$, which is implied by $\Delta^{\text{sel}}_{\text{pre}}(X_i) = \Delta^{\text{sel}}_{\text{post}}(X_i)$. Using the regression-adjusted estimator of $\Delta^{\text{sel}}_{\text{pre}}$ described in Appendix F.2, we obtain (red line in Figure 2b)

$$\widehat{\Delta}_{\text{post}} = -1,333 + (\rho_2 - 1)(-6,113).$$

The estimates of $\rho_1$ and $\rho_2^0$ with covariates are $\widehat{\rho}_1 = 0.566$, which is somewhat smaller than without covariates, and $\widehat{\rho}_2^0 = 0.715$, which is somewhat larger than without covariates.[24] Both benchmark values suggest the same sign and a similar magnitude of the bias as the experimental benchmark.

This analysis demonstrates how the proposed bias decomposition can help empirical practitioners assess the bias of DiD and its sensitivity. This is especially important in applications such as this one, where the (unconditional) pre-trends tests do not reject, even though DiD is biased relative to the experimental benchmark.

## 5.2 Medicaid Expansion

### 5.2.1 Setup and DiD analysis

We revisit the DiD evaluation of Medicaid expansion to illustrate the relevance of our selection-based bias decomposition to DiD settings with aggregate selection. We use the sample from the 2×2 DiD implementation in Baker et al. (2026), but consider one additional pre-treatment period.[25] The treatment group consists of states that have expanded Medicaid in 2014, whereas the control group consists of states that have not expanded by 2019. The pre-treatment periods are 2012 and 2013 ($t = 0, 1$), and the post-treatment period is 2014 ($t = 2$).

In the context of Medicaid expansion, the outcome of interest, observed at the county level, is the crude mortality rate for people aged 20-64 (measured per 100,000). In our conditional DiD analysis, we also include the percentages of a county's population that are

[24] Under the linear relaxation of the martingale assumption, the yearly persistence in the pre-treatment period, $\tilde{\rho}_1$, can be estimated by regressing $\ddot{Y}_{i1975}$ on $\ddot{Y}_{i1974}$. The resulting estimate is $\widehat{\tilde{\rho}}_1 = 0.827$. Adjusting for the difference in periodicity yields $\widehat{\rho}_1 = (\hat{\tilde{\rho}}_1)^3 = 0.566$.

[25] The data are currently available in the following GitHub repository `https://github.com/pedrohcgs/JEL-DiD` and will soon be posted on OPENICPSR as part of the official replication package.

female, white, or Hispanic; the unemployment rate; the poverty rate; and county-level median income (in thousands of dollars)—all measured in 2012—in our regression adjustment.[26] All estimates are weighted by county population in 2013.

We first examine the unconditional DiD estimate using 2013 and 2014, which equals −2.6 (s.e. 1.5), indicating a reduction in mortality due to Medicaid expansion that is statistically significant at the 10% level. Once we account for covariates, however, the results are no longer significant with a regression-adjusted DiD estimate of −2.1 (s.e. 2.2).

Before we proceed to the bias decomposition, we conduct the pre-trends tests. We find that unconditional and conditional pre-trends tests are not rejected at the 5% level, with differences in pre-trends of −2.8 (1.5) and −2.6 (s.e. 2.5), respectively. However, note that the pre-trends for the unconditional DiD are significant at the 10% level.

### 5.2.2 Decomposing the bias of DiD

We next present the sample analogues of the bias decomposition, as described in Section 5.1.2. For the unconditional DiD case, the sample analogue of the bias can be decomposed as follows

$$\begin{aligned}\widehat{\Delta}_{\text{post}} = \widehat{\Delta}_{\text{post}}\left(\Delta^{\text{sel}}_{\text{post}}, \rho_2\right) &= \Delta^{\text{sel}}_{\text{post}} + (\rho_2 - 1)(E_n[Y_{i1}|G_i = 1] - E_n[Y_{i1}|G_i = 0]), \\ &= \Delta^{\text{sel}}_{\text{post}} + (\rho_2 - 1)(-53.7).\end{aligned}$$

where −53.7 denotes the pre-treatment difference in means between the treatment and control group, statistically significant at the 1% level.

Figure 3a demonstrates that this substantial pre-treatment difference translates to the bias of DiD being very sensitive to martingale violations. When considering the benchmark values for $\rho_2$, however, we note that both $\hat{\rho}_1$ and $\hat{\rho}^0_2$ are fairly close to 1, and therefore, the bias of DiD due to martingale deviations is relatively small for those values. If one is willing to assume that $\Delta^{\text{sel}}_{\text{post}} = 0$, then our analysis suggests a positive bias of DiD for these benchmark values of $\rho_2$ (Figure 3a). If we instead assume that $\Delta^{\text{sel}}_{\text{post}} = \Delta^{\text{sel}}_{\text{pre}}$, then our analysis indicates a negative bias of DiD for these benchmark values.

Once we adjust for covariates, the pre-treatment difference is no longer significant. Indeed, it is a negligible difference yielding the following sample analogue of the bias of the

[26] We use covariate values from 2012, so we can treat them as time-invariant in our analysis, simplifying the exposition and avoiding the strong, possibly unrealistic assumption that our covariates are strictly exogenous in this application. In line with Callaway and Sant'Anna (2021)'s implementation in the `did R` package, Baker et al. (2026) fixed covariate values at the 2013 values for post-treatment analysis, and at the 2012 values for pre-treatment periods, 2012-2013. We refer the reader to Caetano et al. (2022), as well as to Ghanem et al. (2026), for additional discussion.

regression-adjusted DiD

$$\widehat{\Delta}_{\text{post}} = \widehat{\Delta}_{\text{post}}\left(\Delta^{\text{sel}}_{\text{post}}, \rho_2\right) \quad = \quad \Delta^{\text{sel}}_{\text{post}} + (\rho_2 - 1)(-3.5).$$

As a result, the bias is insensitive to violations of the martingale condition and mostly driven by the magnitude of $\Delta^{\text{sel}}_{\text{post}}$, as illustrated in Figure 3b. If $\Delta^{\text{sel}}_{\text{post}} = 0$, then our analysis suggests that the bias of DiD is negligible, whereas if $\Delta^{\text{sel}}_{\text{post}} = \Delta^{\text{sel}}_{\text{pre}}$, then the bias is negative.

The bias component $\Delta^{\text{sel}}_{\text{post}}$ captures the bias due to selection on post-treatment unobservables. Selection on post-treatment unobservables is unlikely if the time-varying determinants of mortality are difficult to predict. This is the case, for example, if mortality is determined by factors such as adverse weather shocks and disease outbreaks which are arguably difficult to perfectly foresee, even one year ahead. In this case, a researcher can argue that $\Delta^{\text{sel}}_{\text{post}}$ is close to zero, which implies that the bias of DiD is negligible.

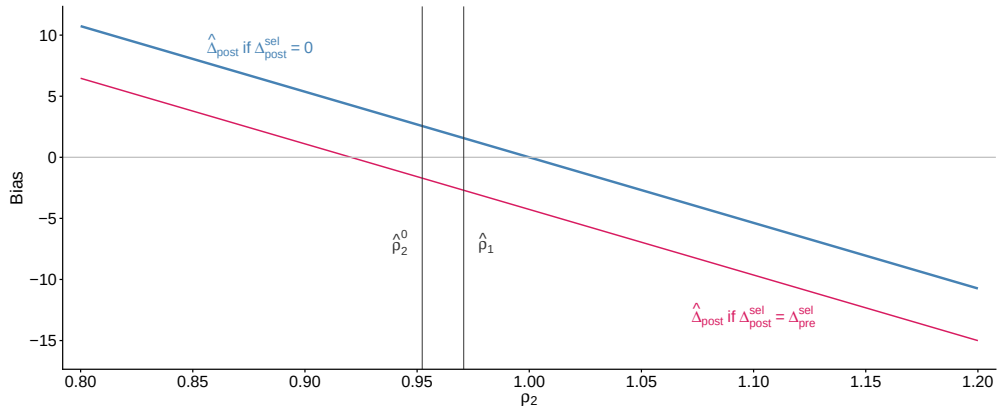
## 6 Implications for empirical practice

In this paper, we study parallel trends assumptions through the lens of selection into treatment. We derive necessary and sufficient conditions that clarify the empirical content of parallel trends, shed light on the trade-offs between assumptions on selection and time series restrictions, motivate DiD bias decompositions and benchmarking strategies, and provide theory-based templates for assessing and justifying parallel trends in applications with and without covariates. Below, we summarize the main implications of our results for practitioners.

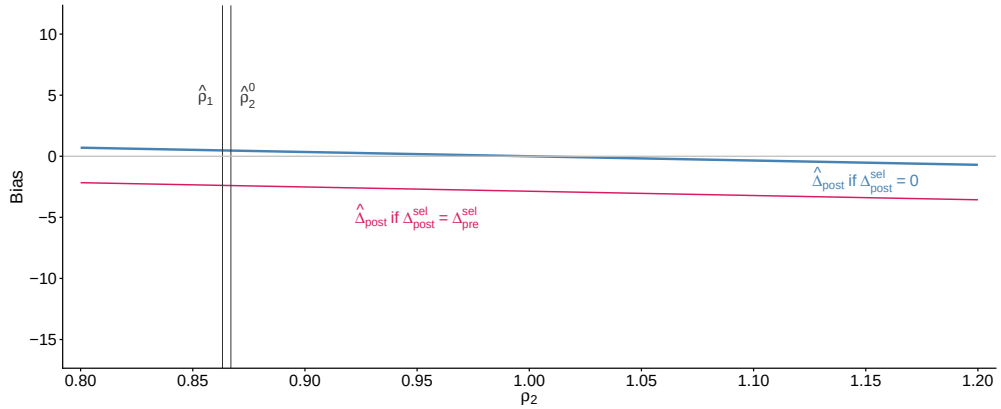
**Restrictions on selection are unavoidable in DiD designs.** The necessary and sufficient condition in Corollary 3.1 underscores that if researchers are not willing to impose any restrictions on selection, then parallel trends is equivalent to the untreated potential outcomes being constant over time up to deterministic mean shifts. Therefore, in realistic settings, relying on parallel trends assumptions implicitly imposes restrictions on the time-varying unobservables and how selection depends on them.

**Contextual and economic knowledge about selection can be used to assess and justify parallel trends.** Our analysis provides a general approach to derive necessary and sufficient conditions for parallel trends with and without covariates. Importantly, these conditions do not require the researchers to specify explicit selection mechanisms, which may be difficult in practice. Instead, the researchers only need to specify what the units select on. When doing so, it is crucial for researchers to consider the periodicity of the data, the timing of the selection decision, the information set available to the units, who make the

Figure 3: Benchmarking the Bias of DiD: Application to Medicaid Expansion




(a) Without covariates




(b) With covariates

*Notes:* Figure 3a displays the results from the bias decomposition without covariates. Figure 3b shows the results from the bias decomposition with regression adjustment using the percentages of a county's population that are female, white, or Hispanic, the unemployment rate, the poverty rate, and county-level median income (in thousands of dollars)—all measured in 2012. Data: Baker et al. (2026).

selection decision (e.g., the individuals themselves or caseworkers in the training program example), as well as at which level the selection decision is made (e.g., at the level of an individual economic agent vs. at the aggregate level via the political process).[27] Another practical byproduct of our analysis is a menu of selection-based templates for assessing and justifying parallel trends with and without covariates, see Section 3.3.5 and Appendix D.

**Selection-based bias decompositions are useful to sign and benchmark the bias of DiD.** In Section 4, we provide a general selection-based decomposition of the bias of DiD. We then apply this decomposition to settings where selection on pre-treatment unobservables is likely. Exploiting a martingale relaxation, we show that the bias of DiD can be decomposed

[27]The importance of the information available to units is underscored by the results in Marx et al. (2024), who study specific economic models of selection including learning and optimal stopping.

into two components: (i) the bias due to selection on post-treatment unobservables, (ii) the bias due to deviations from the martingale property necessary and sufficient for parallel trends under imperfect foresight. This characterization can be used in practice to sign and benchmark these two bias components, as we demonstrate in Section 5. A practical implication of this characterization is that the pre-treatment difference between the treatment and control group is a key determinant of the bias of DiD when parallel trends is violated.

## References


Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Altonji, J. G. and Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, 73(4):1053–1102.

Angrist, J. D. and Krueger, A. B. (1999). Chapter 23 - empirical strategies in labor economics. In Ashenfelter, O. C. and Card, D., editors, *Handbook of Labor Economics*, volume 3, pages 1277–1366. Elsevier.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Arellano, M. and Bonhomme, S. (2011). Nonlinear panel data analysis. *Annual Review of Economics*, 3:395–424.

Arellano, M. and Bonhomme, S. (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, 79(3):987–1020.

Arellano, M. and Bonhomme, S. (2016). Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal*, 19(3):C61–C94.

Arellano, M. and Honoré, B. (2001). Panel data models: Some recent developments. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 5. Elsevier Science.

Arkhangelsky, D. and Imbens, G. W. (2022). Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674.

Arkhangelsky, D., Imbens, G. W., Lei, L., and Luo, X. (2021). Double-robust two-way-fixed-effects regression for panel data. *arXiv:2107.13737 [econ]*.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 60(1):47–57.

Ashenfelter, O. C. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, 116(536):1716–1730.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., and Sant'Anna, P. (2026). Difference-in-differences designs: A practitioner's guide. *Journal of Economic Literature*.

Ban, K. and Kédagni, D. (2023). Generalized difference-in-differences models: Robust bounds. arXiv preprint arXiv:2211.06710.

Bester, C. A. and Hansen, C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business and Economic Statistics*, 27(2):235–250.

Blundell, R. and Dias, M. C. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 44(3):565–640.

Bonhomme, S. (2025). Back to feedback dynamics and heterogeneity in panel data. Working Paper.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies*, 91(6):3253–3285.

Caetano, C., Callaway, B., Payne, S., and Rodrigues, H. S. (2022). Difference in Differences with Time-Varying Covariates. *arXiv:2202.02903*.

Callaway, B. and Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review*, 43(2):245–257.

Card, D. and Hyslop, D. R. (2005). Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica*, 73(6):1723–1770.

Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4):772–793.

Chabé-Ferret, S. (2015). Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *Journal of Econometrics*, 185(1):110–123.

Chabé-Ferret, S. (2025). Should we combine difference in differences with conditioning on pre-treatment outcomes? Working Paper.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18(1):5–46.

Chamberlain, G. (1984). Chapter 22: Panel data. In *Handbook of Econometrics*, volume 2, pages 1247–1318. Elsevier.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, 60(3):567–596.

Chamberlain, G. (2022). Feedback in panel data models. *Journal of Econometrics*, 226(1):4–20. Annals Issue in Honor of Gary Chamberlain.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

de Chaisemartin, C. and D'Haultfœuille, X. (2018). Fuzzy Differences-in-Differences. *The Review of Economic Studies*, 85(2):999–1028.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996.

de Chaisemartin, C. and D'Haultfœuille, X. (2022). Not all differences-in-differences are equally compatible with outcome-based selection models. SSRN Working Paper.

de Chaisemartin, C. and D'Haultfœuille, X. (2023). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey. *Econometrics Journal*, 26(3):C1–C30.

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Dehejia, R. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161.

Ding, P. and Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, 27(4):605–615.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*, 1.

Freyberger, J. (2017). Non-parametric Panel Data Models with Interactive Fixed Effects. *The Review of Economic Studies*, 85(3):1824–1851.

Gardner, J. (2021). Two-stage differences in differences. *Working Paper*.

Ghanem, D. (2017). Testing identifying assumptions in nonseparable panel data models. *Journal of Econometrics*, 197(2):202–217.

Ghanem, D., Sant'Anna, P. H., and Wüthrich, K. (2026). When should pre-trends be parallel? Prepared for AEA Papers and Proceedings 2026, https://psantanna.com/files/GSW2026_AEAPP.pdf.

Graham, B. S. and Powell, J. L. (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, B. E. (2022). *Econometrics*. Princeton University Press.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4):605–654.

Heckman, J. J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1):239–267.

Heckman, J. J. and Vytlacil, E. J. (2007). Chapter 71 econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. volume 6 of *Handbook of Econometrics*, pages 4875–5143. Elsevier.

Hoderlein, S. and White, H. (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, 168(2):300–314.

Honoré, B. and Kyriazidou, E. (2000a). Estimation of Tobit-type models with individual specific effects. *Econometric Reviews*, 19:341–366.

Honoré, B. E. (1993). Orthogonality conditions for Tobit models with fixed effects and lagged dependent variables. *Journal of Econometrics*, 59(1–2):35–61.

Honoré, B. E. and Kyriazidou, E. (2000b). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68(4):839–874.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65(6):1335–1364.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620.

Lechner, M. (2010). The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends in Econometrics*, 4(3):165–224.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55(2):357–362.

Manski, C. F. and Pepper, J. V. (2018). How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *The Review of Economics*

*and Statistics*, 100(2):232–244.

Marcus, M. and Sant'Anna, P. H. C. (2021). The role of parallel trends in event study settings: An application to environmental economics. *Journal of the Association of Environmental and Resource Economists*, 8(2):235–275.

Marx, P., Tamer, E., and Tang, X. (2024). Parallel trends and dynamic choices. *Journal of Political Economy Microeconomics*, 2(1):129–171.

McKenzie, D. (2022). A new synthesis and key lessons from the recent difference-in-differences literature. World Bank Blogs (Link). Accessed: 2022-02-22.

Meyer, B. D., Viscusi, W. K., and Durbin, D. L. (1995). Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *The American Economic Review*, 85(3):322–340.

Miller, S., Johnson, N., and Wherry, L. R. (2021). Medicaid and mortality: New evidence from linked survey and administrative data*. *The Quarterly Journal of Economics*, 136(3):1783–1829.

Mundlak, Y. (1961). Empirical production function free of management bias. *Journal of Farm Economics*, 43(1):44–56.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Nicholls, D. and Quinn, B. (1982). *Random Coefficient Autoregressive Models: An Introduction*. Lecture Notes in Statistics. Springer New York.

Rambachan, A. and Roth, J. (2023). A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, 90(5):2555–2591.

Regis, M., Serra, P., and van den Heuvel, E. R. (2022). Random autoregressive models: A structured overview. *Econometric Reviews*, 41(2):207–230.

Roth, J. (2022). Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends. *American Economic Review: Insights*, 4(3):305–322.

Roth, J. and Sant'Anna, P. H. C. (2023). When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

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

Sant'Anna, P. H. C. and Zhao, J. (2020b). DRDID: Doubly robust difference-in-differences. R package version 1.0.6, https://psantanna.com/DRDID/.

Sommers, B. D. and Epstein, A. M. (2013). U.S. Governors and the Medicaid Expansion —

No Quick Resolution in Sight. *New England Journal of Medicine*, 368(6):496–499.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Verdier, V. (2020). Average treatment effects for stayers with correlated random coefficient models of panel data. *Journal of Applied Econometrics*, 35(7):917–939.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. The MIT Press, Cambridge, MA and London, England.

Wooldridge, J. M. (2025). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. *Empirical Economics*, 69:2545–2587.

# Appendix (for online publication)



## A Disaggregate data and aggregate decisions

In some DiD applications, the data are available at the disaggregate level (e.g., at the individual or firm level), while the decision to select into the treatment is made at the

aggregate level (e.g., at the county or state level). The results in the main text directly apply to such settings by interpreting $i$ as indexing the aggregate unit making the selection decision and the unobservables and potential outcomes as aggregate quantities. However, to justify restrictions about selection into treatment, it can be helpful to be more explicit about how selection at the aggregate level is related to the disaggregate level. In the following, we provide a formal framework for doing so. A leading example is when aggregate decisions are based on aggregating preferences at the disaggregate level (e.g., based on voting mechanisms).

Consider a canonical DiD setting with $S$ groups, indexed by $s \in \{1, \ldots, S\}$. Each group contains $n_s$ units, indexed by $i \in \{1, \ldots n_s\}$. To simplify the exposition, suppose that all groups are the same size, $n_s = n$ for $s \in \{1, \ldots, S\}$. Following the analysis in the main text, we impose general nonseparable models for the disaggregate potential outcomes,

$$Y_{ist}(0) = \xi_{st}(\alpha_{is}, \varepsilon_{ist}).$$

The aggregate potential outcomes are given by

$$Y_{st}(0) = A_{Y(0)}(Y_{1st}(0), \ldots, Y_{nst}(0)),$$

where $A_{Y(0)}(\cdot)$ is a potentially nonlinear aggregation function that can depend on $n$. A simple example is when the aggregate outcomes are averages of the disaggregate outcomes, $Y_{st}(0) = n^{-1} \sum_{i=1}^{n} Y_{ist}(0)$.

Consider a sharp DiD setting in which the treatment decisions are made at the group level, so that $G_s = G_{is}$ for all $i \in \{1, \ldots, n\}$, and researchers rely on parallel trends at the group level,

$$E[Y_{s2}(0) - Y_{s1}(0)|G_s = 1] = E[Y_{s2}(0) - Y_{s1}(0)|G_s = 0]. \tag{10}$$

The aggregate selection decision can depend on all unit-level unobservables,

$$G_s = g(\omega_s, \nu_s). \tag{11}$$

Here $\omega_s$ is a function or subvector of $(\alpha_s, \varepsilon_{s1}, \varepsilon_{s2}, \mu_s, \eta_{s1}, \eta_{s2})$, where $\alpha_s = (\alpha_{1s}, \ldots, \alpha_{ns})$, $\varepsilon_{s1} = (\varepsilon_{1s1}, \ldots, \varepsilon_{ns1})$, and $\varepsilon_{s2} = (\varepsilon_{1s2}, \ldots, \varepsilon_{ns2})$. The vectors $\mu_s = (\mu_{1s}, \ldots, \mu_{ns})$, $\eta_{s1} = (\eta_{1s1}, \ldots, \eta_{ns1})$, and $\eta_{s2} = (\eta_{1s2}, \ldots, \eta_{ns2})$ contain additional time-invariant and time-varying unobservables. The scalar unobservable $\nu_s$ captures determinants of selection that are independent of $\omega_s$.

All results in the main text directly apply in this setting with $i$ replaced by $s$, such that there are no additional theoretical complications. However, being explicit about the disaggregate level can help "microfound" restrictions on the aggregate selection mechanism,

as we illustrate in the following example.

**Example A.1** (Simple majority voting)**.** *Suppose that the aggregate selection decision is based on simple majority voting. Each unit submits a vote* $V_{is} \in \{0, 1\}$*,*

$$V_{is} = v(\omega_{is}, \nu_{is}). \tag{12}$$

*where* $\omega_{is}$ *is a function or subvector of* $(\alpha_{is}, \varepsilon_{is1}, \varepsilon_{is2}, \mu_{is}, \eta_{is1}, \eta_{is2})$*. The voting mechanism* (12) *accommodates voting based on group-level unobservables and outcomes since the additional unobservables* $(\mu_{is}, \eta_{is1}, \eta_{is2})$ *are unrestricted and can contain group-level quantities. Votes can be based on potential outcomes, expected gains, and fixed effects (as in Examples 2.1, 2.2, and 2.3), or other considerations.*

*The aggregate selection decision under simple majority voting is*

$$G_s = 1\left\{\frac{1}{n}\sum_{i=1}^{n} V_{is} \geq 0.5\right\}. \tag{13}$$

*This selection mechanism is a special case of mechanism* (11)*. Restrictions on the aggregate mechanism* (13) *can be directly motivated based on assumptions on the units' voting behavior, their information sets, and discount factors.* □

## B Multiple periods and multiple groups

Here we generalize our results to DiD designs with multiple periods and multiple groups. The setup and notation are based on Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Roth et al. (2023).

Let $t \in \{1, 2, \ldots, T\}$ index the periods. Suppose that at time $t = 1$, no units are treated, at $t = 2$, some units become treated, while others remain untreated, and so on. Previously treated units remain treated for all periods. Units can be categorized based on their treatment adoption pattern $D_i = (D_{i1}, \ldots, D_{iT})$. We define the group indicator $G_i$ as the first period in which units are treated, $G_i = \min\{t \in \{1, \ldots, T\} : D_{it} = 1\}$, and set $G_i = \infty$ for the never-treated units so that $G_i \in \{2, \ldots, T, \infty\}$.[28]

Potential outcomes are indexed by the entire treatment sequence $(d_1, \ldots, d_T) \in \{0, 1\}^T$, $Y_{it}(d_1, \ldots, d_T)$. Since treatment is an absorbing state, the potential outcomes can be indexed by the first treatment period only. Define $Y_{it}(g) = Y_{it}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ for $g \in \{2, \ldots, T\}$ and $Y_{it}(\infty) = Y_{it}(\mathbf{0}_T)$, where $\mathbf{0}_s \equiv (0, \ldots, 0) \in \mathbb{R}^s$ and $\mathbf{1}_s \equiv (1, \ldots, 1) \in \mathbb{R}^s$. We maintain a standard no-anticipation assumption (e.g., Roth et al., 2023).

[28] Since $G_i$ is a random variable with finite support, we emphasize that $\{\infty\}$ is merely a label.

**Assumption NA.** *For $g \in \{2, \ldots, T, \infty\}$ and $t < g$, $Y_{it}(g) = Y_{it}(\infty)$.*

Our objects of interest are the group-time ATTs,

$$\text{ATT}(g,t) = E[Y_{it}(g) - Y_{it}(\infty)|G_i = g]. \tag{14}$$

We impose the following parallel trends assumption to identify the $\text{ATT}(g,t)$.[29]

**Assumption PT-MP.** *For $(g,t) \in \{2, \ldots, T\}^2$,*

$$E[Y_{it}(\infty) - Y_{i(t-1)}(\infty) \,|\, G_i = g] = E[Y_{it}(\infty) - Y_{i(t-1)}(\infty)|G_i = \infty] \tag{15}$$

We consider a general nonseparable outcome model,

$$Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it}), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T.$$

Selection into treatment can depend on the unobservable determinants of $Y_{it}(\infty)$ as well as additional unobservables,

$$G_i = g(\omega_i, \nu_i),$$

where $\omega_i$ is a function or a subvector of $(\alpha_i, \varepsilon_{i1}, \ldots, \varepsilon_{iT}, \mu_i, \eta_{i1}, \ldots, \eta_{iT})$

**Assumption SEL-MP.** *Assume that $\nu_i \perp\!\!\!\perp (\alpha_i, \varepsilon_{i1}, \ldots, \varepsilon_{iT}, \mu_i, \eta_{i1}, \ldots, \eta_{iT})$. Furthermore, there exists a non-overlapping partition of the support of $v_i$, $\{B_g\}_{g \in \{2,\ldots,T,\infty\}}$, such that $P(v_i \in B_g) \in (0,1)$ for $g \in \{2, \ldots, T, \infty\}$.*

The following theorem presents the necessary and sufficient condition for Assumption PT-MP.

**Theorem B.1** (Necessary and sufficient condition for PT-MP to hold for all $g \in \mathcal{G}_\omega$)**.** *Suppose that Assumptions NA and SEL-MP hold. Suppose further that for $t \in \{2, \ldots, T\}$ either $P(E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty)|\omega_i] > 0) < 1$ or $P(E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty)|\omega_i] < 0) < 1$. Then, Assumption PT-MP holds for all $g \in \mathcal{G}_\omega$ satisfying $P(G_i = g) \in (0,1)$ for $g \in \{2, \ldots, T, \infty\}$ if and only if $E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty)|\omega_i] = 0$ a.s. for $t \in \{2, \ldots, T\}$.*

Following the logic of Section 3, Theorem B.1 implies necessary and sufficient conditions for many different classes of selection mechanisms. For example, if selection is unrestricted,

[29] In our setting, this parallel trends assumption corresponds to the ones made by Callaway and Sant'Anna (2021), Gardner (2021), Sun and Abraham (2021), Borusyak et al. (2024), and Wooldridge (2025); see also de Chaisemartin and D'Haultfœuille (2020) and Marcus and Sant'Anna (2021) for related assumptions.

so that $\omega_i = (\alpha_i, \varepsilon_{i1}, \ldots, \varepsilon_{iT}, \mu_i, \eta_{i1}, \ldots, \eta_{iT})$, we obtain the following necessary and sufficient condition as a corollary of Theorem B.1,

$$\dot{Y}_{i1}(\infty) = \dot{Y}_{i2}(\infty) = \cdots = \dot{Y}_{iT}(\infty),$$

which is a natural extension of Corollary 3.1.

Similarly, we can consider various classes of restricted selection mechanisms. For example, if $\omega_i = (\alpha_i, \mu_i)$, we obtain the following necessary and sufficient condition,

$$E[\dot{Y}_{i1}(\infty)|\alpha_i, \mu_i] = \cdots = E[\dot{Y}_{iT}(\infty)|\alpha_i, \mu_i],$$

which is a multi-period version of Corollary 3.5. There are many other corollaries of Theorem B.1 depending on the choice of $\omega_i$.

# C Sufficient conditions for parallel trends beyond Theorem 3.1

Theorem 3.1 allows researchers to derive a broad set of sufficient (and necessary) conditions for parallel trends, depending on what unit select on. However, in some applications, researchers might be interested in imposing other types of restrictions not covered by Theorem 3.1.

To illustrate, consider the separable model (5). The following is a sufficient condition for parallel trends based on a symmetry condition on the selection mechanism:

(e) Symmetric selection:

(i) $g$ is a symmetric function in $\varepsilon_{i1}$ and $\varepsilon_{i2}$

(ii) $\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i \overset{d}{=} \varepsilon_{i2}, \varepsilon_{i1}|\alpha_i$ and $(\mu_i, \eta_{i1}, \eta_{i2})|\alpha_i, \varepsilon_{i1}, \varepsilon_{i2} \overset{d}{=} (\mu_i, \eta_{i1}, \eta_{i2})|\alpha_i, \varepsilon_{i2}, \varepsilon_{i1}$

Condition (e) can be shown to imply parallel trends. In addition to the symmetry of the selection mechanism, this sufficient condition imposes two different types of exchangeability restrictions. First, it requires that the conditional distribution of $(\mu_i, \eta_{i1}, \eta_{i2})$ is exchangeable in $\varepsilon_{i1}$ and $\varepsilon_{i2}$ after conditioning on $\alpha_i$. This notion of exchangeability has been employed, for example, in Altonji and Matzkin (2005). Second, it requires the distribution of $(\varepsilon_{i1}, \varepsilon_{i2})$ to be exchangeable conditional on $\alpha_i$.

To illustrate condition (e), consider the selection mechanism in Example 2.1 and suppose that the units have complete information, so that $\omega_i = (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \kappa_{i2})$. In this case, the selection mechanism (2) simplifies to $G_i = 1\{\alpha_i(1+\beta) + \varepsilon_{i1} + \beta\varepsilon_{i2} \leq \kappa_{i2}\}$. This selection mechanism is symmetric if there is no discounting ($\beta = 1$), which may be plausible if the

time span between the pre- and post-treatment period is short. In addition, the condition (ii) requires $(\varepsilon_{i1}, \varepsilon_{i2})$ to be exchangeable and the conditional distribution of the costs to be symmetric in $\varepsilon_{i1}$ and $\varepsilon_{i2}$, $\kappa_{i2}|\alpha_i, \varepsilon_{i1}, \varepsilon_{i2} \overset{d}{=} \kappa_{i2}|\alpha_i, \varepsilon_{i2}, \varepsilon_{i1}$.

# D Covariates

In many applications, parallel trends may only be plausible conditional on covariates (e.g., Heckman et al., 1997; Abadie, 2005; Sant'Anna and Zhao, 2020a; Callaway and Sant'Anna, 2021). In this appendix, we therefore study the role of covariates through the lens of selection into treatment. While many existing approaches focus on time-invariant covariates, we explicitly allow for a vector of both time-invariant and time-varying covariates, $X_{it}$, assuming that $X_{it}$ is not affected by the treatment.

In Appendix D.1, we derive necessary and sufficient conditions for parallel trends to hold conditional on covariates and show that these conditions imply separability restrictions on how covariates enter the outcome model. In Appendix D.3, we propose a modified parallel trends assumption that accommodates nonseparable models and provide sufficient conditions for this modified parallel trends assumption.

## D.1 Covariates and the role of separability

Suppose that parallel trends holds conditional on the time series of covariates, $X_i = (X_{i1}, X_{i2})$.

**Assumption PT-X.** *The conditional parallel trends assumption holds:*

$$E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, X_i] = E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0, X_i] \ a.s.$$

Under Assumption PT-X, the unconditional ATT is identified as

$$E[Y_{i2}(1) - Y_{i2}(0)|G_i = 1] = E\left[\mathrm{ATT}(X_i)|G_i = 1\right] = E\left[\mathrm{DiD}(X_i)|G_i = 1\right],$$

where $\mathrm{ATT}(X_i) \equiv E[Y_{i2}(1) - Y_{i2}(0)|G_i = 1, X_i]$ and $\mathrm{DiD}(X_i) \equiv E[Y_{i2} - Y_{i1}|G_i = 1, X_i] - E[Y_{i2} - Y_{i1}|G_i = 0, X_i]$.

In the presence of covariates, potential outcomes and selection into treatment may naturally depend on them. We therefore consider the following outcome model and selection mechanism,

$$\begin{aligned} Y_{it}(0) &= \xi_t(X_{it}, \alpha_i, \varepsilon_{it}), \quad i = 1, \ldots, n, \quad t = 1, 2, \\ G_i &= g(X_i, \omega_i, \nu_i), \quad i = 1, \ldots, n. \end{aligned}$$

Denote by $\mathcal{G}_\omega^x$ the set of all selection mechanisms $g$ mapping from the support of $(X_i, \omega_i, \nu_i)$ to $\{0, 1\}$.

The necessary and sufficient condition in Theorem 3.1 generalizes straightforwardly to Assumption PT-X as we show in the following theorem. Before we proceed, we introduce the notation $\ddot{Y}_{it}(0) \equiv Y_{it}(0) - E[Y_{it}(0)|X_i]$ and the following regularity condition on $\nu_i$.

**Assumption SEL-X.** *$P(\nu_i > c|X_i) \in (0, 1)$ for some $c \in \mathbb{R}$ and $\nu_i \perp\!\!\!\perp (X_i, \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \mu_i, \eta_{i1}, \eta_{i2})$.*

**Theorem D.1** (Necessary and sufficient condition for Assumption PT-X for all $g \in \mathcal{G}_\omega$)**.** *Suppose that Assumption SEL-X holds. Suppose further that either $P(E[\ddot{Y}_{i2}(0) - \ddot{Y}_{i1}(0)|\omega_i, X_i] > 0) < 1$ or $P(E[\ddot{Y}_{i2}(0) - \ddot{Y}_{i1}(0)|\omega_i, X_i] < 0) < 1$. Then, Assumption PT-X holds for all $g \in \mathcal{G}_\omega^x$ satisfying $P(G_i = 1|X_i) \in (0, 1)$ if and only if $E[\ddot{Y}_{i2}(0) - \ddot{Y}_{i1}(0)|\omega_i, X_i] = 0$ a.s.*

The proof follows from the same arguments as in the proof of Theorem 3.1, conditional on covariates $X_i$. Given the necessary and sufficient condition in Theorem D.1, the results in the main text generalize straightforwardly to settings with covariates.

An important practical implication of the necessary and sufficient condition in Theorem D.1 is that it implies separability requirements on how the covariates can enter the outcome model. To illustrate, consider the simple case where $\omega_i = \alpha_i$. In this case, the necessary and sufficient condition in Theorem D.1 can be written as

$$E[Y_{i2}(0)|X_i, \alpha_i] - E[Y_{i1}(0)|X_i, \alpha_i] = E[Y_{i2}(0)|X_i] - E[Y_{i1}(0)|X_i].$$

To illustrate the separability restrictions, consider a generalized random coefficient model (e.g., Chamberlain, 1992) in which $\alpha_i$ interacts with $X_{it}$,

$$Y_{it}(0) = \alpha_i \chi_t(X_{it}) + \lambda_t + \varepsilon_{it}. \tag{16}$$

Here $\chi_t(\cdot)$ is an arbitrary time-varying function. Even under the assumption that $E[\varepsilon_{it}|X_i, \alpha_i] = 0$, this model generally violates the necessary condition due to the combination of nonseparability between $\alpha_i$ and $X_{it}$ and the time variability in the structural function through $\chi_t(\cdot)$,

$$E[Y_{i2}(0)|X_i, \alpha_i] - E[Y_{i1}(0)|X_i, \alpha_i] = \alpha_i(\chi_2(X_{i2}) - \chi_1(X_{i1})) + \lambda_2 - \lambda_1.$$

This example demonstrates that for parallel trends to hold in the presence of interactions between covariates and $\alpha_i$, it is not sufficient to focus on subpopulations with $X_{i1} = X_{i2}$. We additionally require that the component that interacts with $\alpha_i$, $\chi_t(\cdot)$, does not vary across time.

Allowing for interactions between the unobservable determinants of selection and some covariates is important in applications. We therefore consider a weaker conditional parallel trends assumption that allows for such interactions in Section D.3.

### D.2 Selection-based templates for justifying parallel trends in separable models with covariates

The necessary and sufficient conditions with covariates can serve as theory-based templates for assessing and justifying parallel trends in DiD applications with covariates. The exact form of these conditions depends on the model for $Y_{it}(0)$. Consider, for example, the separable model

$$Y_{it}(0) = \alpha_i + \chi_t(X_{it}) + \lambda_t + \varepsilon_{it},$$

where $\chi_t(\cdot)$ is a potentially time-varying function. For this model, we can provide a menu of necessary and sufficient conditions that specify the determinants of selection as well as conditions on the idiosyncratic shocks $\varepsilon_{it}$. For instance,

(a) Imperfect foresight (case 1): (i) $\omega_i = (\alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1})$ and (ii) $E[\varepsilon_{i2} - \varepsilon_{i1}|X_i, \alpha_i, \varepsilon_{i1}, \mu_i, \eta_{i1}] = E[\varepsilon_{i2} - \varepsilon_{i1}|X_i]$

(b) Imperfect foresight (case 2): (i) $\omega_i = (\mu_i, \eta_{i1})$ and (ii) $E[\varepsilon_{i2} - \varepsilon_{i1}|X_i, \mu_i, \eta_{i1}] = E[\varepsilon_{i2} - \varepsilon_{i1}|X_i]$

(c) Roy-style selection: (i) $\omega_i = (\tau_{i2}, \kappa_{i2})$ and (ii) $E[\varepsilon_{i2} - \varepsilon_{i1}|X_i, \tau_{i2}, \kappa_{i2}] = E[\varepsilon_{i2} - \varepsilon_{i1}|X_i]$

(d) Selection on fixed effects: (i) $\omega_i = (\alpha_i, \mu_i)$ and (ii) $E[\varepsilon_{i2} - \varepsilon_{i1}|X_i, \alpha_i, \mu_i] = E[\varepsilon_{i2} - \varepsilon_{i1}|X_i]$

### D.3 A parallel trends assumption for nonseparable models

Motivated by Section D.1, we consider a weaker (than Assumption PT-X) conditional parallel trends assumption that accommodates nonseparable models. To state this assumption, we explicitly differentiate between two types of covariates: (i) $X_{it}^{\mu}$ are covariates that interact with the unobservable determinants of selection in the outcome model; (ii) $X_{it}^{\lambda}$ are covariates that do not interact with these unobservables. Both types of covariates can enter the selection mechanism in an arbitrary way. The following conditional parallel trends assumption holds for subpopulations that experience no change in $X_{it}^{\mu}$ and the same trajectory in $X_{it}^{\lambda}$.

**Assumption PT-NSP.** *The (modified) conditional parallel trends assumption holds:*

$$E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, X_i^{\lambda}, X_{i1}^{\mu} = X_{i2}^{\mu}] = E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0, X_i^{\lambda}, X_{i1}^{\mu} = X_{i2}^{\mu}] \ a.s.$$

Under Assumption PT-NSP, we can no longer identify the ATT, $E[Y_{i2}(1)-Y_{i2}(0)|G_i=1]$, because we cannot identify the conditional ATT, $E[Y_{i2}(1)-Y_{i2}(0)|G_i=1,X_i^\lambda,X_i^\mu]$. Instead, we can identify $E[Y_{i2}(1)-Y_{i2}(0)|G_i=1,X_i^\lambda,X_{i1}^\mu=X_{i2}^\mu]$.[30] After integrating out with respect to the distribution of covariates, we can identify the ATT for subpopulations that do not experience changes in $X_{it}^\mu$,

$$E[Y_{i2}(1)-Y_{i2}(0)|G_i=1,X_{i1}^\mu-X_{i2}^\mu=0].$$

Note that if $X_{it}^\mu$ is time-invariant, then $X_{i1}^\mu=X_{i2}^\mu$ holds by definition and Assumptions PT-X and PT-NSP are equivalent.

Next, we provide different sets of sufficient conditions for Assumption PT-NSP under the following nonseparable outcome model.[31]

**Assumption NSP-X.**

$$Y_{it}(0)=\mu(X_{it}^\mu,\alpha_i^\mu,\varepsilon_{it}^\mu)+\lambda_t(X_{it}^\lambda,\alpha_i^\lambda,\varepsilon_{it}^\lambda),\quad i=1,\ldots,n,\quad t=1,2,$$

*where $X_{it}^\mu$, $X_{it}^\lambda$, $\alpha_i^\mu$, $\alpha_i^\lambda$, $\varepsilon_{it}^\mu$, and $\varepsilon_{it}^\lambda$ are finite-dimensional random vectors.*

Let $\mathcal{X}_\mu$, $\mathcal{X}_\lambda$, $\mathcal{A}$, and $\mathcal{E}$ denote the supports of $X_{it}^\mu$, $X_{it}^\lambda$, $\alpha_i^\mu$, and $\varepsilon_{it}^\mu$, respectively.

It is natural to consider selection based on the unobservables entering $\mu(\cdot)$. We therefore impose the following condition on the projected selection mechanism.

**Assumption SEL-CI.**

$$E[G_i|X_{i1}^\mu,X_{i2}^\mu,X_{i1}^\lambda,X_{i2}^\lambda,\alpha_i^\mu,\alpha_i^\lambda,\varepsilon_{i1}^\mu,\varepsilon_{i2}^\mu,\varepsilon_{i1}^\lambda,\varepsilon_{i2}^\lambda]=E[G_i|X_{i1}^\mu,X_{i2}^\mu,X_{i1}^\lambda,X_{i2}^\lambda,\alpha_i^\mu,\varepsilon_{i1}^\mu,\varepsilon_{i2}^\mu].$$

Assumption SEL-CI allows the projected selection mechanism to depend on all covariates, but only on the unobservables that enter $\mu(\cdot)$. In view of Assumption SEL-CI, we define

$$\begin{aligned}&\bar{g}(x_1^\mu,x_2^\mu,x_1^\lambda,x_2^\lambda,a^\mu,e_1^\mu,e_2^\mu)\\&\equiv E[G_i|X_{i1}^\mu=x_1^\mu,X_{i2}^\mu=x_2^\mu,X_{i1}^\lambda=x_1^\lambda,X_{i2}^\lambda=x_2^\lambda,\alpha_i^\mu=a^\mu,\varepsilon_{i1}^\mu=e_1^\mu,\varepsilon_{i2}^\mu=e_2^\mu].\end{aligned}$$

In the following, we present different sets of sufficient conditions for Assumption PT-NSP, inspired by the sufficient conditions in Section 3.3.5 and Appendix C. For concreteness, we

[30] With a slight abuse of notation, we use $(X_{i1}^\mu=X_{i2}^\mu)$ in the conditioning set as a short-hand for $(X_{i1}^\mu,X_{i2}^\mu=X_{i1}^\mu)$.

[31] Without further restrictions on the unobservables, the additive structure is without loss of generality and the superscripts $\mu$ and $\lambda$ are merely labels. Indeed, if $X_{it}^\mu=X_{it}^\lambda$, $\alpha_i^\mu=\alpha_i^\lambda$, and $\varepsilon_{it}^\mu=\varepsilon_{it}^\lambda$, the model is fully nonseparable and time-varying in an arbitrary way.

focus on three sets of sufficient conditions; other sets of sufficient conditions could also be considered. Each set of conditions consists of assumptions on the projected selection mechanism as well as distributional restrictions on the unobservables. Our first sufficient condition allows selection to depend on all covariates as well as the unobservables that enter the time-invariant component of the structural function, while imposing a symmetry restriction on the projected selection mechanism.

**Assumption SC1-NSP.** *The following conditions hold:*

*(i)* $\bar{g}(x_1^\mu, x_2^\mu, x_1^\lambda, x_2^\lambda, a^\mu, e_1^\mu, e_2^\mu)$ *is a symmetric function in* $e_1^\mu$ *and* $e_2^\mu$.

*(ii)* $(\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)|X_i^\mu, X_i^\lambda, \alpha_i^\mu \overset{d}{=} (\varepsilon_{i2}^\mu, \varepsilon_{i1}^\mu)|X_i^\mu, X_i^\lambda, \alpha_i^\mu$.

*(iii)* $(\alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu) \perp\!\!\!\perp (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda)|X_i^\mu, X_i^\lambda$.

Here we require the conditional distribution of $(\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)|X_i^\mu, X_i^\lambda, \alpha_i^\mu$ to be exchangeable. Since the projected selection mechanism depends on $(\alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)$, we require them to be independent of the unobservables entering $\lambda_t(\cdot)$ conditional on $(X_i^\mu, X_i^\lambda)$.

The exchangeability restriction in Assumption SC1-NSP is different from the exchangeability assumption in Altonji and Matzkin (2005). The exchangeability assumption in Altonji and Matzkin (2005) requires the conditional distribution of all unobservables that enter $\mu(\cdot)$ and $\lambda_t(\cdot)$ to be invariant to permutations of covariates in the conditioning set, which is a nonparametric correlated random effects restriction (Ghanem, 2017). By contrast, we assume that the time-varying unobservables are exchangeable conditional on $(X_i^\mu, X_i^\lambda, \alpha_i^\mu)$ without imposing any restrictions on the distribution of $\alpha_i^\mu|G_i, X_i^\mu, X_i^\lambda$.

Next, we consider a projected selection mechanism that is a trivial function of $\varepsilon_{i2}^\mu$ in the following sufficient condition.

**Assumption SC2-NSP.** *The following conditions hold:*

*(i)* $\bar{g}(x_1^\mu, x_2^\mu, x_1^\lambda, x_2^\lambda, a^\mu, e_1^\mu, e_2^\mu)$ *is a trivial function of* $e_2^\mu$.

*(ii)* $(\alpha_i^\mu, \varepsilon_{i1}^\mu) \perp\!\!\!\perp \Delta_{\mu,i}|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu$, *where* $\Delta_{\mu,i} \equiv \mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)$.

*(iii)* $(\alpha_i^\mu, \varepsilon_{i1}^\mu) \perp\!\!\!\perp (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda)|X_i^\mu, X_i^\lambda$.

Assumption SC2-NSP.ii implicitly imposes separability conditions on $\mu(\cdot)$ (but not on $\lambda_t(\cdot)$) and restrictions on time series dependence.[32] The independence condition in Assumption SC2-NSP.iii requires that the unobservable determinants of selection are independent of the unobservables that enter $\lambda_t(\cdot)$ conditional on the time series of covariates.

[32]To see this, note that since $\Delta_{\mu,i} = \mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)$, for $\Delta_{\mu,i}$ to be conditionally independent of $(\alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)$, a sufficient condition would be that $\Delta_{\mu,i}$ is separable in $\alpha_i^\mu$ and $\varepsilon_{it}^\mu$, such that $\Delta_{\mu,i} = \gamma(X_{i2}^\mu) - \gamma(X_{i1}^\mu) + \varepsilon_{i2}^\mu - \varepsilon_{i1}^\mu$, as well as that $(\varepsilon_{i2}^\mu - \varepsilon_{i1}^\mu)$ and $(\alpha_i^\mu, \varepsilon_{i1}^\mu)$ are conditionally independent.

The last sufficient condition restricts the projected selection mechanism to only depend on covariates and the time-invariant unobservables.

**Assumption SC3-NSP.** *The following conditions hold:*

*(i)* $\bar{g}(x_1^\mu, x_2^\mu, x_1^\lambda, x_2^\lambda, a^\mu, e_1^\mu, e_2^\mu)$ *is a trivial function of* $e_1^\mu$ *and* $e_2^\mu$.

*(ii)* $\varepsilon_{i1}^\mu | X_i^\mu, X_i^\lambda, \alpha_i^\mu \overset{d}{=} \varepsilon_{i2}^\mu | X_i^\mu, X_i^\lambda, \alpha_i^\mu$.

*(iii)* $\alpha_i^\mu \perp\!\!\!\perp (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda) | X_i^\mu, X_i^\lambda$.

Assumption SC3-NSP requires the distribution of $\varepsilon_{it}^\mu$, which enters $\mu(\cdot)$, to be time-invariant conditional on $(\alpha_i^\mu, X_i^\mu, X_i^\lambda)$. The unobservables entering $\lambda_t(\cdot)$, $(\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda)$, are required to be independent of the unobservables that determine selection, $\alpha_i^\mu$, conditional on $(X_i^\mu, X_i^\lambda)$.

Each of the sufficient conditions consists of three components: (i) a restriction on how/which unobservables enter the projected selection mechanism, (ii) a restriction on the unobservables entering the time-invariant component of the structural function, and (iii) an independence assumption that ensures that the time-varying component of the structural function is independent of $G_i$ conditional on the time series of covariates.

The following proposition formally establishes sufficiency of each set of conditions.

**Proposition D.1** (Sufficient conditions)**.** *Suppose that Assumptions NSP-X and SEL-CI hold and* $P(G_i = 1 | X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu) \in (0, 1)$ *a.s. Then (i) Assumption SC1-NSP implies Assumption PT-NSP, (ii) Assumption SC2-NSP implies Assumption PT-NSP, and (iii) Assumption SC3-NSP implies Assumption PT-NSP.*

**Remark D.1** (Connection to unconfoundedness)**.** *All sufficient conditions in Proposition D.1 allow for selection on unobservable determinants of the untreated potential outcome. This is in contrast to the unconfoundedness assumptions commonly used in cross-sectional studies (e.g., Imbens, 2004; Imbens and Wooldridge, 2009). Therefore, these results elucidate the differences between conditional parallel trends and unconfoundedness-type assumptions.* □

## E Connections to identification assumptions in panel models

Here we relate the selection-based sufficient conditions in Section PT-NSP to the identification assumptions in the nonseparable panel literature.[33] The literature on nonseparable

[33] See, e.g., Altonji and Matzkin (2005); Athey and Imbens (2006); Bester and Hansen (2009); Hoderlein and White (2012); Chernozhukov et al. (2013); Arellano and Bonhomme (2016); Ghanem (2017). This work extends notions of fixed effects and correlated random effects that originated in the linear model (Mundlak, 1961, 1978; Chamberlain, 1982, 1984). Surveys (Arellano and Honoré, 2001; Arellano and Bonhomme, 2011)

panel models has considered two broad categories of identification assumptions. First, time homogeneity conditions (e.g., Hoderlein and White, 2012; Chernozhukov et al., 2013) require the distribution of time-varying unobservables to be stationary across time while allowing for unrestricted individual heterogeneity (fixed effects). Second, nonparametric correlated random effects restrictions (e.g., Altonji and Matzkin, 2005; Bester and Hansen, 2009) allow for unrestricted time heterogeneity by imposing restrictions on individual heterogeneity, generalizing the classical notion of correlated random effects (e.g., Mundlak, 1978; Chamberlain, 1984). However, neither category of assumptions is explicit about the selection mechanism and, in particular, about how unobservables determine selection.

The existing identification results based on time homogeneity or correlated random effects assumptions suggest a trade-off between restrictions on time and individual heterogeneity. Here we show that our sufficient conditions for Assumption PT-NSP constitute interpretable primitive conditions on the selection mechanism that imply *combinations* of time homogeneity and correlated random effects restrictions from the nonseparable panel literature.

The following assumption is the time homogeneity assumption from Chernozhukov et al. (2013) imposed on $\varepsilon_{it}^{\mu}$ in Assumption NSP-X, conditional on the time series of all covariates that enter the outcome equation.

**Assumption TH.** $\varepsilon_{i1}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu} \overset{d}{=} \varepsilon_{i2}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}$

Assumption TH requires the distribution of $\varepsilon_{it}^{\mu}$ to be homogeneous across time conditional on $G_i$, $X_i^{\mu}$, $X_i^{\lambda}$, and $\alpha_i^{\mu}$. However, it does not impose any restrictions on the conditional distribution of $\varepsilon_{it}^{\mu}$. Furthermore, there are no restrictions imposed on the distribution of $\alpha_i^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}$, consistent with the notion of fixed effects.

The next assumption is a nonparametric correlated random effects assumption (e.g., Altonji and Matzkin, 2005; Ghanem, 2017).

**Assumption CRE.** $(\alpha_i^{\lambda}, \varepsilon_{i1}^{\lambda}, \varepsilon_{i2}^{\lambda})|G_i, X_i^{\mu}, X_i^{\lambda} \overset{d}{=} (\alpha_i^{\lambda}, \varepsilon_{i1}^{\lambda}, \varepsilon_{i2}^{\lambda})|X_i^{\mu}, X_i^{\lambda}$.

Assumption CRE is a conditional independence condition between $G_i$ and the unobservables that enter the time-varying component of the structural function, $\lambda_t(\cdot)$. This assumption does not imply conditional random assignment, $(Y_{i1}(0), Y_{i2}(0)) \perp\!\!\!\perp G_i|X_i^{\mu}, X_i^{\lambda}$, since selection into treatment can depend on the unobservables entering the time-invariant component $\mu(\cdot)$.

and textbook treatments (Arellano, 2003; Wooldridge, 2010) further describe the role of restrictions on time and individual heterogeneity in linear and nonlinear models. Such restrictions have been imposed in the context of identification in limited dependent variable models (e.g. Manski, 1987; Honoré, 1993; Kyriazidou, 1997; Honoré and Kyriazidou, 2000a,b) and random coefficient models (e.g. Chamberlain, 1992; Graham and Powell, 2012; Arellano and Bonhomme, 2012). Nonparametric identification of panel models with additivity restrictions has been examined, e.g., in Evdokimov (2010) and Freyberger (2017).

Together, Assumptions TH and CRE imply Assumption PT-NSP.[34]

**Proposition E.1** (TH and CRE imply PT-NSP)**.** *Suppose that Assumption NSP-X holds and* $P(G_i = 1|X_{i1}^{\mu} = X_{i2}^{\mu}, X_i^{\lambda}) \in (0,1)$ *a.s. Then Assumptions TH and CRE imply Assumption PT-NSP.*

In view of Proposition E.1, it is interesting to explore the connection between selection, time homogeneity, and correlated random effects in the nonseparable DiD framework. To this end, Proposition E.2 shows that Assumptions SC1-NSP and SC3-NSP are primitive sufficient conditions on the selection mechanism for the nonseparable model satisfying Assumptions TH and CRE.[35] In the following, with a slight abuse of notation, we let $\mathcal{A}$ denote the support of $\alpha_i^{\mu}$. For $t = 1, 2$, let $\mathcal{E}$ denote the support of $\varepsilon_{it}^{\mu}$, $\mathcal{X}_{\mu}$ the support of $X_{it}^{\mu}$ and $\mathcal{X}_{\lambda}$ the support of $X_{it}^{\lambda}$.

**Proposition E.2** (Connection between selection, time homogeneity, and correlated random effects)**.** *Suppose that Assumption NSP-X holds and* $G_i = g(X_{i1}^{\mu}, X_{i2}^{\mu}, X_{i1}^{\lambda}, X_{i2}^{\lambda}, \alpha_i^{\mu}, \varepsilon_{i1}^{\mu}, \varepsilon_{i2}^{\mu})$.[36] *Assume that the conditional density of* $(\varepsilon_{i1}^{\mu}, \varepsilon_{i2}^{\mu})|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}$, $f_{\varepsilon_{i1}^{\mu}, \varepsilon_{i2}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1, e_2|x^{\mu}, x^{\lambda}, a)$, *exists and the conditional density of* $\varepsilon_{it}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}$, $f_{\varepsilon_{it}|X_i^{\mu}, X_i^{\lambda}, \alpha_i}(e|x^{\mu}, x^{\lambda}, a)$, *is strictly positive for* $(e, x^{\mu}, x^{\lambda}, a) \in \mathcal{E} \times \mathcal{X}_{\mu}^2 \times \mathcal{X}_{\lambda}^2 \times \mathcal{A}$, $t = 1, 2$. *Then (i) Assumption SC1-NSP with* $g(\cdot)$ *in lieu of* $\bar{g}(\cdot)$ *implies Assumptions TH and CRE if* $P(G_i = 1|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}) \in (0,1)$ *a.s., (ii) Assumption SC3-NSP with* $g(\cdot)$ *in lieu of* $\bar{g}(\cdot)$ *implies Assumptions TH and CRE.*

Proposition E.2 demonstrates how restrictions on selection can be used to justify combinations of Assumptions TH and CRE.

# F Bias decomposition with covariates

Here we extend the analysis in Section 4 to the setup with covariates in Appendix D.

## F.1 Decomposition

The necessary and sufficient condition in Corollary 3.2 with covariates and one additional pretreatment period is $E[\ddot{Y}_{i2}(0)|X_i, \alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1] = \ddot{Y}_{i1}(0)$, where $\ddot{Y}_{it}(0) \equiv Y_{it}(0) - E[Y_{it}(0)|X_i]$.

[34] Ghanem (2017, Appendix B) discusses the nonparametric identification of the ATT through DiD either through time homogeneity or random effects assumptions.

[35] In the context of correlated random coefficient models, Graham and Powell (2012) impose a similar structure on their model.

[36] For simplicity, we assume that the selection mechanism only depends on the unobservables that enter the time-invariant component of the structural function $\mu(\cdot)$ in addition to covariates.

Consider the following linear relaxation of the martingale property,

$$E[\ddot{Y}_{it}(0)|X_i, \alpha_i, \varepsilon_i^{t-1}, \mu_i, \eta_i^{t-1}] = \rho_t \ddot{Y}_{i(t-1)}(0), \quad i = 1, \ldots, n, \quad t = 1, 2, \tag{17}$$

and define $\zeta_{it} = \ddot{Y}_{it}(0) - \rho_t \ddot{Y}_{i(t-1)}$.

By the same arguments as in Proposition 4.1, conditional on $X_i$, we can decompose the bias of the conditional ATT as follows

$$\Delta_{\text{post}}(X_i) \equiv \text{DiD}(X_i) - \text{ATT}(X_i) = \Delta_{\text{post}}^{\text{sel}}(X_i) + \Delta_{\text{post}}^{\text{dev}}(X_i),$$

where

$$\begin{aligned}
\Delta_{\text{post}}^{\text{sel}}(X_i) &\equiv E[\zeta_{i2}|G_i = 1, X_i] - E[\zeta_{i2}|G_i = 0, X_i], \\
\Delta_{\text{post}}^{\text{dev}}(X_i) &\equiv (\rho_2 - 1)(E[Y_{i1}|G_i = 1, X_i] - E[Y_{i1}|G_i = 0, X_i]).
\end{aligned}$$

Therefore, the unconditional bias is (with a slight abuse of notation)

$$\begin{aligned}
\Delta_{\text{post}} &\equiv E[\Delta_{\text{post}}(X_i)|G_i = 1] \\
&= E[E[\zeta_{i2}|G_i = 1, X_i] - E[\zeta_{i2}|G_i = 0, X_i]|G_i = 1] \\
&\quad +(\rho_2 - 1)(E[Y_{i1}|G_i = 1] - E[E[Y_{i1}|G_i = 0, X_i]|G_i = 1]) \\
&\equiv \Delta_{\text{post}}^{\text{sel}} + \Delta_{\text{post}}^{\text{dev}}
\end{aligned}$$

Analogous to Section 4.3, we rely on pre-treament and control group counterparts of $\rho_2$ and the pre-treatment counterpart of $\Delta_{\text{post}}^{\text{sel}}$, $\Delta_{\text{pre}}^{\text{sel}} = E[\zeta_{i1}|G_i = 1] - E[E[\zeta_{i1}|G_i = 0, X_i]|G_i = 1]$, to benchmark the bias.

### F.2 Implementation

For a random variable $W_{it}$, let $E_n[W_{it}|G_i = 1] = \sum_{i=1}^n G_i W_{it} / \sum_{i=1}^n G_i$ and $\widehat{E}[W_{it}|G_i = g, X_i]$ ($\widehat{E}[W_{it}|X_i]$) be a linear regression estimator of $E[W_{it}|G_i = g, X_i]$ ($E[W_{it}|X_i]$).

Given $\Delta_{\text{post}}^{\text{sel}}$ and $\rho_2$, an estimator of $\Delta_{\text{post}}$ is given by

$$\widehat{\Delta}_{\text{post}} = \widehat{\Delta}_{\text{post}}\left(\Delta_{\text{post}}^{\text{sel}}, \rho_2\right) = \Delta_{\text{post}}^{\text{sel}} + (\rho_2 - 1)\left(E_n[Y_{i1}|G_i = 1] - E_n[\widehat{E}[Y_{i1}|G_i = 0, X_i]|G_i = 1]\right).$$

If $\Delta_{\text{post}}^{\text{sel}} = 0$, then

$$\widehat{\Delta}_{\text{post}} = (\rho_2 - 1)\left(E_n[Y_{i1}|G_i = 1] - E_n[\widehat{E}[Y_{i1}|G_i = 0, X_i]|G_i = 1]\right).$$

Alternatively, if $\Delta^{\text{sel}}_{\text{post}} = \Delta^{\text{sel}}_{\text{pre}}$, then

$$\widehat{\Delta}_{\text{post}} = \widehat{\Delta}^{\text{sel}}_{\text{pre}} + (\rho_2 - 1)\left(E_n[Y_{i1}|G_i = 1] - E_n[\widehat{E}[Y_{i1}|G_i = 0, X_i]|G_i = 1]\right),$$

where $\widehat{\Delta}^{\text{sel}}_{\text{pre}} = E_n[\hat{\zeta}_{i1}|G_i = 1] - E_n\left[\widehat{E}\left[\widehat{\zeta}_{i1}|G_i = 0, X_i\right]|G_i = 1\right]$, $\widehat{\zeta}_{i1} = \ddot{\widehat{Y}}_{i1} - \widehat{\rho}_1 \ddot{\widehat{Y}}_{i0}$, $\ddot{\widehat{Y}}_{it} = Y_{it} - \widehat{E}[Y_{it}|X_i]$, and $\widehat{\rho}_1$ is obtained from a regression of $\ddot{\widehat{Y}}_{i1}$ on $\ddot{\widehat{Y}}_{i0}$.

# G Proofs of the results in the main text

## G.1 Auxiliary lemmas

**Lemma G.1.** *Let $W_i$ denote a vector of random variables. Suppose that $P(G_i = 1|W_i) \in (0, 1)$ a.s. Then $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, W_i] = E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0, W_i]$ if and only if $E[G_i(Y_{i2}(0) - Y_{i1}(0))|W_i] = E[G_i|W_i]E[Y_{i2}(0) - Y_{i1}(0)|W_i]$ a.s.*

*Proof.* In the following, all equalities involving conditional expectations are understood as a.s. equalities.

"$\Longrightarrow$": First, note that by the law of total probability, $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, W_i] = E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0, W_i]$ implies

$$E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, W_i] = E[Y_{i2}(0) - Y_{i1}(0)|W_i].$$

The result follows from noting that $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, W_i] = \frac{E[G_i(Y_{i2}(0) - Y_{i1}(0))|W_i]}{P(G_i=1|W_i)}$ by definition.

"$\Longleftarrow$": Since $P(G_i = 1|W_i) \in (0, 1)$, it follows that $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, W_i] = E[Y_{i2}(0) - Y_{i1}(0)|W_i]$. It then follows that

$$\begin{aligned}&E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, W_i]P(G_i = 1|W_i) + E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0, W_i]P(G_i = 0|W_i)\\ &= E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1, W_i].\end{aligned}$$

The result follows from subtracting the first term on the left-hand side and dividing by $P(G_i = 0|W_i)$. □

**Lemma G.2.** *For a scalar random variable $W_i$, let $\dot{W}_i = W_i - E[W_i]$. If $E[\dot{W}_i 1\{\dot{W}_i \leq 0\}] = 0$ or $E[\dot{W}_i 1\{\dot{W}_i \geq 0\}] = 0$, then $W_i = E[W_i]$ a.s.*

*Proof.* We prove the result for the case where $E[\dot{W}_i 1\{\dot{W}_i \leq 0\}] = 0$, since the proof for the other case follows by identical arguments. First, note that by definition $E[\dot{W}_i] = 0$, which is

equivalent to

$$E[\dot{W}_i^+] = E[\dot{W}_i^-], \tag{18}$$

where $\dot{W}_i^+ = |\dot{W}_i| 1\{\dot{W}_i > 0\}$ and $\dot{W}_i^- = |\dot{W}_i| 1\{\dot{W}_i < 0\}$.

Now suppose that $E[\dot{W}_i 1\{\dot{W}_i \leq 0\}] = 0$ holds, which is equivalent to

$$E[\dot{W}_i^+ 1\{\dot{W}_i \leq 0\}] = E[\dot{W}_i^- 1\{\dot{W}_i \leq 0\}], \tag{19}$$

since, by definition, $\dot{W}_i = \dot{W}_i^+ - \dot{W}_i^-$. Note that the left-hand side equals zero by the definition of $\dot{W}_i^+$. As a result, $E[\dot{W}_i^- 1\{\dot{W}_i \leq 0\}] = E[\dot{W}_i^-] = 0$. Since $\dot{W}_i^- \geq 0$, this implies that $P(\dot{W}_i^- = 0) = 1$. Now note that $P(\dot{W}_i^- = 0) = P(|\dot{W}_i| 1\{\dot{W}_i < 0\} = 0) = P(1\{\dot{W}_i < 0\} = 0) = 1$, which implies $P(\dot{W}_i < 0) = 0$.

Since $E[\dot{W}_i] = 0$, (18) further implies that $E[\dot{W}_i^-] = E[\dot{W}_i^+] = 0$. Since $\dot{W}_i^+ \geq 0$, it follows that $P(\dot{W}_i^+ = 0) = 1$. Now note that $P(\dot{W}_i^+ = 0) = P(|\dot{W}_i| 1\{\dot{W}_i > 0\} = 0) = P(1\{\dot{W}_i > 0\} = 0) = 1$, which implies $P(\dot{W}_i > 0) = 0$.

Together, $P(\dot{W}_i < 0) = 0$ and $P(\dot{W}_i > 0) = 0$ imply that $P(\dot{W}_i = 0) = 1 - (P(\dot{W}_i < 0) + P(\dot{W}_i > 0)) = 1$, which completes the proof. □

### G.2 Proof of Theorem 3.1

"$\Longrightarrow$": We prove the result for the case where $P(E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] > 0) < 1$. The proof for the case where $P(E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] < 0) < 1$ follows from the same arguments.

Under Assumption SEL and because $P(E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] > 0) < 1$, the selection mechanism

$$G_i = 1\{\nu_i > c\} 1\{E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] \leq 0\} \tag{20}$$

is nondegenerate, that is,

$$P(1\{\nu_i > c\} 1\{E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] \leq 0\} = 1) \in (0, 1).$$

If Assumption PT holds for all non-degenerate selection mechanisms, then it holds for the mechanism (20). By Lemma G.1, Assumption PT holding for the mechanism in (20) is equivalent to

$$E[1\{\nu_i > c\} 1\{E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] \leq 0\} (\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))] = 0,$$

which, by Assumption SEL, is equivalent to

$$E[1\{E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] \leq 0\}(\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))] = 0.$$

By the law of iterated expectations (LIE), this is further equivalent to

$$E[1\{E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i] \leq 0\}E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i]] = 0$$

Since $E[E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i]] = 0$, the result follows by Lemma G.2.
"$\Longleftarrow$": By the LIE,

$$\begin{aligned} E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|G_i] &= E[E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i, \nu_i]|G_i] \\ &= E[E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i]|G_i] = 0, \end{aligned}$$

where the penultimate equality follows from Assumption SEL.

### G.3 Proof of Lemma 4.1

We rewrite $\Delta_{\text{post}}$ as follows,

$$\begin{aligned} \Delta_{\text{post}} &\equiv E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0] \\ &= \frac{E[G_i(Y_{i2}(0) - Y_{i1}(0))]}{P(G_i = 1)} - \frac{E[(1 - G_i)(Y_{i2}(0) - Y_{i1}(0)]}{P(G_i = 0)} \\ &= \frac{(1 - P(G_i = 1))E[G_i(Y_{i2}(0) - Y_{i1}(0))] - P(G_i = 1)E[Y_{i2}(0) - Y_{i1}(0)]}{P(G_i = 1)P(G_i = 0)} \\ &\quad + \frac{P(G_i = 1)E[G_i(Y_{i2}(0) - Y_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} \\ &= \frac{E[G_i(\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} \\ &= \frac{E[(G_i - E[G_i|\omega_i])(\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} + \frac{E[E[G_i|\omega_i](\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} \\ &= \frac{E[(G_i - E[G_i|\omega_i])(\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} + \frac{E[E[G_i|\omega_i]E[\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0)|\omega_i]]}{P(G_i = 1)P(G_i = 0)} \end{aligned}$$

The first equality follows by definition of $\Delta_{\text{post}}$. The second equality follows by the LIE. The penultimate equality follows from subtracting and adding $E[G_i|\omega_i](\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))$ inside the expectation in the numerator. The last equality follows from the LIE. □

### G.4 Proof of Proposition 4.1

We first simplify $\Delta^{\text{sel}}_{\text{post}}$ in Lemma 4.1 with $\omega_i = (\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1)$ as follows,

$$\begin{aligned}
\Delta^{\text{sel}}_{\text{post}} &= \frac{E[(G_i - E[G_i|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1])(\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)}, \\
&= \frac{E[G_i(\dot{Y}_{i2}(0) - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} \\
&\quad - \frac{E[E[G_i|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1](E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1] - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} \\
&= \frac{E[G_i(\dot{Y}_{i2}(0) - E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1])]}{P(G_i = 1)P(G_i = 0)} \\
&= E[\zeta_{i2}|G_i = 1] - E[\zeta_{i2}|G_i = 0]
\end{aligned}$$

The second and the third equality follow from the LIE. The last equality follows from Assumption REL and because $G_i \in \{0, 1\}$.

Next, we simplify $\Delta^{\text{dev}}_{\text{post}}$ in Lemma 4.1 with $\omega_i = (\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1)$ as follows,

$$\begin{aligned}
\Delta^{\text{dev}}_{\text{post}} &= \frac{E[E[G_i|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1](E[\dot{Y}_{i2}(0)|\alpha_i, \varepsilon_i^1, \mu_i, \eta_i^1] - \dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} \\
&= (\rho_2 - 1)\frac{E[G_i\dot{Y}_{i1}(0))]}{P(G_i = 1)P(G_i = 0)} \\
&= (\rho_2 - 1)(E[Y_{i1}(0)|G_i = 1] - E[Y_{i1}(0)|G_i = 0].
\end{aligned}$$

The second equality follows from Assumption REL and the LIE. The last equality follows because $G_i \in \{0, 1\}$. □

## H Proofs of results in the Appendix

### H.1 Auxiliary lemmas

**Lemma H.1** (Equivalence with multiple periods)**.** *Suppose that Assumption NA holds and* $P(G_i = g) \in (0, 1)$ *for* $g \in \{2, \ldots, T, \infty\}$*. Then Assumption PT-MP is equivalent to* $E[1\{G_i = g\}(\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty))] = 0$ *for* $g \in \{2, \ldots, T, \infty\}$ *and* $t \in \{2, \ldots, T\}$*.*

*Proof.* Assumption PT-MP is equivalent to

$$E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty)|G_i = g] = E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty)|G_i = \infty] \quad \text{for } (g, t) \in \{2, \ldots, T\}^2,$$

which, since $E[\dot{Y}_{it}(\infty)] = 0$, is also equivalent to

$$E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty)|G_i = g] = 0 \quad \text{for } (g,t) \in \{2,\ldots,T,\infty\} \times \{2,\ldots,T\}. \tag{21}$$

Thus, we need to show that (21) is equivalent to $E[1\{G_i = g\}(\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty))] = 0$ for $g \in \{2,\ldots,T,\infty\}$ and $t \in \{2,\ldots,T\}$. This follows because

$$E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty)|G_i = g] = \frac{E[1\{G_i = g\}(\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty))]}{P(G_i = g)}$$

for $(g,t) \in \{2,\ldots,T,\infty\} \times \{2,\ldots,T\}$, since $P(G_i = g) \in (0,1)$ for $g \in \{2,\ldots,T,\infty\}$ by assumption. □

**Lemma H.2.** *For a scalar random variable $W_i$, let $\ddot{W}_i = W_i - E[W_i|X_i]$. If $E[\ddot{W}_i 1\{\ddot{W}_i \leq 0\}|X_i] = 0$ or $E[\ddot{W}_i 1\{\ddot{W}_i \geq 0\}|X_i] = 0$ a.s., then $W_i = E[W_i|X_i]$ a.s.*

*Proof.* We prove the result for the case where $E[\ddot{W}_i 1\{\ddot{W}_i \leq 0\}|X_i] = 0$, since the proof for the other case follows by identical arguments. In the following, all statements involving conditional expectations hold a.s.

First, note that by definition $E[\ddot{W}_i|X_i] = 0$, which is equivalent to

$$E[\ddot{W}_i^+|X_i] = E[\ddot{W}_i^-|X_i], \tag{22}$$

where $\ddot{W}_i^+ = |\ddot{W}_i|1\{\ddot{W}_i > 0\}$ and $\ddot{W}_i^- = |\ddot{W}_i|1\{\ddot{W}_i < 0\}$.

Now suppose that $E[\ddot{W}_i 1\{\ddot{W}_i \leq 0\}|X_i] = 0$ holds, which is equivalent to

$$E[\ddot{W}_i^+ 1\{\ddot{W}_i \leq 0\}|X_i] = E[\ddot{W}_i^- 1\{\ddot{W}_i \leq 0\}|X_i], \tag{23}$$

since, by definition, $\ddot{W}_i = \ddot{W}_i^+ - \ddot{W}_i^-$. Note that the left-hand side equals zero by the definition of $\ddot{W}_i^+$. As a result, $E[\ddot{W}_i^- 1\{\ddot{W}_i \leq 0\}|X_i] = E[\ddot{W}_i^-|X_i] = 0$. Since $\ddot{W}_i^- \geq 0$, this implies that $P(\ddot{W}_i^- = 0|X_i) = 1$. Now note that $P(\ddot{W}_i^- = 0|X_i) = P(|\ddot{W}_i|1\{\ddot{W}_i < 0\} = 0|X_i) = P(1\{\ddot{W}_i < 0\} = 0|X_i) = 1$, which implies $P(\ddot{W}_i < 0|X_i) = 0$.

Since $E[\ddot{W}_i|X_i] = 0$, (22) further implies that $E[\ddot{W}_i^-|X_i] = E[\ddot{W}_i^+|X_i] = 0$. Since $\ddot{W}_i^+ \geq 0$, it follows that $P(\ddot{W}_i^+ = 0|X_i) = 1$. Now note that $P(\ddot{W}_i^+ = 0|X_i) = P(|\ddot{W}_i|1\{\ddot{W}_i > 0\} = 0|X_i) = P(1\{\ddot{W}_i > 0\} = 0|X_i) = 1$, which implies $P(\ddot{W}_i > 0|X_i) = 0$.

Together, $P(\ddot{W}_i < 0|X_i) = 0$ and $P(\ddot{W}_i > 0|X_i) = 0$ imply that $P(\ddot{W}_i = 0|X_i) = 1 - (P(\ddot{W}_i < 0|X_i) + P(\ddot{W}_i > 0|X_i)) = 1$, which further implies that $P(\ddot{W}_i = 0) = 1$ and thereby completes the proof. □

**Lemma H.3.** *Let $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ denote a vector of random variables on $\mathcal{A} \times \mathcal{E}^2$. Suppose that $\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i \overset{d}{=} \varepsilon_{i2}, \varepsilon_{i1}|\alpha_i$ holds. Then,*

*(i)* $F_{\varepsilon_{i1}|\alpha_i}(e|a) = F_{\varepsilon_{i2}|\alpha_i}(e|a)$ *a.e.* $(a, e) \in \mathcal{A} \times \mathcal{E}$.

*(ii) Suppose further that the conditional density of* $(\varepsilon_{i1}, \varepsilon_{i2})|\alpha_i$, $f_{\varepsilon_{i1},\varepsilon_{i2}|\alpha_i}(e_1, e_2|a)$, *exists and the conditional density of* $\varepsilon_{it}|\alpha_i$, $f_{\varepsilon_{it}|\alpha_i}(e|a)$, *is strictly positive for* $(e, a) \in \mathcal{E} \times \mathcal{A}$, $t = 1, 2$. *Then,* $F_{\varepsilon_{i1}|\varepsilon_{i2},\alpha_i}(e_1|e_2, a) = F_{\varepsilon_{i2}|\varepsilon_{i1},\alpha_i}(e_1|e_2, a)$ *a.e.* $(a, e_1, e_2) \in \mathcal{A} \times \mathcal{E}^2$.

*Proof.* **(i)** By the definition of the marginal distribution, the conditional exchangeability restriction implies (i) by the following *a.e.*

$$F_{\varepsilon_{i1}|\alpha_i}(e_1|a) = \lim_{e_2 \to \infty} F_{\varepsilon_{i1},\varepsilon_{i2}|\alpha_i}(e_1, e_2|a) = \lim_{e_2 \to \infty} F_{\varepsilon_{i1},\varepsilon_{i2}|\alpha_i}(e_2, e_1|a) = F_{\varepsilon_{i2}|\alpha_i}(e_1|a). \tag{24}$$

**(ii)** To simplify notation, we prove this result assuming $\mathcal{E} = \mathbb{R}^k$, where $k \equiv \dim(\varepsilon_{it})$. Let $\tilde{e} = (\tilde{e}^1, \ldots \tilde{e}^k)$ be the constant of integration and $e_1 = (e_1^1, \ldots, e_1^k)$. By the definition of the conditional distribution and (i) of this lemma, the conditional exchangeability restriction implies (ii) by the following

$$\begin{aligned} F_{\varepsilon_{i1}|\varepsilon_{i2},\alpha_i}(e_1|e_2, a) &= \int_{-\infty}^{e_1^1} \cdots \int_{-\infty}^{e_1^k} f_{\varepsilon_{i1}|\varepsilon_{i2},\alpha_i}(\tilde{e}|e_2, a) d\tilde{e}^1 \ldots d\tilde{e}^k = \int_{-\infty}^{e_1^1} \cdots \int_{-\infty}^{e_1^k} \frac{f_{\varepsilon_{i1},\varepsilon_{i2}|\alpha_i}(\tilde{e}, e_2|a)}{f_{\varepsilon_{i2}|\alpha_i}(e_2|a)} d\tilde{e}^1 \ldots d\tilde{e}^k \\ &= \int_{-\infty}^{e_1^1} \cdots \int_{-\infty}^{e_1^k} \frac{f_{\varepsilon_{i2},\varepsilon_{i1}|\alpha_i}(\tilde{e}, e_2|a)}{f_{\varepsilon_{i1}|\alpha_i}(e_2|a)} d\tilde{e}^1 \ldots d\tilde{e}^k = F_{\varepsilon_{i2}|\varepsilon_{i1},\alpha_i}(e_1|e_2, a). \end{aligned}$$

□

## H.2 Proof of Theorem B.1

"$\Longrightarrow$": We first consider the case where $P(E[\dot{Y}_{it}(\infty) - \dot{Y}_{i(t-1)}(\infty))|\omega_i] > 0) < 1$ for $t \in \{2, \ldots, T\}$. Since Assumption PT-MP holds for all $g \in \mathcal{G}_\omega$, it holds for the following selection mechanism, where $\mathcal{G}_S = \{2, \ldots, T\}$ denotes the set of switcher groups,

$$\check{g}(\omega_i, \nu_i) = \begin{cases} g \text{ if } 1\{E[\dot{Y}_{ig}(\infty) - \dot{Y}_{i(g-1)}(\infty)|\omega_i] \leq 0\} 1\{\nu_i \in B_g\} = 1, g \in \mathcal{G}_S \\ \infty \text{ otherwise.} \end{cases}$$

Next, we show that $P(\check{g}(\omega_i, \nu_i) = g) \in (0, 1)$ for $g \in \{2, \ldots, T, \infty\}$, such that $G_i = \check{g}(\omega_i, \nu_i)$ is a nondegenerate selection mechanism. For $g \in \mathcal{G}_S$,

$$P(\check{g}(\omega_i, \nu_i) = g) = P(E[\dot{Y}_{ig}(\infty) - \dot{Y}_{i(g-1)}(\infty)|\omega_i] \leq 0) P(\nu_i \in B_g) \in (0, 1), \tag{25}$$

where the first equality follows from the independence between $\omega_i$ and $\nu_i$ implied by Assumption SEL-MP. Since $P(E[\dot{Y}_{ig}(\infty) - \dot{Y}_{i(g-1)}(\infty)|\omega_i] \leq 0) > 0$ and $P(\nu_i \in B_g) \in (0, 1)$

for $\mathcal{G}_S$, it follows that $P(\check{g}(\omega_i,\nu_i)=g)\in(0,1)$ for $g\in\mathcal{G}_S$. This further implies that $\sum_{g\in\mathcal{G}_S}P(\check{g}(\omega_i,\nu_i)=g)>0$ and therefore $P(\check{g}(\omega_i,\nu_i)=\infty)=1-\sum_{g\in\mathcal{G}_S}P(\check{g}(\omega_i,\nu_i)=g)<1$. It remains to show that $P(\check{g}(\omega_i,\nu_i)=\infty)>0$.

$$P(\check{g}(\omega_i,\nu_i)=\infty) \;=\; 1-\sum_{g\in\mathcal{G}_S}P(\check{g}(\omega_i,\nu_i)=g)\geq 1-\sum_{g\in\mathcal{G}_S}P(\nu_i\in B_g)=P(\nu_i\in B_\infty)>0,$$

where the weak inequality follows from $P(\check{g}(\omega_i,\nu_i)=g)\leq P(\nu_i\in B_g)$ in (25). The last equality follows from Assumption SEL-MP, in particular that $\{B_g\}_{g=2,\ldots,T,\infty}$ is a non-overlapping partition of the support of $\nu_i$ and $P(\nu_i\in B_g)\in(0,1)$ for $g\in\{2,\ldots,T,\infty\}$. It follows that $P(\check{g}(\omega_i,\nu_i)=g)\in(0,1)$ for $g\in\{2,\ldots,T,\infty\}$.

Now we can invoke Lemma H.1 to show the implication of Assumption PT-MP with $G_i=\check{g}(\omega_i,\nu_i)$; specifically for any $g\in\mathcal{G}_S$

$$\begin{aligned}
0 &= E[1\{\check{g}(\omega_i,\nu_i)=g\}(\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty))] \\
&= E[1\{E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]\leq 0\}1\{\nu_i\in B_g\}(\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty))] \\
&= P(\nu_i\in B_g)E[1\{E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]\leq 0\}(\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty))] \\
&= P(\nu_i\in B_g)E[1\{E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]\leq 0\}E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]].
\end{aligned}$$

The third equality follows by the independence condition in Assumption SEL-MP and the definition of $\omega_i$. The last equality follows by the LIE. Since $P(\nu_i\in B_g)>0$ by Assumption SEL-MP, it follows that $E[1\{E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]\leq 0\}E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]]=0$ for $g\in\mathcal{G}_S$. By Lemma G.2 with $W_i=E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]$ for each $g\in\mathcal{G}_S$, it follows that $E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]=0$ a.s. for each $g\in\mathcal{G}_S=\{2,\ldots,T\}$, which implies the result.

The proof for the case where $P(E[\dot{Y}_{it}(\infty)-\dot{Y}_{i(t-1)}(\infty)|\omega_i]<0)<1$ for $t\in\{2,\ldots,T\}$ follows symmetrically using the selection mechanism,

$$\check{g}(\omega_i,\nu_i)=\begin{cases} g \text{ if } 1\{E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]\geq 0\}1\{\nu_i\in B_g\}=1, g\in\mathcal{G}_S \\ \infty \text{ otherwise.}\end{cases}$$

The proof for the case where $P(E[\dot{Y}_{it}(\infty)-\dot{Y}_{i(t-1)}(\infty)|\omega_i]>0)<1$ for $t\in\mathcal{G}_1\subset\mathcal{G}_S$ and $P(E[\dot{Y}_{is}(\infty)-\dot{Y}_{i(s-1)}(\infty)|\omega_i]<0)<1$ for $s\in\mathcal{G}_2=\mathcal{G}_1^c\cap\mathcal{G}_S$ follows from using the following

selection mechanism

$$\check{g}(\omega_i,\nu_i)=\begin{cases} g \text{ if } 1\{E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]\leq 0\}1\{\nu_i\in B_g\}=1, g\in\mathcal{G}_1,\\ g \text{ if } 1\{E[\dot{Y}_{ig}(\infty)-\dot{Y}_{i(g-1)}(\infty)|\omega_i]\geq 0\}1\{\nu_i\in B_g\}=1, g\in\mathcal{G}_2,\\ \infty \text{ otherwise.}\end{cases}$$

"$\Longleftarrow$": This direction is immediate by the LIE.

### H.3 Proof of Theorem D.1

"$\Longrightarrow$": We prove the result for the case where $P(E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]>0)<1$. The proof for the case where $P(E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]<0)<1$ follows from the same arguments.

Under Assumption SEL-X and because $P(E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]>0)<1$, the selection mechanism

$$G_i=1\{\nu_i>c\}1\{E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]\leq 0\} \tag{26}$$

is nondegenerate, that is,

$$P(1\{\nu_i>c\}1\{E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]\leq 0\}=1|X_i)\in(0,1).$$

If Assumption PT-X holds for all non-degenerate selection mechanisms $g\in\mathcal{G}^x_\omega$, then it holds for the mechanism (26). By Lemma G.1, Assumption PT-X holding for the mechanism in (26) is equivalent to

$$E[1\{\nu_i>c\}1\{E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]\leq 0\}(\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0))|X_i]=0,$$

which, by Assumption SEL-X, is equivalent to

$$E[1\{E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]\leq 0\}(\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0))|X_i]=0.$$

By the law of iterated expectations (LIE), this is further equivalent to

$$E[1\{E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]\leq 0\}E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]|X_i]=0$$

Since $E[E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]|X_i]=0$ by construction, the result follows by Lemma H.2.
"$\Longleftarrow$": By the LIE,

$$\begin{aligned}E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|G_i,X_i] &= E[E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,\nu_i,X_i]|G_i,X_i]\\ &= E[E[\ddot{Y}_{i2}(0)-\ddot{Y}_{i1}(0)|\omega_i,X_i]|G_i,X_i]=0,\end{aligned}$$

where the penultimate equality follows from Assumption SEL-X.

### H.4 Proof of Proposition D.1

In this proof, all equalities involving random variables are understood to hold a.s.

First, by Lemma G.1, Assumption PT-NSP under Assumption NSP-X holds if and only if

$$\begin{aligned}&E[G_i(Y_{i2}(0)-Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\\ =&E[G_i|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0)-Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu].\end{aligned} \tag{27}$$

Next, we state some preliminary observations and then proceed to show each statement separately.

Note that, by the LIE, Assumption SEL-CI and the definition of $\bar{g}(\cdot)$, the LHS of (27) equals the following,

$$\begin{aligned}&E[G_i(Y_{i2}(0)-Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\\ =&E[E[G_i|X_i^\mu, X_i^\lambda, \alpha_i^\mu, \alpha_i^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda](Y_{i2}(0)-Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\\ =&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(Y_{i2}(0)-Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu].\end{aligned} \tag{28}$$

Similarly, by the LIE, the RHS of (27) equals the following,

$$\begin{aligned}&E[G_i|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0)-Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\\ =&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0)-Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\end{aligned} \tag{29}$$

As a result, in the following, to show that Assumptions SC1-NSP, SC2-NSP, and SC3-NSP are sufficient for Assumption PT-NSP, it suffices to show that each assumption implies the following equality,

$$\begin{aligned}&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(Y_{i2}(0)-Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\\ =&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0)-Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\end{aligned}$$

**(i)** By Assumption NSP-X, it follows that

$$\begin{aligned}&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(Y_{i2}(0)-Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\\ =&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu)-\mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]\\ &+E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda)-\lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu],\end{aligned} \tag{30}$$

We first examine the first term on the RHS of the above equality. Note that by the symmetry restrictions in Assumptions SC1-NSP.i and SC1-NSP.ii, it follows that a.e. $(a, x^\mu, x_1^\lambda, x_2^\lambda) \in \mathcal{A} \times \mathcal{X}_\mu \times \mathcal{X}_\lambda^2$

$$
\begin{aligned}
&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)\mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda = (x_1^\lambda, x_2^\lambda), X_{i1}^\mu = X_{i2}^\mu = x^\mu, \alpha_i^\mu = a] \\
=&\int \bar{g}(x^\mu, x^\mu, x_1^\lambda, x_2^\lambda, a, e_1, e_2)\mu(x^\mu, a, e_1) dF_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu | X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu, \alpha_i^\mu}(e_1, e_2|(x_1^\lambda, x_2^\lambda), x^\mu, a) \\
=&\int \bar{g}(x^\mu, x^\mu, x_1^\lambda, x_2^\lambda, a, e_2, e_1)\mu(x^\mu, a, e_1) dF_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu | X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu, \alpha_i^\mu}(e_2, e_1|(x_1^\lambda, x_2^\lambda), x^\mu, a) \\
=&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu)|X_i^\lambda = (x_1^\lambda, x_2^\lambda), X_{i1}^\mu = X_{i2}^\mu = x^\mu, \alpha_i^\mu = a]. \qquad (31)
\end{aligned}
$$

As a result, the first summand in (30) equals zero by (31) and the LIE.

Next, we consider the second summand in (30),

$$
\begin{aligned}
&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\bar{g}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]. \qquad (32)
\end{aligned}
$$

The first equality follows from the conditional independence assumption in Assumption SC1-NSP.iii. The last equality follows from the time homogeneity of $F_{\varepsilon_{it}^\mu | X_i^\mu, X_i^\lambda, \alpha_i^\mu}$, which follows from the exchangeability restriction in Assumption SC1-NSP.ii by Lemma H.3.i, and implies that $E[\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu, \alpha_i^\mu] = 0$ and

$$
E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] = E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]
$$

by the LIE. As a result, the above implies that Assumption PT-NSP holds.

**(ii)** By Assumption SC2-NSP.i, we can define $\check{\bar{g}}(x_1^\mu, x_2^\mu, x_1^\lambda, x_2^\lambda, a^\mu, e_1^\mu) = \bar{g}(x_1^\mu, x_2^\mu, x_1^\lambda, x_2^\lambda, a^\mu, e_1^\mu, e_2^\mu)$. By Assumption NSP-X, it follows that

$$
\begin{aligned}
&E[\check{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu)(Y_{i2}(0) - Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\check{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu)\Delta_{\mu,i}|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&+ E[\check{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu)(\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\check{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\Delta_{\mu,i}|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&+ E[\check{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\check{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \qquad (33)
\end{aligned}
$$

The second equality follows from the conditional independence conditions in Assumptions SC2-NSP.ii and SC2-NSP.iii. The last equality follows from Assumption NSP-X. Equation (33) then implies Assumption PT-NSP.

**(iii)** By Assumption SC3-NSP.i, we can define $\breve{\bar{g}}(x_1^\mu, x_2^\mu, x_1^\lambda, x_2^\lambda, a^\mu) = \bar{g}(x_1^\mu, x_2^\mu, x_1^\lambda, x_2^\lambda, a^\mu, e_1^\lambda, e_2^\lambda)$. Now by Assumption NSP-X and SC3-NSP.i, it follows that

$$
\begin{aligned}
&E[\breve{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu)(Y_{i2}(0) - Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\breve{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu)(\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&+ E[\breve{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu)(\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\breve{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu)E[\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu, \alpha_i^\mu]|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&+ E[\breve{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\breve{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\breve{\bar{g}}(X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \alpha_i^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu],
\end{aligned}
$$

where the first equality follows from Assumption NSP-X. The second equality follows by applying the LIE to the first term and the conditional independence imposed in Assumption SC3-NSP.iii to the second term. The first term on the RHS of the second equality equals zero by the conditioning on $X_{i1}^\mu = X_{i2}^\mu$ and the time homogeneity condition in Assumption SC3-NSP.ii. The last equality follows from noting, similar to the proof of (i), that since $E[\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu, \alpha_i^\mu] = 0$,

$$
E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] = E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]
$$

by the LIE. This completes the proof. □

### H.5 Proof of Proposition E.1

Under Assumption NSP-X,

$$
\begin{aligned}
&E[Y_{i2}(0) - Y_{i1}(0)|G_i, X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
=&E[\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)|G_i, X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \qquad (34) \\
&+ E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|G_i, X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]. \qquad (35)
\end{aligned}
$$

The remainder of the proof follows in two steps. First, we show that the term in (34) equals zero under our assumptions. Second, we show that the second term is conditionally mean independent of $G_i$, which implies Assumption PT-NSP.

We proceed to show that under Assumption TH the term in (34) equals zero by the

following,

$$
\begin{aligned}
&E[\mu(X_{i1}^{\mu}, \alpha_i^{\mu}, \varepsilon_{i1}^{\mu})|G_i = g, X_i^{\lambda} = (x_1^{\lambda}, x_2^{\lambda}), X_{i1}^{\mu} = X_{i2}^{\mu} = x^{\mu}] \\
=&\int \mu(x^{\mu}, a^{\mu}, e^{\mu}) dF_{\alpha_i^{\mu}, \varepsilon_{i1}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}}(a^{\mu}, e^{\mu}|g, (x^{\mu}, x^{\mu}), (x_1^{\lambda}, x_2^{\lambda})) \\
=&\int \mu(x^{\mu}, a^{\mu}, e^{\mu}) dF_{\alpha_i^{\mu}, \varepsilon_{i2}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}}(a^{\mu}, e^{\mu}|g, (x^{\mu}, x^{\mu}), (x_1^{\lambda}, x_2^{\lambda})) \\
=&E[\mu(X_{i2}^{\mu}, \alpha_i^{\mu}, \varepsilon_{i2}^{\mu})|G_i = g, X_i^{\lambda} = (x_1^{\lambda}, x_2^{\lambda}), X_{i1}^{\mu} = X_{i2}^{\mu} = x^{\mu}], \qquad (36)
\end{aligned}
$$

where the first and last equalities follow by definition, whereas the penultimate equality follows from Assumption TH noting that it implies $\alpha_i^{\mu}, \varepsilon_{i1}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda} \overset{d}{=} \alpha_i^{\mu}, \varepsilon_{i2}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}$.

Finally, we show that Assumption CRE implies the following for (35)

$$
\begin{aligned}
&E[\lambda_2(X_{i2}^{\lambda}, \alpha_i^{\lambda}, \varepsilon_{i2}^{\lambda}) - \lambda_1(X_{i1}^{\lambda}, \alpha_i^{\lambda}, \varepsilon_{i1}^{\lambda})|G_i = g, X_i^{\lambda} = (x_1^{\lambda}, x_2^{\lambda}), X_{i1}^{\mu} = X_{i2}^{\mu} = x^{\mu}] \\
=&\int (\lambda_2(x_2^{\lambda}, a^{\lambda}, e_2^{\lambda}) - \lambda_1(x_1^{\lambda}, a^{\lambda}, e_1^{\lambda})) dF_{\alpha_i^{\lambda}, \varepsilon_{i1}^{\lambda}, \varepsilon_{i2}^{\lambda}|G_i, X_i^{\mu}, X_i^{\lambda}}(a^{\lambda}, e_1^{\lambda}, e_2^{\lambda}|g, (x^{\mu}, x^{\mu}), (x_1^{\lambda}, x_2^{\lambda})) \\
=&\int (\lambda_2(x_2^{\lambda}, a^{\lambda}, e_2^{\lambda}) - \lambda_1(x_1^{\lambda}, a^{\lambda}, e_1^{\lambda})) dF_{\alpha_i^{\lambda}, \varepsilon_{i1}^{\lambda}, \varepsilon_{i2}^{\lambda}|X_i^{\mu}, X_i^{\lambda}}(a^{\lambda}, e_1^{\lambda}, e_2^{\lambda}|(x^{\mu}, x^{\mu}), (x_1^{\lambda}, x_2^{\lambda})) \\
=&E[\lambda_2(X_{i2}^{\lambda}, \alpha_i^{\lambda}, \varepsilon_{i2}^{\lambda}) - \lambda_1(X_{i1}^{\lambda}, \alpha_i^{\lambda}, \varepsilon_{i1}^{\lambda})|X_i^{\lambda} = (x_1^{\lambda}, x_2^{\lambda}), X_{i1}^{\mu} = X_{i2}^{\mu} = x^{\mu}], \qquad (37)
\end{aligned}
$$

where the penultimate equality follows by Assumption CRE. This completes the proof. □

### H.6 Proof of Proposition E.2

Throughout this proof, equalities involving conditioning statements are understood to hold *a.e.* We proceed to show each result separately.

**(i)** It suffices to show (i.a) Assumptions SC1-NSP.i and SC1-NSP.ii imply Assumption TH and (i.b) Assumptions SC1-NSP.i and SC1-NSP.iii imply Assumption CRE.

**(i.a)** First, we define the conditional joint (mixed) density function of $\varepsilon_{i1}^{\mu}$ and $G_i$

$$
\begin{aligned}
&f_{\varepsilon_{i1}^{\mu}, G_i|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1, g|x^{\mu}, x^{\lambda}, a) \\
=\;& P(G_i = g|\varepsilon_{i1}^{\mu} = e_1, X_i^{\mu} = x^{\mu}, X_i^{\lambda} = x^{\lambda}, \alpha_i^{\mu} = a) f_{\varepsilon_{i1}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1|x^{\mu}, x^{\lambda}, a)
\end{aligned}
$$

Assumption SC1-NSP.ii implies $f_{\varepsilon_{i1}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e|x^{\mu}, x^{\lambda}, a) = f_{\varepsilon_{i2}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e|x^{\mu}, x^{\lambda}, a)$ as well as $f_{\varepsilon_{i1}^{\mu}|\varepsilon_{i2}^{\mu}, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1|e_2, x^{\mu}, x^{\lambda}, a) = f_{\varepsilon_{i2}^{\mu}|\varepsilon_{i1}^{\mu}, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1|e_2, x^{\mu}, x^{\lambda}, a)$ by Lemma H.3. As a result, the second term on the RHS of the last equality equals $f_{\varepsilon_{i2}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1|x^{\mu}, x^{\lambda}, a)$. To complete the proof that Assumption TH holds, it remains to show the following

$$
\begin{aligned}
& P(G_i = g|\varepsilon_{i1}^{\mu} = e_1, X_i^{\mu} = x^{\mu}, X_i^{\lambda} = x^{\lambda}, \alpha_i^{\mu} = a) \\
= & \int 1\{g(x_1^{\mu}, x_2^{\mu}, x_1^{\lambda}, x_2^{\lambda}, a, e_1, e_2) = g\} f_{\varepsilon_{i2}^{\mu}|\varepsilon_{i1}^{\mu}, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_2|e_1, x^{\mu}, x^{\lambda}, a) de_2 \\
= & \int 1\{g(x_1^{\mu}, x_2^{\mu}, x_1^{\lambda}, x_2^{\lambda}, a, e_2, e_1) = g\} f_{\varepsilon_{i1}^{\mu}|\varepsilon_{i2}^{\mu}, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_2|e_1, x^{\mu}, x^{\lambda}, a) de_2 \\
= & P(G_i = g|\varepsilon_{i2}^{\mu} = e_1, X_i^{\mu} = x^{\mu}, X_i^{\lambda} = x^{\lambda}, \alpha_i^{\mu} = a)
\end{aligned}
$$

where the penultimate equality follows from Assumptions SC1-NSP.i and the implication of SC1-NSP.ii that follows from Lemma H.3.

As a result,

$$
f_{\varepsilon_{i1}^{\mu}, G_i|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1, g|x^{\mu}, x^{\lambda}, a) \quad = \quad f_{\varepsilon_{i2}^{\mu}, G_i|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e_1, g|x^{\mu}, x^{\lambda}, a).
$$

Dividing the last equality by $P(G_i = g|x^{\mu}, x^{\lambda}, a) \in (0, 1)$ for $g \in \{0, 1\}$, it follows that $f_{\varepsilon_{i1}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}} = f_{\varepsilon_{i2}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}$, and therefore Assumption TH holds.

**(i.b)** This statement follows in a straightforward manner from the definition of $G_i$ in Assumption SC1-NSP.i and the conditional independence condition in Assumption SC1-NSP.iii which together imply Assumption CRE. This completes the proof of (i).

**(ii)** To show the result, it suffices to show that (ii.a) Assumptions SC3-NSP.i and SC3-NSP.ii imply Assumption TH and (ii.b) Assumptions SC3-NSP.i and SC3-NSP.iii imply Assumption CRE.

**(ii.a)** Under Assumptions SC3-NSP.i and SC3-NSP.ii, conditional on $X_i^{\mu}$, $X_i^{\lambda}$ and $\alpha_i^{\mu}$, $G_i = g(X_{i1}^{\mu}, X_{i2}^{\mu}, X_{i1}^{\lambda}, X_{i2}^{\lambda}, \alpha_i^{\mu})$ is a degenerate random variable equaling either zero or one with probability one. As a result,

$$
\begin{aligned}
& F_{\varepsilon_{it}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e|g, x^{\mu}, x^{\lambda}, a) \\
= & \sum_{h=0,1} P(\varepsilon_{it}^{\mu} \leq e|G_i = g(x_1^{\mu}, x_2^{\mu}, x_1^{\lambda}, x_2^{\lambda}, a), X_i^{\mu} = x^{\mu}, X_i^{\lambda} = x^{\lambda}, \alpha_i^{\mu} = a) 1\{g(x_1^{\mu}, x_2^{\mu}, x_1^{\lambda}, x_2^{\lambda}, a) = h\} \\
= & \sum_{h=0,1} P(\varepsilon_{it}^{\mu} \leq e|X_i^{\mu} = x^{\mu}, X_i^{\lambda} = x^{\lambda}, \alpha_i^{\mu} = a) 1\{g(x_1^{\mu}, x_2^{\mu}, x_1^{\lambda}, x_2^{\lambda}, a) = h\} \\
= & \sum_{h=0,1} F_{\varepsilon_{it}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}(e|x^{\mu}, x^{\lambda}, a) 1\{g(x_1^{\mu}, x_2^{\mu}, x_1^{\lambda}, x_2^{\lambda}, a) = h\}. \qquad (38)
\end{aligned}
$$

As a result, Assumption SC3-NSP.i together with the time homogeneity of $F_{\varepsilon_{it}^{\mu}|X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}$ in Assumption SC3-NSP.ii is sufficient for the time homogeneity of $F_{\varepsilon_{it}^{\mu}|G_i, X_i^{\mu}, X_i^{\lambda}, \alpha_i^{\mu}}$, which yields Assumption TH.

**(ii.b)** The statement (ii.b) is immediate from noting that Assumption SC3-NSP.iii to-

gether with $G_i = g(X_{i1}^{\mu}, X_{i2}^{\mu}, X_{i1}^{\lambda}, X_{i2}^{\lambda}, \alpha_i^{\mu})$ imply that $g(X_{i1}^{\mu}, X_{i2}^{\mu}, X_{i1}^{\lambda}, X_{i2}^{\lambda}, \alpha_i^{\mu}) \perp\!\!\!\perp (\alpha_i^{\lambda}, \varepsilon_{i1}^{\lambda}, \varepsilon_{i2}^{\lambda}) | X_i^{\mu}, X_i^{\lambda}$, which is equivalent to Assumption CRE. This completes the proof of (ii). □