

Selection and parallel trends*

Dalia Ghanem[†] Pedro H. C. Sant’Anna[‡] Kaspar Wüthrich[§]

First draft on arXiv: March 17, 2022. This draft: May 5, 2022

Comments welcome!

Abstract

One of the perceived advantages of difference-in-differences (DiD) methods is that they do not explicitly restrict how units select into treatment. However, when justifying DiD, researchers often argue that the treatment is “quasi-randomly” assigned. We investigate what selection mechanisms are compatible with the parallel trends assumptions underlying DiD. We derive necessary conditions for parallel trends that clarify whether and how selection can depend on time-invariant and time-varying unobservables. Motivated by these necessary conditions, we suggest a menu of interpretable sufficient conditions for parallel trends, thereby providing the formal underpinnings for justifying DiD based on contextual information about selection into treatment. We provide results for both separable and nonseparable outcome models and show that this distinction has implications for the use of covariates in DiD analyses.

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 Brantly Callaway, Clément de Chaisemartin, Gordon Dahl, Stefan Hoderlein, Nikolay Kudrin, David McKenzie, Vitor Possebom, Niklas Potrafke, Jonathan Roth, Yuya Sasaki, and seminar participants at UCSD for comments. The usual disclaimer applies.

[†]Department of Agricultural & Resource Economics, University of California, Davis. dghanem@ucdavis.edu

[‡]Microsoft and Department of Economics, Vanderbilt University. pedro.h.santanna@vanderbilt.edu

[§]Department of Economics, University of California San Diego, 9500 Gilman Dr. La Jolla, CA 92093; CESifo; Ifo Institute. kwuthrich@ucsd.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

Difference-in-differences (DiD) designs are widely used in practice to estimate causal effects. One of the perceived advantages of DiD is that it does not require explicit assumptions on how units select into treatment but instead relies on parallel trends assumptions. However, when justifying DiD in empirical applications, researchers often argue that the treatment is “quasi-randomly” assigned. Although these discussions allude to potential selection mechanisms, they are often not explicit about what constitutes a “quasi-random” assignment, arguably due to the lack of formal guidance. In this paper, we investigate the connection between selection and parallel trends assumptions and thereby establish formal underpinnings for justifying parallel trends in practice.

To analyze the connection between selection and parallel trends, this paper makes three contributions. First, we provide necessary conditions for parallel trends that characterize trade-offs between restrictions on the selection mechanism and the distribution of unobservables. Our necessary conditions show that parallel trends cannot hold unless one is willing to impose restrictions on the selection mechanism or on the distribution of time-varying unobservables. Second, motivated by these necessary conditions, we derive a menu of interpretable sufficient conditions for parallel trends. Our sufficient conditions provide several avenues for justifying parallel trends in applications based on assumptions on the selection mechanism. Third, we examine the role of covariates in making parallel trends assumptions more plausible. This is important due to the wide use of covariates in empirical practice. Our results provide formal guidance on how to justify the inclusion of covariates in DiD analyses for identification purposes.

We consider the classical DiD setting, where we observe N units over two time periods. In the first time period, none of the units is treated; in the second period, some units select into treatment (treatment group), while others remain untreated (control group). Let $Y_{it}(0)$ denote the untreated potential outcome for unit $i = 1, \dots, N$ in time period $t = 1, 2$. The main identifying assumption of DiD is the parallel trends assumption, which requires that the expected change across time in the untreated potential outcome, $Y_{it}(0)$, is identical in

the treatment and control group, formally

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

where $G_i = 1$ indicates the treatment group and $G_i = 0$ indicates the control group.

Our goal within this classical framework is to understand what conditions on selection and unobservables are compatible with parallel trends. In addition to being of independent interest, our results are relevant for more general DiD settings with multiple periods and groups because these more general settings can often be viewed as a sequence of two-period, two-group problems (e.g., de Chaisemartin and D’Haultfoeuille, 2020; Callaway and Sant’Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021).¹

We begin our analysis with a separable model for the untreated potential outcome, $Y_{it}(0)$, and a selection mechanism that depends on all unobservable determinants of $Y_{it}(0)$,

$$Y_{it}(0) = \alpha_i + \lambda_t + \varepsilon_{it}, \tag{1}$$

$$G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}), \tag{2}$$

where α_i and ε_{it} are time-invariant and time-varying unobservables, respectively, and $g(\cdot)$ is an arbitrary function. Equation (1) imposes separability in the unobservable determinants of the *untreated* potential outcome and allows for a transparent discussion of our main theoretical results.² Equation (2) is a general selection mechanism in which selection into treatment may depend on the unobservables that determine the untreated potential outcomes in *both* periods. Our results extend to general selection mechanisms,

$$G_i = \check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}). \tag{3}$$

where (η_{i1}, η_{i2}) are additional unobservables that do not determine the untreated potential outcomes. This selection mechanism accommodates selection based on treatment effects (Roy-style selection) and other economic models of selection.

Our first contribution is to provide necessary conditions for parallel trends. This analysis allows us to characterize the trade-offs between restrictions on selection into treatment and time-varying unobservables. If researchers are not willing to impose any restrictions on how the unobservables affect selection, for parallel trends to hold, it is necessary that all unob-

¹Other work examining identification and estimation in the multiple-period, multiple-group problem includes, e.g., Borusyak, Jaravel, and Spiess (2021); Gardner (2021); Marcus and Sant’Anna (2021); Wooldridge (2021).

²Note that, even under the separable outcome model (1), parallel trends may not hold without further restrictions on the selection mechanism.

servables are time-invariant. Conversely, if they are not willing to impose any restrictions on the distribution of (time-varying) unobservables, then selection needs to be independent of the time-varying unobservables.

Next, motivated by our necessary conditions, we propose a menu of interpretable sufficient conditions on selection and the time-varying unobservables. We specifically consider three selection mechanisms that differ in terms of how they depend on the time-varying unobservables $(\varepsilon_{i1}, \varepsilon_{i2})$: (i) selection into treatment does not depend on $(\varepsilon_{i1}, \varepsilon_{i2})$ but only on the time-invariant unobservable α_i ; (ii) selection depends on α_i and symmetrically on ε_{i1} and ε_{i2} ; (iii) selection only depends on $(\alpha_i, \varepsilon_{i1})$ but not on ε_{i2} . Each selection mechanism is accompanied by distributional restrictions on the unobservables.

We then examine the role of (time-varying) covariates in DiD analyses.³ We start by incorporating them into the separable model,

$$Y_{it}(0) = \alpha_i + \lambda_t + \gamma_t(X_{it}) + \varepsilon_{it}, \quad (4)$$

where $\gamma_t(\cdot)$ is an arbitrary, potentially time-varying nonparametric function of the covariates. We provide interpretable conditions that imply conditional parallel trends. These conditions generalize the sufficient conditions for unconditional parallel trends by allowing covariates to enter the selection mechanism and by conditioning on them in all distributional restrictions. Our analysis highlights the importance of time-varying covariates in weakening the sufficient conditions for parallel trends. In all our sufficient conditions, time-varying covariates can enter the selection mechanism in an unrestricted way. In particular, they do not have to obey the symmetry condition imposed on the time-varying unobservables. Furthermore, conditioning on both time-invariant and time-varying covariates makes the restrictions on the distribution of unobservables more plausible.

Finally, we generalize our analysis to a nonseparable model, where covariates and unobservables can interact,

$$Y_{it}(0) = \mu(X_{it}^\mu, \alpha_i^\mu, \varepsilon_{it}^\mu) + \lambda_t(X_{it}^\lambda, \alpha_i^\lambda, \varepsilon_{it}^\lambda). \quad (5)$$

We show that the basic insights from our analysis of the separable model remain valid. However, nonseparability between the covariates and the unobservables determining selection implies that parallel trends can only hold within subpopulations for which these covariates do not vary across time. This analysis highlights that the role of covariates in DiD analyses depends on how they enter the outcome model.

³Here we assume that covariates are not affected by the treatment. See Caetano, Callaway, Payne, and Rodrigues (2022) for some recent results relaxing this assumption.

Our analysis has important implications for empirical practice. First, our necessary conditions demonstrate that (implicit) assumptions on selection are unavoidable in DiD analyses. Second, we expand the set of formally grounded arguments for justifying DiD based on contextual knowledge about selection, thereby providing formal guidance on what constitutes “quasi-random” assignment. We specifically provide a menu of interpretable conditions that restrict (i) how selection depends on unobservables and (ii) how the distribution of unobservables varies across time. Finally, our results provide guidance on which covariates should be included in DiD analyses to render the required identification conditions more plausible. We discuss these implications in greater detail in Section 6.

This paper contributes to several branches of the causal inference literature. Our first contribution is to the classical literature on canonical DiD setups without covariates. See, e.g., Ashenfelter (1978), Ashenfelter and Card (1985), Heckman and Robb (1985), Card (1990), Card and Krueger (1994), Meyer, Viscusi, and Durbin (1995), and Angrist and Krueger (1999) for early developments, and Section 2 of Lechner (2010) for a historical perspective. Our contribution is to provide foundations for the parallel trends assumption to hold in non-experimental settings, where selection into treatment may depend on time-invariant and time-varying unobservables.

Our second contribution is to the more recent literature on DiD methods. See, e.g., de Chaisemartin and D’Haultfœuille (2021) and Roth, Sant’Anna, Bilinski, and Poe (2022) for recent surveys. Within this strand of the literature, our paper is most closely related to Roth and Sant’Anna (2021), Arkhangelsky and Imbens (2022), and Arkhangelsky, Imbens, Lei, and Luo (2021), though our focus greatly differs from theirs. Roth and Sant’Anna (2021) discuss necessary and sufficient conditions under which a parallel trends assumption is satisfied for all (monotonic) transformations of the untreated potential outcomes. We, on the other hand, focus on highlighting conditions that allow selection mechanisms to depend on unobservables and be compatible with parallel trends. Arkhangelsky and Imbens (2022) and Arkhangelsky, Imbens, Lei, and Luo (2021) propose doubly robust estimation methods that leverage restrictions on outcome models and/or selection models with unconfoundedness-type restrictions; see also Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021). Our results complement theirs as we maintain the parallel trends assumption (and the DiD estimand) and discuss the types of restrictions on selection that are compatible with it.

Our third contribution is to the literature imposing explicit selection and/or outcome models to develop and compare different methods for estimating treatment effects, including DiD (e.g., Ashenfelter and Card, 1985; Heckman and Robb, 1985; Chabé-Ferret, 2015; Verdier, 2020). These selection mechanisms were developed for economic models in applications such as job training and technology adoption. Our results complement this strand

of the literature in several ways. First, our necessary conditions clarify the trade-offs between assumptions on selection and time-varying unobservables that arise in the models considered in this literature. Second, our general sufficient conditions nest several of the existing application-specific restrictions. Finally, we provide results for both separable and nonseparable models and consider the role of covariates in the context of parallel trends assumptions.

Finally, we connect the DiD assumptions to the literature on nonparametric identification in panel models.⁴ A strand in this literature has analyzed the identification of average effects either by allowing for fixed effects and imposing time homogeneity (e.g. Hoderlein and White, 2012; Chernozhukov, Fernández-Val, Hahn, and Newey, 2013) or restricting individual heterogeneity via nonparametric correlated random effects assumptions (e.g. Altonji and Matzkin, 2005; Bester and Hansen, 2009).⁵ Here we establish an explicit connection between DiD and the literature on nonseparable panel models. We show that our sufficient conditions for parallel trends imply combinations of time homogeneity and correlated random effects restrictions. Our results demonstrate how restrictions on the selection mechanism can be used to justify existing identification assumptions in the nonseparable panel data literature.

2 Setup and parallel trends assumptions

Consider the classical DiD setup where we observe two groups and two periods. While several recent papers have considered more general setups with multiple periods and groups, they typically show that these more general setups can be thought of as a sequence of DiD problems with two groups and two periods.⁶ Thus, for expositional simplicity and clarity, we focus on the classical two-period, two-group setup.

Let D_{it} and Y_{it} denote the treatment status and outcome for individual (or unit) i in period t . For a random variable A_{it} , we denote its time series as $A_i \equiv (A_{i1}, A_{i2})$. The

⁴See, e.g., Altonji and Matzkin (2005); Bester and Hansen (2009); Hoderlein and White (2012); Chernozhukov, Fernández-Val, Hahn, and Newey (2013); 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). Recent surveys (Arellano and Honoré, 2001; Arellano and Bonhomme, 2011) 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).

⁵It is worth noting that by imposing correlated random effects assumptions, Altonji and Matzkin (2005); Bester and Hansen (2009) can accommodate arbitrary time heterogeneity.

⁶See, e.g., de Chaisemartin and D’Haultfœuille (2020); Callaway and Sant’Anna (2021); Goodman-Bacon (2021); Sun and Abraham (2021).

treatment group ($G_i = 1$) selects the following 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

Identification in DiD settings relies on parallel trends assumptions.⁸ In Section 3, we abstract from covariates and consider the following unconditional parallel trends assumption.

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 unconditional average treatment effect on the treated (ATT) is identified as

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

In many applications, parallel trends may only be plausible conditional on covariates (e.g., Heckman, Ichimura, and Todd, 1997; Abadie, 2005; Sant’Anna and Zhao, 2020). 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 they are not affected by the treatment.

In Section 4, we examine the model in (4), which is separable in observables and unobservables. We consider the following parallel trends assumption, which is conditional on the time series of the covariates.

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

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

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

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

The unconditional ATT can then be obtained by integrating out with respect to the distri-

⁷We assume that the units do not anticipate their treatment. As a result, at period $t = 1$ we observe untreated outcomes for all units, $Y_{i1}(0)$, while at $t = 2$ we observe $Y_{i2}(1)$ for treated and $Y_{i2}(0)$ for untreated units.

⁸For identification strategies that rely on alternative assumptions, see, e.g., Athey and Imbens (2006); Bonhomme and Sauder (2011); Callaway, Li, and Oka (2018); de Chaisemartin and D’Haultfoeuille (2017); Callaway and Li (2019); D’Haultfoeuille, Hoderlein, and Sasaki (2021).

bution of X_i conditional on $G_i = 1$,

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

In Section 5, we examine the model in (5), which is nonseparable in observables and unobservables. In this context, it is crucial to differentiate between the covariates that interact with the unobservables determining selection, X_{it}^μ , and those that do not, X_{it}^λ . Intuitively, this is because in general we cannot have parallel trends between treatment and control subpopulations that experience changes in X_{it}^μ over time. We therefore examine the following modified version of Assumption PT-X.

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

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

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 the following conditional ATT,

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

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}^μ ,

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

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

3 Selection and parallel trends in separable models

In this section, we examine the trade-off between restrictions on the selection mechanism and the distribution of unobservables in the context of the parallel trends assumption. In order to keep the presentation transparent, we start with a separable outcome model without covariates here and extend the analysis to covariates in Section 4 and to nonseparable models in Section 5.

3.1 Model

We consider a model for the potential outcome without the treatment that is separable in the time-invariant and time-varying unobservables.

Assumption SP.

$$Y_{it}(0) = \alpha_i + \lambda_t + \varepsilon_{it}, \quad E[\varepsilon_{it}] = 0, \quad i = 1, \dots, N, \quad t = 1, 2.$$

In Assumption SP, α_i is the time-invariant unobservable, λ_t is the (non-stochastic) time fixed effect, and ε_{it} is the time-varying, individual-specific unobservable.⁹ The assumption that the time-varying unobservables have mean zero, $E[\varepsilon_{it}] = 0$, is a normalization. It is without loss of generality since we can always redefine λ_t such that this assumption holds. In what follows, we denote the supports of α_i and ε_{it} as \mathcal{A} and \mathcal{E} , respectively.¹⁰

Remark 1 (Two-way fixed effects model). *Assumption SP does not impose the standard two-way fixed effects model for the realized outcomes,*

$$Y_{it} = \delta D_{it} + \alpha_i + \lambda_t + \varepsilon_{it},$$

which implicitly imposes treatment effect homogeneity. Since Assumption SP does not restrict the potential outcome with the treatment, $Y_{it}(1)$, it is consistent with arbitrary treatment effect heterogeneity. □

To analyze the role of selection, it is useful to express Assumption PT equivalently as an orthogonality condition.

Lemma 3.1 (Equivalence). *Suppose that Assumption SP holds and that $P(G_i = 1) \in (0, 1)$. Then Assumption PT is equivalent to $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$.*

Recall that under Assumption SP, the counterfactual trend for each unit i , $Y_{i2}(0) - Y_{i1}(0)$, consists of two components: a common component, $\lambda_2 - \lambda_1$, and an individual-specific component, $\varepsilon_{i2} - \varepsilon_{i1}$. Lemma 3.1 allows us to state Assumption PT as an orthogonality condition between the selection indicator and the individual-specific component.

The key implication of Lemma 3.1 is that for Assumption PT to hold, we need to impose additional restrictions on the selection mechanism and/or the distribution of the time-varying unobservables. To formalize these additional restrictions, we consider a general

⁹The assumption that λ_t is non-stochastic is w.l.o.g. since we can always reparametrize the model as $Y_{it}(0) = \alpha_i + \tilde{\lambda}_t + \tilde{\varepsilon}_{it}$, where $\tilde{\lambda}_t$ is stochastic and $E[\tilde{\lambda}_t] = \lambda_t$ and $\tilde{\varepsilon}_{it} = \varepsilon_{it} - (\tilde{\lambda}_t - \lambda_t)$.

¹⁰For simplicity, we assume that the supports do not depend on (i, t) .

selection mechanism in which units select into treatment based on both the time-invariant unobservable, α_i , and the time-varying unobservables, $(\varepsilon_{i1}, \varepsilon_{i2})$.

Assumption SEL. $G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$, where $P(G_i = 1) \in (0, 1)$.

Without further restrictions, Assumptions SP and SEL do not imply the PT assumption. In Section 3.2, we examine the trade-off between restrictions on the selection mechanism and the distribution of unobservables by deriving necessary conditions for PT in terms of restrictions on $g(\cdot)$ and the distribution of $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$.

Remark 2 (General selection mechanisms). *In Assumption SEL, selection into treatment depends on the unobservable determinants of the untreated potential outcome, $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$. Our results extend to more general selection mechanisms that also depend on a vector of time-invariant and time-varying unobservables, η_{it} , such that $G_i = \check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2})$.¹¹ This general selection mechanism nests several existing economic models of treatment selection. We provide formal results and an example of Roy-style selection based on treatment effects in Section 3.4. \square*

3.2 Necessary conditions for parallel trends

To better understand the implications of the PT assumption for selection, we derive necessary conditions for this assumption in two scenarios.¹² The first is where researchers are not willing to make any assumptions on how the selection mechanism depends on the time-invariant and time-varying unobservables; in the second, researchers are not willing to restrict the distribution of unobservables. These two scenarios clarify the trade-offs between restrictions on selection into treatment and time-varying unobservables.

We first provide a necessary condition for the scenario where researchers are not willing to make any assumptions on the selection mechanism $g(\cdot)$.

Proposition 3.1 (Necessary condition for parallel trends for any selection mechanism). *Suppose that Assumptions SP and SEL hold. Suppose further that $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ is a vector of non-degenerate random variables satisfying $\alpha_i \perp (\varepsilon_{i1}, \varepsilon_{i2})$, $P(\varepsilon_{i1} \leq \varepsilon_{i2}) > 0$, and $P(\varepsilon_{i1} \geq \varepsilon_{i2}) > 0$. If Assumption PT holds for all $g : \mathcal{A} \times \mathcal{E}^2 \mapsto \{0, 1\}$, then $\varepsilon_{i1} = \varepsilon_{i2}$ w.p.1.*

Proposition 3.1 shows that for Assumption PT to hold for any selection mechanism $g(\cdot)$, it is necessary that the time-varying unobservables are in fact time-invariant, even in the “ideal” case where $\alpha_i \perp (\varepsilon_{i1}, \varepsilon_{i2})$. Put simply, if one were to allow for an *unrestricted* selection

¹¹Random assignment of the treatment is nested as a special case of the general selection mechanism, where $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2})$ is a trivial function of $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ and $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) \perp (\eta_{i1}, \eta_{i2})$.

¹²Similar conditions hold for the nonseparable model; see Appendix A.

mechanism, one would need to rule out time-varying shocks. This assumption is too strong in many applications, motivating restrictions on the selection mechanism, which we examine in Section 3.3.

Next, consider a scenario where researchers are not willing to impose any restrictions on the distribution of unobservables, $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$. The following proposition shows that for Assumption PT to hold for any $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$, selection needs to be independent of the time-varying unobservables $(\varepsilon_{i1}, \varepsilon_{i2})$. To state the proposition, we recall the definition of a complete class of distributions (Equations (4.8)–(4.9) on p.115 in Lehmann and Romano, 2005).

Definition 3.1 (Completeness of a class of distributions). *A family of distributions \mathcal{F} is complete if*

$$E[f(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})] = 0 \quad \text{for all } F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}$$

implies

$$f(a, e_1, e_2) = 0 \quad \text{almost everywhere (a.e.) } \mathcal{F}.$$

Proposition 3.2 (Necessary condition for parallel trends for any distribution of unobservables). *Suppose that Assumptions SP and SEL hold. Suppose further that $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}$, where \mathcal{F} is a complete family of distributions satisfying $P(\varepsilon_{i1} \neq \varepsilon_{i2}) = 1$, $E[\varepsilon_{i1}] = E[\varepsilon_{i2}] = 0$, and $P(G_i = 1) \in (0, 1)$. If Assumption PT holds for all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}$, then $P(G_i = 1 | \varepsilon_{i1}, \varepsilon_{i2}) = P(G_i = 1)$ a.s. for all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}$.*

In Proposition 3.2, we require $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$ to belong to a complete family of distributions, \mathcal{F} . Completeness requires that the class of possible distributions of unobservables is rich enough. This condition allows us to obtain a necessary condition for Assumption PT holding for all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}$.

Taken together, Propositions 3.1 and 3.2 show that the PT assumption cannot hold absent additional restrictions on the selection mechanism and the distribution of unobservables. In particular, these results highlight the role of restrictions on time-varying unobservables, either in terms of how they vary over time or how they determine selection. As a result, researchers using DiD approaches cannot avoid making meaningful and nontrivial assumptions on selection and time-varying unobservables.

A natural question is which combinations of restrictions on the selection mechanism and unobservables are consistent with parallel trends. Since Propositions 3.1 and 3.2 are stated for all $g(\cdot)$ and all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$, respectively, they do not provide an answer to this question. In Section 3.3, we therefore complement our necessary conditions with a menu of restrictions on $g(\cdot)$ and $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$ that imply parallel trends.

3.3 Sufficient conditions for parallel trends

The goal in this section is to provide a menu of interpretable restrictions on the selection mechanism and the distribution of unobservables that are sufficient for the PT assumption.

To illustrate our general sufficient conditions, we will use a simple threshold-crossing selection mechanism inspired by the selection mechanisms considered by Ashenfelter and Card (1985). Suppose that individuals select into treatment if the discounted sum of untreated potential outcomes is below a certain cutoff.

$$G_i = 1 \{Y_{i1}(0) + \beta Y_{i2}(0) \leq c\} = 1 \{(1 + \beta)\alpha_i + \varepsilon_{i1} + \beta\varepsilon_{i2} \leq \tilde{c}\}, \quad (6)$$

where $\beta \in [0, 1]$ is a discount factor and $\tilde{c} = c - \lambda_1 - \beta\lambda_2$. In Section 3.4, we consider general selection mechanisms, which also accommodate Roy-type selection on treatment effects.

In the first sufficient condition, Assumption SC1, selection can only depend on the time-invariant unobservable. In the following, for random variables Z and W with supports \mathcal{Z} and \mathcal{W} , respectively, 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 $(w, w') \in \mathcal{W}^2$.

Assumption SC1. *Suppose that the following conditions hold:*

1. $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ is a trivial function of ε_{i1} and ε_{i2} .
2. Either (i) $(\varepsilon_{i1}, \varepsilon_{i2}) \perp \alpha_i$ or (ii) $\varepsilon_{i1} | \alpha_i \stackrel{d}{=} \varepsilon_{i2} | \alpha_i$.

Assumption SC1.1 requires the selection mechanism to be a trivial function of the time-varying unobservables. Assumption SC1.2 requires the time-varying unobservables to be either (i) independent of α_i or (ii) time homogeneous.¹³ It is well-understood that the key advantage of panel data is that they allow researchers to account for time-invariant unobservable confounders, i.e. unobservables that can determine selection. Assumption SC1 corresponds to this classical type of selection. A simple example of a selection mechanism satisfying Assumption SC1.1 is $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = 1\{\alpha_i \leq c\}$, which corresponds to the selection mechanism in Ashenfelter and Card (1985, p.650).

It is interesting to compare Assumption SC1 to the strict exogeneity condition, which is the main identification condition in linear static panel models (e.g., Wooldridge, 2010, Section 10). In our setting, strict exogeneity on the potential outcome model is given by $E[\varepsilon_{it} | D_i, \alpha_i] = E[\varepsilon_{it} | G_i, \alpha_i] = E[\varepsilon_{it}] = 0$ for $t = 1, 2$, where the first equality follows by the definition of G_i . Since selection only depends on α_i , independence between time-invariant

¹³We borrow the terminology “time homogeneity” from the nonseparable panel data literature (e.g., Chernozhukov, Fernández-Val, Hahn, and Newey, 2013; Ghanem, 2017).

and time-varying unobservables, $\alpha_i \perp (\varepsilon_{i1}, \varepsilon_{i2})$, implies strict exogeneity. On the other hand, the time homogeneity restriction, $\varepsilon_{i1}|\alpha_i \stackrel{d}{=} \varepsilon_{i2}|\alpha_i$, implies $E[\varepsilon_{i1}|G_i, \alpha_i] = E[\varepsilon_{i2}|G_i, \alpha_i]$, which does not require mean independence of ε_{it} and (G_i, α_i) , $E[\varepsilon_{it}|G_i, \alpha_i] = E[\varepsilon_{it}]$.

Selection on time-varying unobservables is a major concern in DiD analyses. Therefore, we also consider sufficient conditions that allow selection into treatment to depend on time-varying unobservables. In the following sufficient condition, selection is determined by the time-invariant as well as the time-varying unobservables.

Assumption SC2. *The following conditions hold:*

1. $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = g(\alpha_i, \varepsilon_{i2}, \varepsilon_{i1})$.
2. $\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i \stackrel{d}{=} \varepsilon_{i2}, \varepsilon_{i1}|\alpha_i$.

While Assumption SC2.1 allows for selection based on time-varying unobservables, it requires them to enter the selection mechanism symmetrically. For the simple selection mechanism in Equation (6), Assumption SC2.1 requires that there is no discounting, $\beta = 1$, so that

$$G_i = 1 \{Y_{i1}(0) + Y_{i2}(0) \leq c\} = 1 \{2\alpha_i + \varepsilon_{i1} + \varepsilon_{i2} \leq \tilde{c}\}. \quad (7)$$

Assumption SC2.2 imposes that the joint distribution of the time-varying unobservables conditional on α_i is exchangeable. We emphasize that the exchangeability restriction we impose here is different from the restrictions previously exploited in Altonji and Matzkin (2005); see also Arkhangelsky and Imbens (2019, 2022) for related assumptions. We further discuss the connection to Altonji and Matzkin (2005) in Section 5.

The last sufficient condition we consider allows selection into treatment to depend on the time-invariant unobservable and the time-varying unobservable in the first period only.

Assumption SC3. *The following conditions hold:¹⁴*

1. $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ is a trivial function of ε_{i2} .
2. $(\varepsilon_{i2} - \varepsilon_{i1}) \perp (\alpha_i, \varepsilon_{i1})$.

In Assumption SC3.1, the time-varying unobservable in the second period does not affect selection. For the selection mechanism in Equation (6), Assumption SC2.1 requires that $Y_{i2}(0)$ does not affect selection into treatment, $\beta = 0$, so that

$$G_i = 1 \{Y_{i1}(0) \leq c\} = 1 \{\alpha_i + \varepsilon_{i1} \leq \tilde{c}\}. \quad (8)$$

¹⁴Note that the roles of ε_{i1} and ε_{i2} can be reversed so that selection depends on ε_{i2} but not on ε_{i1} .

This corresponds to the selection mechanism considered in Ashenfelter and Card (1985, p.651).

Assumption SC3.2 requires the change in the time-varying unobservables to be independent of the determinants of selection $(\alpha_i, \varepsilon_{i1})$. The conditional mean version of this assumption, $E[\varepsilon_{i2} - \varepsilon_{i1} | \alpha_i, \varepsilon_{i1}] = 0$, implies a martingale property and, thus, restricts the time series dependence.

The following proposition shows that either of these three sufficient conditions implies the PT assumption.

Proposition 3.3 (Sufficient conditions for parallel trends). *Suppose that Assumptions SP and SEL hold. Then (i) Assumption SC1 implies Assumption PT, (ii) Assumption SC2 implies Assumption PT, and (iii) Assumption SC3 implies Assumption PT.*

In sum, this section provides a menu of interpretable sufficient conditions for PT that take the form of combinations of restrictions on the selection mechanism and the distribution of unobservables. In empirical practice, these conditions can be used to justify parallel trends based on contextual knowledge about the selection mechanism. See Section 6 for a discussion.

3.4 General selection mechanisms

Our menu of sufficient conditions is tailored to the case where selection is function of the unobservable determinants of the untreated potential outcomes. This is a consequence of the PT assumption being imposed on the untreated potential outcomes, since most DiD applications focus on the ATT. Here we extend our analysis to more general selection mechanisms

$$G_i = \check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}), \quad (9)$$

where (η_{i1}, η_{i2}) are vectors of time-invariant and time-varying unobservables that do not determine the untreated potential outcomes. This allows us to accommodate additional selection mechanisms, including Roy-style selection based on treatment effects and other selection mechanisms based on economic decision problems (e.g. Heckman and Robb, 1985; Chabé-Ferret, 2015).

The following corollary to Proposition 3.3 shows that Assumptions SC1, SC2, and SC3 imply the PT assumption where $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ is a conditional expectation function, specifically,

$$g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) \equiv \int \check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, t_1, t_2) dF_{\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}(t_1, t_2). \quad (10)$$

Corollary 3.1 (Sufficient conditions for parallel trends). *Suppose that Assumption SP holds. Suppose further that $P(G_i = 1) \in (0, 1)$. Then (i) Assumption SC1 implies Assumption PT, (ii) Assumption SC2 implies Assumption PT, and (iii) Assumption SC3 implies Assumption PT, where in each assumption $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) \equiv E[G_i | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}]$ as defined in (10).*

The corollary clarifies that Assumptions SC1.1, SC2.1, and SC3.1 extend in a straightforward manner to the case where $g(\cdot)$ is a conditional expectation function. These assumptions can therefore be interpreted as high-level conditions on $g(\cdot)$, the projected selection mechanism. For parallel trends to hold, the restrictions on $g(\cdot)$ have to be accompanied by the distributional restrictions on $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ in Assumptions SC1.2, SC2.2, and SC3.2, respectively.

Corollary 3.1 does not provide primitive conditions on $\check{g}(\cdot)$ and $F_{\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$ that imply the high-level conditions on the projected selection mechanism $g(\cdot)$. The next proposition gives such primitive sufficient conditions for Assumptions SC1.1, SC2.1, and SC3.1.

Proposition 3.4 (General selection mechanisms). *Suppose Assumption SP holds. Suppose further that*

$$G_i = \check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}), \quad (11)$$

where $P(G_i = 1) \in (0, 1)$.

- (i) *If $\check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2})$ is a trivial function of $(\varepsilon_{i1}, \varepsilon_{i2})$ and $(\eta_{i1}, \eta_{i2}) \perp (\varepsilon_{i1}, \varepsilon_{i2}) | \alpha_i$, then $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ is a trivial function of ε_{i1} and ε_{i2} .*
- (ii) *If $\check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}) = \check{g}(\alpha_i, \varepsilon_{i2}, \varepsilon_{i1}, \eta_{i1}, \eta_{i2})$ and $\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2} \stackrel{d}{=} \eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i2}, \varepsilon_{i1}$, then $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = g(\alpha_i, \varepsilon_{i2}, \varepsilon_{i1})$.*
- (iii) *If $\check{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2})$ is a trivial function of ε_{i2} and $\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2} \stackrel{d}{=} \eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}$, then $g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$ is a trivial function of ε_{i2} .*

The conditions in Proposition 3.4(i) ensure that neither the selection mechanism nor the conditional distribution of (η_{i1}, η_{i2}) depend on $(\varepsilon_{i1}, \varepsilon_{i2})$. Together, they imply that the conditional expectation function $g(\cdot)$ is a trivial function of $(\varepsilon_{i1}, \varepsilon_{i2})$. Proposition 3.4(ii) imposes symmetry in ε_{i1} and ε_{i2} on both the selection mechanism $\check{g}(\cdot)$ and the conditional distribution $F_{\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$, implying symmetry of $g(\cdot)$ in $(\varepsilon_{i1}, \varepsilon_{i2})$. Finally, the conditions in Proposition 3.4(iii) ensure that neither the selection mechanism $\check{g}(\cdot)$ nor the conditional distribution $F_{\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$ depend on ε_{i2} so that $g(\cdot)$ is a trivial function of ε_{i2} . Importantly, none of the sufficient conditions in Proposition 3.4 imposes any restrictions on how η_{i1} and η_{i2} determine selection.

Finally, we illustrate the sufficient conditions on $\check{g}(\cdot)$ using a simple example of Roy-style selection on treatment effects.

Example 1 (Roy-style selection). *Consider the following random coefficients model for the observed outcome*

$$Y_{it} = \alpha_i + \delta_{it}D_{it} + \lambda_t + \epsilon_{it}, \quad (12)$$

and suppose that selection depends on the treatment effects (η_{i1}, η_{i2}) as well as an individual-specific cost to treatment, c_i , that may depend on α_i

$$G_i = 1\{f(\delta_{i1}, \delta_{i2}) > c_i\}. \quad (13)$$

Since the selection mechanism in (13) does not depend on $(\epsilon_{i1}, \epsilon_{i2})$, the conditions on $\check{g}(\cdot)$ in Proposition 3.4(i)-(iii) hold. We therefore only have to impose the distributional restrictions in Proposition 3.4(i)-(iii). Specifically, for Assumption SC1.1 to hold, it is sufficient that the treatment effects and costs are independent of the time-varying unobservable determinants of the untreated potential outcome conditional on α_i , that is $(\delta_{i1}, \delta_{i2}, c_i) \perp (\epsilon_{i1}, \epsilon_{i2}) | \alpha_i$. For Assumption SC2.1 to hold, the conditional distribution of the treatment effects and costs has to be exchangeable in $(\epsilon_{i1}, \epsilon_{i2})$, i.e., $(\delta_{i1}, \delta_{i2}, c_i) | (\alpha_i, \epsilon_{i1}, \epsilon_{i2}) \stackrel{d}{=} (\delta_{i1}, \delta_{i2}, c_i) | (\alpha_i, \epsilon_{i2}, \epsilon_{i1})$. Finally, for Assumption SC3.1 to hold, the treatment effects and costs have to be independent of ϵ_{i2} conditional on $(\alpha_i, \epsilon_{i1})$, formally $(\delta_{i1}, \delta_{i2}, c_i) | (\alpha_i, \epsilon_{i1}, \epsilon_{i2}) \stackrel{d}{=} (\delta_{i1}, \delta_{i2}, c_i) | (\alpha_i, \epsilon_{i1})$.

Overall, the example with Roy-style selection illustrates that the time-varying treatment effects can enter the selection mechanism in an unrestricted way. In fact, $f(\cdot)$ can depend on δ_{i1} and δ_{i2} asymmetrically (e.g., $f(\cdot)$ could be a trivial function of one or the other). As a result, in the context of Roy-style selection, researchers do not have to impose any restrictions on how the selection mechanism depends on the treatment effects or costs. Instead, they have to justify the required distributional restrictions. \square

Remark 3 (Covariates and nonseparable models). *The results here can be extended to accommodate covariates (Section 4). In this case, the conditional expectation of G_i would be taken after conditioning on the covariates in addition to $(\alpha_i, \epsilon_{i1}, \epsilon_{i2})$. By similar arguments, our results on the nonseparable models (Section 5) extend to more general selection mechanisms.* \square

4 Covariates in the separable model

In this section, we introduce covariates in a separable model. The necessary conditions in Propositions 3.1 and 3.2 can be extended in a straightforward manner to incorporate covariates. Therefore, we focus on sufficient conditions here. We first provide conditions under which the unconditional parallel trends assumption (Assumption PT) continues to hold. We then allow covariates to enter the selection mechanism and give primitive conditions under which the conditional parallel trends assumption (Assumption PT-X) holds. The results have implications for the choice of covariates to be included in DiD analyses, which we discuss in detail in Section 6.

4.1 Model

Throughout this section, we consider the following separable model, which includes covariates, X_{it} .

Assumption SP-X.

$$Y_{it}(0) = \alpha_i + \lambda_t + \gamma_t(X_{it}) + \varepsilon_{it}, \quad E[\varepsilon_{it}] = 0, \quad i = 1, \dots, N, \quad t = 1, 2. \quad (14)$$

Assumption SP-X extends the separable model to include covariates, both time-invariant and time-varying. The model allows for nonparametric covariate-specific trends, which is a key reason for incorporating covariates in DiD analyses. It nests commonly-used parametric specifications such as $\gamma_t(X_{it}) = X'_{it}\beta_t$ and $\gamma_t(X_{it}) = X'_{it}\beta$. Recall that we assume that the treatment does not affect X_{it} . In what follows, we denote the support of X_{it} as \mathcal{X} .

4.2 Sufficient conditions for unconditional parallel trends

A natural first question is whether and when unconditional parallel trends (Assumption PT) continues to hold despite covariates entering the outcome equation as in Assumption SP-X. To this end, we maintain the selection mechanism in Assumption SEL, which does not depend on covariates.

Proposition 4.1. *Suppose that Assumptions SP-X and SEL hold. If $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$ and $X_i \perp (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$, then Assumption PT holds.*

Proposition 4.1 shows that if selection is orthogonal to the change in the time-varying unobservables, and covariates are independent of all unobservables, then unconditional par-

allel trends continues to hold under Assumption SP-X.¹⁵ While Proposition 4.1 demonstrates a case where covariates may not be required for identification, even if there are covariate-specific trends, the conditions are restrictive. Not only do they rule out covariates determining selection, but they also require them to be independent of all unobservable determinants of the outcome and selection models.¹⁶

4.3 Sufficient conditions for conditional parallel trends

Here we allow covariates to determine selection into treatment and give interpretable sufficient conditions for Assumption PT-X. Specifically, we consider the following selection mechanism.

Assumption SEL-X. $G_i = g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})$, where $P(G_i = 1|X_i) \in (0, 1)$ a.s.

In the following, we weaken the three sets of sufficient conditions discussed in Section 3.3 by including covariates in the selection mechanism and conditioning on them in the distributional assumptions.

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

1. $g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})$ is a trivial function of ε_{i1} and ε_{i2} .
2. Either (i) $(\varepsilon_{i1}, \varepsilon_{i2}) \perp \alpha_i | X_i$ or (ii) $\varepsilon_{i1} | \alpha_i, X_i \stackrel{d}{=} \varepsilon_{i2} | \alpha_i, X_i$.

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

1. $g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2}) = g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i2}, \varepsilon_{i1})$.
2. $\varepsilon_{i1}, \varepsilon_{i2} | \alpha_i, X_i \stackrel{d}{=} \varepsilon_{i2}, \varepsilon_{i1} | \alpha_i, X_i$.

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

1. $g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})$ is a trivial function of ε_{i2} .
2. $(\varepsilon_{i2} - \varepsilon_{i1}) \perp (\alpha_i, \varepsilon_{i1}) | X_i$.

Assumptions SC1-X, SC2-X, and SC3-X demonstrate that incorporating time-varying covariates is crucial for making the restrictions on the selection mechanism more plausible.

¹⁵The independence between covariates and unobservables can be relaxed to the arguably less interpretable high-level condition $E[G_i|X_i] = E[G_i]$ a.s.

¹⁶This independence condition and the assumptions on selection have testable implications: $(X_{i1}, X_{i2}) \perp G_i$. The implied tests are reminiscent of “balance tests” in randomized experiments. However, here the testable restrictions concern only the covariates and not the pre-treatment outcomes.

None of the assumptions impose any restrictions on how the time-varying covariates determine selection, making them more plausible in applications than Assumptions SC1, SC2, and SC3. For instance, in Assumption SC1-X.1, while we rule out selection on time-varying unobservables, we can allow the selection mechanism to depend on time-varying covariates. Similarly, in Assumption SC2-X.1, the time-varying unobservables have to enter the selection mechanism symmetrically, whereas the time-varying covariates do not have to obey any symmetry restrictions.

Conditioning on covariates also weakens the distributional restrictions in Assumptions SC1-X, SC2-X, and SC3-X relative to Assumptions SC1, SC2, and SC3. For example, the independence and time homogeneity conditions in Assumption SC1-X.2 are more likely to be satisfied once we focus on subpopulations with the same time-invariant covariates and the same evolution of time-varying covariates.

The following proposition shows that Assumptions SC1-X, SC2-X, and SC3-X are sufficient for the PT-X assumption.

Proposition 4.2. *Suppose that Assumptions SP-X and SEL-X hold. Then (i) Assumption SC1-X implies Assumption PT-X, (ii) Assumption SC2-X implies Assumption PT-X, and (iii) Assumption SC3-X implies Assumption PT-X.*

Proposition 4.2 provides several avenues for justifying the inclusion of covariates in DiD analyses. In particular, it shows that time-invariant and time-varying covariates play different roles in satisfying the PT-X assumption. Any time-varying factors that asymmetrically affect selection should be included as covariates. In addition, practitioners should include time-invariant and time-varying covariates that render the distributional restrictions plausible in their application.

The conclusions in this section crucially depend on the separability between covariates and the unobservables that determine selection. Assumption PT-X will generally not hold in models where α_i and X_{it} interact. A simple example is a correlated random coefficients model (e.g., Chamberlain, 1992), $Y_{it}(0) = \alpha_i X_{it} + \lambda_t + \varepsilon_{it}$, where the scalar X_{it} and α_i enter the selection mechanism. In the next section, we relax the separability restriction and demonstrate the implications of relaxing separability for the type of conditional parallel trends assumptions that can hold in this setting.

5 Selection in a nonseparable model with covariates

So far, we have studied separable models to keep the presentation transparent. However, since DiD is a model-agnostic reduced-form approach, it is crucial to generalize the results

to nonseparable models for practical and theoretical reasons. In doing so, we establish interesting connections between our sufficient conditions for parallel trends and identifying assumptions from the nonseparable panel literature. Our results have important implications for the choice of covariates in DiD analyses. See Section 6 for further discussions.

The necessary conditions we provide in Section 3 illustrate the trade-offs between restrictions on selection and the time-varying unobservables. Because the nonseparable model nests the separable model as a special case, these trade-offs remain relevant for this more general class of models as we show in Appendix A. We therefore focus on sufficient conditions for Assumption PT-NSP in this section.

5.1 Model

We consider the following nonseparable model, which nests the models in Assumptions SP and SP-X.

Assumption NSP.

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

where X_{it}^μ , X_{it}^λ , α_i^μ , α_i^λ , ε_{it}^μ , and ε_{it}^λ are finite-dimensional vector-valued random variables.

The above model consists of time-invariant and time-varying nonseparable components. Without further restrictions on the unobservables, the additive structure in Assumption NSP is without loss of generality and the superscripts μ and λ are merely labels. Indeed, it is possible that $X_{it}^\mu = X_{it}^\lambda$, $\alpha_i^\mu = \alpha_i^\lambda$, and $\varepsilon_{it}^\mu = \varepsilon_{it}^\lambda$, which implies that the model is fully nonseparable and time-varying in an arbitrary way. In the following, we denote the supports of X_{it}^μ , X_{it}^λ , α_i^μ , and ε_{it}^μ using \mathcal{X}_μ , \mathcal{X}_λ , \mathcal{A} , and \mathcal{E} , respectively.

To extend our analysis of the separable models, it is natural to consider selection based on the unobservables entering $\mu(\cdot)$, since they can be viewed as the counterparts of the unobservables in the separable model.¹⁷ We therefore consider the following selection mechanism, which can depend on all covariates, whereas it can only depend on the unobservables that enter the time-invariant component of the structural function.

Assumption SEL-NSP.

$$G_i = g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu),$$

where $P(G_i = 1 | X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu) \in (0, 1)$ a.s.

¹⁷To see this, note that the separable model in Assumption SP-X is nested in Assumption NSP by setting $\mu(X_{it}^\mu, \alpha_i^\mu, \varepsilon_{it}^\mu) = \alpha_i^\mu + \varepsilon_{it}^\mu$ and $\lambda_t(X_{it}^\lambda, \alpha_i^\lambda, \varepsilon_{it}^\lambda) = \lambda_t + \gamma_t(X_{it}^\lambda)$.

5.2 Sufficient conditions

Our first sufficient condition generalizes Assumptions SC1 and SC1-X, accounting for the different types of unobservables and covariates in the nonseparable model.

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

1. $g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)$ is a trivial function of ε_{i1}^μ and ε_{i2}^μ .
2. $\varepsilon_{i1}^\mu | \alpha_i^\mu, X_i^\mu, X_i^\lambda \stackrel{d}{=} \varepsilon_{i2}^\mu | \alpha_i^\mu, X_i^\mu, X_i^\lambda$.
3. $\alpha_i^\mu \perp (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda) | X_i^\mu, X_i^\lambda$.

Assumption SC1-NSP imposes different conditions on the unobservables, depending on whether they enter the time-invariant or time-varying component of the structural function, $\mu(\cdot)$ and $\lambda_t(\cdot)$, respectively. The distribution of ε_{it}^μ , which enters the time-invariant component, is required to be time homogeneous conditional on $(\alpha_i^\mu, X_i^\mu, X_i^\lambda)$. On the other hand, the unobservables that enter the time-varying component, $(\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda)$, are required to be independent of the unobservables that determine selection, α_i^μ , conditional on (X_i^μ, X_i^λ) .

The next sufficient condition allows selection to depend on all unobservables that enter the time-invariant component of the structural function, imposing a symmetry restriction on the selection mechanism similar to Assumptions SC2 and SC2-X.

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

1. $g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu) = g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i2}^\mu, \varepsilon_{i1}^\mu)$.
2. $\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu | \alpha_i^\mu, X_i^\mu, X_i^\lambda \stackrel{d}{=} \varepsilon_{i2}^\mu, \varepsilon_{i1}^\mu | \alpha_i^\mu, X_i^\mu, X_i^\lambda$.
3. $(\alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu) \perp (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda) | X_i^\mu, X_i^\lambda$.

Here, we require the distribution of $(\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)$ to be exchangeable conditional on $(\alpha_i^\mu, X_i^\mu, X_i^\lambda)$. Since the 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^μ, X_i^λ) .

The exchangeability restriction in Assumption SC2-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 non-parametric correlated random effects restriction (Ghanem, 2017). By contrast, we assume

that the time-varying unobservables are exchangeable conditional on $(\alpha_i^\mu, X_i^\mu, X_i^\lambda)$ without imposing any restrictions on the distribution of $\alpha_i^\mu | G_i, X_i^\mu, X_i^\lambda$.¹⁸

To generalize Assumptions SC3 and SC3-X, we require the selection mechanism to be a trivial function of ε_{i2}^μ in the following sufficient condition.

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

1. $g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)$ is a trivial function of ε_{i2}^μ .
2. $(\alpha_i^\mu, \varepsilon_{i1}^\mu) \perp \Delta_{\mu,i} | X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu$, where $\Delta_{\mu,i} \equiv \mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)$.
3. $(\alpha_i^\mu, \varepsilon_{i1}^\mu) \perp (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda) | X_i^\mu, X_i^\lambda$.

Assumption SC3-NSP.2 implicitly imposes separability conditions on $\mu(\cdot)$ and restrictions on time series dependence. The independence condition in Assumption SC3-NSP.3 requires that the unobservables that determine selection are independent of the unobservables that enter $\lambda_t(\cdot)$ conditional on the times series of the covariates.

Looking across the three sufficient conditions, we can see that all of them consist of three components: (i) a restriction on selection, (ii) a restriction on the unobservables entering the selection equation, 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 the covariates.

The next proposition shows that each of these conditions is sufficient for Assumption PT-NSP.

Proposition 5.1. *Suppose that Assumptions NSP and SEL-NSP hold. 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.*

The proof is based on two observations. First, the combination of conditions on selection and $(\alpha_i^\mu, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)$ ensures that we can cancel out the components involving $\mu(\cdot)$ when taking

¹⁸In fact, the exchangeability assumption in Altonji and Matzkin (2005) is not suitable for identification in our setting with a never-treated control group. To provide an example, abstract from covariates and consider a setting where we have a richer set of treatment trajectories $D_i = (D_{i1}, D_{i2}) \in \{(0,0), (0,1), (1,0), (1,1)\}$, i.e., there are four instead of two groups. For the general nonseparable model $Y_{it}(0) = \xi_t(\alpha_i, \varepsilon_{it})$, the exchangeability restriction in Altonji and Matzkin (2005) is given by $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) | (D_{i1}, D_{i2}) \stackrel{d}{=} (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) | (D_{i2}, D_{i1})$. As a result, one would identify the average treatment effect for the subpopulation $D_i = (1,0)$ using a simple between-group difference as follows $E[Y_{i1}(1) - Y_{i1}(0) | D_i = (1,0)] = E[Y_{i1} | D_i = (1,0)] - E[Y_{i1} | D_i = (0,1)]$ thereby using the subpopulation $D_i = (0,1)$ as a control group for $D_i = (1,0)$ in period 1, and vice versa in period 2. The reason that taking a simple between-group difference is sufficient for identification is that the exchangeability assumption in Altonji and Matzkin (2005) essentially rules out “fixed effects” by imposing restrictions on the distribution of $\alpha_i | D_i$.

differences between the expected potential outcomes over time within each group conditional on covariates. In the nonseparable model, X_{it}^μ may interact with the unobservables determining selection. As a result, the treatment and the control group may not experience the same change in their untreated potential outcome after conditioning on the time series of X_{it}^μ and X_{it}^λ . In general, parallel trends will hold when we examine subpopulations that do not experience a change in X_{it}^μ . Second, independence between $\lambda_t(X_{it}^\lambda, \alpha_i^\lambda, \varepsilon_{it}^\lambda)$ and G_i ensures that the expected change in $\lambda_t(X_{it}^\lambda, \alpha_i^\lambda, \varepsilon_{it}^\lambda)$ is the same for both groups.

The theoretical results on the nonseparable model presented in this section have important implications for practice, particularly for the choice of covariates in DiD analyses. We further discuss these implications in Section 6.

5.3 Selection, fixed effects, and correlated random 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 in separable and nonseparable models, thereby implicitly allowing for selection on time-invariant unobservables.¹⁹ The main contribution of this paper is to analyze explicitly the connection between selection mechanisms and the parallel trends assumptions underlying DiD. Therefore, a natural question is how our sufficient conditions relate to the identification assumptions in the panel literature, and those pertaining to nonseparable models in particular.

The literature on nonseparable panel models has considered two broad categories of identification assumptions. First, time homogeneity conditions (e.g., Hoderlein and White, 2012; Chernozhukov, Fernández-Val, Hahn, and Newey, 2013) require the distribution of time-varying unobservables to be stationary across time while allowing for unrestricted individual heterogeneity (fixed effects). Second, correlated random effects restrictions (e.g., Altonji and Matzkin, 2005; Bester and Hansen, 2009) impose restrictions on individual heterogeneity, but can allow for unrestricted time heterogeneity. 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 parallel trends constitute interpretable primitive conditions on the selection mechanism that imply combinations of time homogeneity and correlated random effects restrictions from the nonseparable panel literature. To simplify

¹⁹See, e.g., Chamberlain (1984); Arellano (2003); Evdokimov (2010); Wooldridge (2010); Hoderlein and White (2012); Chernozhukov, Fernández-Val, Hahn, and Newey (2013).

exposition, we abstract from covariates.

The following assumption is the time homogeneity assumption from Chernozhukov, Fernández-Val, Hahn, and Newey (2013) imposed on the unobservables that enter $\mu(\cdot)$, the time-invariant component of the structural function in Assumption NSP.

Assumption TH. $\varepsilon_{i1}^\mu | G_i, \alpha_i^\mu \stackrel{d}{=} \varepsilon_{i2}^\mu | G_i, \alpha_i^\mu$

Assumption TH requires the distribution of ε_{it}^μ to be time homogeneous conditional on G_i and α_i^μ . However, it does not impose any restrictions on the conditional distribution of ε_{i1}^μ given G_i and α_i^μ . Furthermore, there are no restrictions imposed on the distribution of $\alpha_i^\mu | G_i$ consistent with the notion of fixed effects in the nonseparable panel literature (Evdokimov, 2010; Hoderlein and White, 2012; Chernozhukov, Fernández-Val, Hahn, and Newey, 2013).

The next assumption is a nonparametric random effects assumption (e.g., Altonji and Matzkin, 2005; Ghanem, 2017), which generalizes the classical notion of (correlated) random effects (e.g., Mundlak, 1978; Chamberlain, 1984).

Assumption RE. $(\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda) | G_i \stackrel{d}{=} (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda)$.

Assumption RE is an independence condition between G_i and the unobservables that enter the time-varying component of the structural function, $\lambda_t(\cdot)$. This does not imply random assignment, $(Y_{i1}(0), Y_{i2}(0)) \perp G_i$, since selection into treatment can depend on the unobservables entering the time-invariant component $\mu(\cdot)$. With covariates, Assumption RE corresponds to a correlated random effects restriction that takes the form of a conditional independence assumption.

It is straightforward to show that Assumptions TH and RE imply Assumption PT. In the following proposition, we show that Assumptions SC1-NSP and SC2-NSP are primitive sufficient conditions on the selection mechanism for the nonseparable model satisfying Assumptions TH and RE.²⁰ This result demonstrates how restrictions on the selection mechanism can be used to justify combinations of Assumptions TH and RE.

Proposition 5.2. *Suppose that Assumptions NSP and SEL-NSP hold with $X_{it}^\mu = X_{it}^\lambda = \emptyset$. Then (i) Assumption SC1-NSP implies Assumptions TH and RE, and (ii) Assumption SC2-NSP implies Assumptions TH and RE if $P(G_i = 1 | \alpha_i^\mu = a) \in (0, 1)$ for all $a \in \mathcal{A}$.*

Given the selection mechanisms in Assumptions SC1-NSP and SC2-NSP, the results in Proposition 5.2 follow from two observations. First, the restrictions on the unobservables that enter $\mu(\cdot)$ constitute primitive conditions for Assumption TH. Second, the independence

²⁰In the context of correlated random coefficient models, Graham and Powell (2012) impose a similar structure on their random coefficient. They assume that the random coefficient consists of a time-invariant and time-varying component.

between the unobservables determining selection and those entering $\lambda_t(\cdot)$ are primitive conditions for Assumption RE.

Proposition 5.2 sheds light on the connection between selection, fixed effects, and correlated random effects in our nonseparable DiD framework. On the one hand, Assumptions SC1-NSP and SC2-NSP allow the distribution of $\alpha_i^\mu|G_i$ to be unrestricted, consistent with the notion of fixed effects. On the other hand, both conditions require $(\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda)$ to be independent of the determinants of selection and therefore independent of G_i , consistent with the notion of random effects.

6 Implications for practice

This paper studies the connection between selection and parallel trends in DiD analyses. We first provide necessary conditions that demonstrate that researchers relying on parallel trends assumptions implicitly impose restrictions on how selection depends on unobservables. We then derive sufficient conditions on selection for parallel trends assumptions with and without covariates. These conditions provide empirical practitioners with new and explicit theory-based templates for justifying parallel trends assumptions based contextual information on the selection mechanism.

Our necessary conditions for parallel trends demonstrate a trade-off between assumptions on selection and the presence of time-varying unobservables. First, we show that absent any restrictions on how the selection mechanism depends on the unobservable determinants of the outcome, parallel trends rules out time-varying unobservables. Conversely, for parallel trends to hold for any distribution of (time-invariant and time-varying) unobservables, selection needs to be independent of time-varying unobservables.

This analysis highlights that to justify parallel trends in applications, researchers have to impose restrictions on selection and time-varying unobservables. We therefore provide different sets of interpretable conditions that imply parallel trends. Our conditions consist of combinations of restrictions on (i) which/how unobservables determine selection and (ii) how their distribution varies over time. We recommend that applied researchers relying on our conditions use contextual information to assess and explicitly discuss which determinants of the outcome affect selection. Once a suitable selection mechanism is identified, the next step is to discuss the plausibility of the corresponding assumption on the distribution of the unobservables. In this context, periodicity is crucial both to distinguish between time-invariant and time-varying factors and to justify the distributional assumptions. These restrictions are typically more plausible the closer the pre- and post-treatment period are.

Our theoretical results have important implications for the role of covariates in DiD anal-

yses. First, they clarify how covariates can weaken restrictions on selection. We show that time-varying covariates do not have to obey the strong symmetry and exclusion restrictions required for time-varying unobservables. Thus, researchers should include time-varying factors that asymmetrically determine selection into treatment. Second, conditioning on covariates makes the distributional restrictions on unobservables weaker/more plausible. However, even after conditioning on covariates, researchers still have to take a stance on how selection depends on the remaining unobservables.

Finally, by analyzing parallel trends through the lens of nonseparable panel models, we demonstrate the implications of separability restrictions on the outcome model for how researchers should condition on covariates in their DiD analyses. If covariates and unobservable determinants of selection enter the outcome model separably, researchers should condition on the entire time series of covariates. If, in addition, there are covariates that interact with the unobservable determinants of selection in the outcome model, researchers have to condition on these covariates not changing over time. Indeed, even if treatment and control groups experience the same change in these covariates, the two groups may not exhibit the same counterfactual trends.

References

- ABADIE, A. (2005): “Semiparametric Difference-in-Differences Estimators,” *The Review of Economic Studies*, 72(1), 1–19.
- ALTONJI, J. G., AND R. L. MATZKIN (2005): “Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors,” *Econometrica*, 73(4), 1053–1102.
- ANGRIST, J. D., AND A. B. KRUEGER (1999): “Chapter 23 - Empirical Strategies in Labor Economics,” vol. 3 of *Handbook of Labor Economics*, pp. 1277–1366. Elsevier.
- ARELLANO, M. (2003): *Panel Data Econometrics*. Oxford University Press.
- ARELLANO, M., AND S. BONHOMME (2011): “Nonlinear Panel Data Analysis,” *Annual Review of Economics*, 3, 395–424.
- ARELLANO, M., AND S. BONHOMME (2012): “Identifying Distributional Characteristics in Random Coefficients Panel Data Models,” *The Review of Economic Studies*, 79(3), 987–1020.
- ARELLANO, M., AND B. HONORÉ (2001): “Panel Data Models: Some Recent Developments,” in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 5. Elsevier Science.
- ARKHANGELSKY, D., AND G. W. IMBENS (2019): “The Role of the Propensity Score in Fixed Effect Models,” *arXiv:1807.02099 [econ]*.

- (2022): “Double-Robust Identification for Causal Panel Data Models,” *arXiv:1909.09412 [econ]*.
- ARKHANGELSKY, D., G. W. IMBENS, L. LEI, AND X. LUO (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 D. CARD (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., M. BAYATI, N. DOUDCHENKO, G. IMBENS, AND K. KHOSRAVI (2021): “Matrix Completion Methods for Causal Panel Data Models*,” *Journal of the American Statistical Association*, 116(536), 1716–1730.
- ATHEY, S., AND G. W. IMBENS (2006): “Identification and Inference in Nonlinear Difference-in-Differences Models,” *Econometrica*, 74(2), 431–497.
- BESTER, C. A., AND C. HANSEN (2009): “Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model,” *Journal of Business and Economic Statistics*, 27(2), 235–250.
- BONHOMME, S., AND U. SAUDER (2011): “Recovering Distributions in Difference-in-Differences models: A Comparison of Selective and Comprehensive Schooling,” *Review of Economics and Statistics*, 93(May), 479–494.
- BORUSYAK, K., X. JARAVEL, AND J. SPIESS (2021): “Revisiting Event Study Designs: Robust and Efficient Estimation,” *arXiv:2108.12419 [econ]*.
- CAETANO, C., B. CALLAWAY, S. PAYNE, AND H. S. RODRIGUES (2022): “Difference in Differences with Time-Varying Covariates,” *arXiv:2202.02903*.
- CALLAWAY, B., AND T. LI (2019): “Quantile treatment effects in difference in differences models with panel data,” *Quantitative Economics*, 10(4), 1579–1618.
- CALLAWAY, B., T. LI, AND T. OKA (2018): “Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods,” *Journal of Econometrics*, 206(2), 395–413, Special issue on Advances in Econometric Theory: Essays in honor of Takeshi Amemiya.
- CALLAWAY, B., AND P. H. C. SANT’ANNA (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 A. B. KRUEGER (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.
- CHAMBERLAIN, G. (1982): “Multivariate regression models for panel data,” *Journal of Econometrics*, 18(1), 5–46.
- (1984): “Chapter 22 Panel data,” vol. 2 of *Handbook of Econometrics*, pp. 1247–1318. Elsevier.
- (1992): “Efficiency Bounds for Semiparametric Regression,” *Econometrica*, 60(3), 567–596.
- CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. HAHN, AND W. NEWEY (2013): “Average and Quantile Effects in Nonseparable Panel Models,” *Econometrica*, 81(2), 535–580.
- DE CHAISEMARTIN, C., AND X. D’HAULTFŒUILLE (2017): “Fuzzy Differences-in-Differences,” *The Review of Economic Studies*, 85(2), 999–1028.
- (2020): “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects,” *American Economic Review*, 110(9), 2964–2996.
- (2021): “Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey,” SSRN Scholarly Paper ID 3980758, Social Science Research Network, Rochester, NY.
- D’HAULTFŒUILLE, X., S. HODERLEIN, AND Y. SASAKI (2021): “Nonparametric difference-in-differences in repeated cross-sections with continuous treatments,” arXiv preprint arXiv:2104.14458.
- 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.
- GOODMAN-BACON, A. (2021): “Difference-in-differences with variation in treatment timing,” *Journal of Econometrics*, 225(2), 254–277.
- GRAHAM, B. S., AND J. L. POWELL (2012): “Identification and estimation of average partial effects in “irregular” correlated random coefficient panel data models,” *Econometrica*, 80(5), 2105–2152.

- HECKMAN, J. J., H. ICHIMURA, AND P. E. TODD (1997): “Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *The Review of Economic Studies*, 64(4), 605–654.
- HECKMAN, J. J., AND R. ROBB (1985): “Alternative methods for evaluating the impact of interventions: An overview,” *Journal of Econometrics*, 30(1), 239–267.
- HODERLEIN, S., AND H. WHITE (2012): “Nonparametric identification in nonseparable panel data models with generalized fixed effects,” *Journal of Econometrics*, 168(2), 300–314.
- HONORÉ, B., AND E. KYRIAZIDOU (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 E. KYRIAZIDOU (2000b): “Panel Data Discrete Choice Models with Lagged Dependent Variables,” *Econometrica*, 68(4), 839–874.
- KYRIAZIDOU, E. (1997): “Estimation of a Panel Data Sample Selection Model,” *Econometrica*, 65(6), 1335–1364.
- LECHNER, M. (2010): “The Estimation of Causal Effects by Difference-in-Difference Methods,” *Foundations and Trends in Econometrics*, 4(3), 165–224.
- LEHMANN, E., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. Springer Texts in Statistics.
- MANSKI, C. F. (1987): “Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data,” *Econometrica*, 55(2), 357–362.
- MARCUS, M., AND P. H. C. SANT’ANNA (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.
- 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., W. K. VISCUSI, AND D. L. DURBIN (1995): “Workers’ Compensation and Injury Duration: Evidence from a Natural Experiment,” *The American Economic Review*, 85(3), 322–340.
- MUNDLAK, Y. (1961): “Empirical Production Function Free of Management Bias,” *Journal of Farm Economics*, 43(1), 44–56.
- (1978): “On the Pooling of Time Series and Cross Section Data,” *Econometrica*, 46(1), 69–85.
- ROTH, J., AND P. H. C. SANT’ANNA (2021): “When Is Parallel Trends Sensitive to

- Functional Form?,” *arXiv:2010.04814 [econ, stat]*.
- ROTH, J., P. H. C. SANT’ANNA, A. BILINSKI, AND J. POE (2022): “What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature,” *arXiv:2201.01194 [econ]*.
- SANT’ANNA, P. H. C., AND J. ZHAO (2020): “Doubly robust difference-in-differences estimators,” *Journal of Econometrics*, 219(1), 101–122.
- SUN, L., AND S. ABRAHAM (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. (2021): “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” *Working Paper*.

Appendix to *Selection and Parallel Trends*

Dalia Ghanem Pedro H. C. Sant’Anna Kaspar Wüthrich

A	Necessary conditions for nonseparable models	1
B	Proofs	4
B.1	Proof of Lemma 3.1	4
B.2	Proof of Proposition 3.1	5
B.3	Proof of Proposition 3.2	6
B.4	Proof of Proposition 3.3	7
B.5	Proof of Corollary 3.1	8
B.6	Proof of Proposition 3.4	8
B.7	Proof of Proposition 4.1	9
B.8	Proof of Proposition 4.2	10
B.9	Proof of Proposition 5.1	11
B.10	Proof of Proposition 5.2	14
B.11	Supplementary lemmas	15

A Necessary conditions for nonseparable models

Here we provide necessary conditions for parallel trends in the nonseparable model. To simplify the exposition, we abstract from covariates. We derive the necessary conditions in the context of a fully nonseparable, time-varying outcome model and a general selection mechanism that can depend on all unobservable determinants of the outcome as well as additional unobservables (η_{i1}, η_{i2}) .

Assumption A.1 (Nonseparable model).

$$Y_{it}(0) = \xi_t(\alpha_i, \varepsilon_{it}), \quad i = 1, \dots, N, \quad t = 1, 2,$$

where α_i , ε_{i1} and ε_{i2} are finite-dimensional vector-valued random variables.

Assumption A.2 (Selection mechanism). $G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2})$, where $P(G_i = 1) \in (0, 1)$.

To simplify exposition, we use $\tilde{Y}_{it}(0)$ to denote the centered potential outcome without the treatment, $\tilde{Y}_{it}(0) = Y_{it}(0) - E[Y_{it}(0)]$ for $t = 1, 2$. Furthermore, let \mathcal{T} denote the support of η_{it} for $t = 1, 2$.

Proposition A.1 (Necessary condition for parallel trends for any selection mechanism). *Suppose that Assumptions A.1 and A.2 hold. Suppose further that $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2})$ is a vector of non-degenerate random variables satisfying $(\eta_{i1}, \eta_{i2}) \perp (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$, $P(\tilde{Y}_{i1}(0) \leq \tilde{Y}_{i2}(0)) > 0$, and $P(\tilde{Y}_{i1}(0) \geq \tilde{Y}_{i2}(0)) > 0$. If Assumption PT holds for all $g : \mathcal{A} \times \mathcal{E}^2 \times \mathcal{T}^2 \mapsto \{0, 1\}$, then $\tilde{Y}_{i1}(0) = \tilde{Y}_{i2}(0)$ w.p.1.*

Proof. First, note that, by definition, $\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) = (\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+ - (\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-$, where $(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+ = |\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0)|1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) > 0\}$ and $(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^- = |\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0)|1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) < 0\}$. Now we can re-write $E[G_i(Y_{i2}(0) - Y_{i1}(0))] = E[G_i]E[Y_{i2}(0) - Y_{i1}(0)]$ (the equivalent condition for Assumption PT by Lemma B.2) as

$$E[G_i(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))] = 0,$$

which is equivalent to

$$E[G_i(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] = E[G_i(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-]. \quad (15)$$

If Assumption PT holds for any $g : \mathcal{A} \times \mathcal{E}^2 \times \mathcal{T}^2 \mapsto \{0, 1\}$, then it holds for (i) $G_i = 1\{\bar{\eta}_i > c\}1\{\xi_2(\alpha_i, \varepsilon_{i2}) - \xi_1(\alpha_i, \varepsilon_{i1}) \geq E[Y_{i2}(0) - Y_{i1}(0)]\} = 1\{\bar{\eta}_i > c\}1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}$ and (ii) $G_i = 1\{\bar{\eta}_i > c\}1\{\xi_2(\alpha_i, \varepsilon_{i2}) - \xi_1(\alpha_i, \varepsilon_{i1}) \leq E[Y_{i2}(0) - Y_{i1}(0)]\} = 1\{\bar{\eta}_i > c\}1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \leq 0\}$, where $\bar{\eta}_i = (\eta_{i1} + \eta_{i2})/2$ and c is chosen such that $P(\bar{\eta}_i > c) \in (0, 1)$. We consider both cases separately.

(i) Consider

$$\begin{aligned} & E[1\{\bar{\eta}_i > c\}1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] \\ &= E[1\{\bar{\eta}_i > c\}1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-], \\ & P(\bar{\eta}_i > c)E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] \\ &= P(\bar{\eta}_i > c)E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-], \\ & E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] \\ &= E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-], \end{aligned} \quad (16)$$

where the second equality follows by the independence of $(\eta_{i1}, \eta_{i2}) \perp (\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$. The last equality follows from $P(\bar{\eta}_i > c) > 0$. Since the term on the right-hand side equals zero, $E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \geq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] = E[(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] = 0$. Since by definition $(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+ \geq 0$, it follows that $P(\tilde{Y}_{i2}(0) > \tilde{Y}_{i1}(0)) = 0$.

(ii) Similarly, consider

$$\begin{aligned}
& E[1\{\bar{\eta}_i > c\}1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \leq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] \\
& = E[1\{\bar{\eta}_i > c\}1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \leq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-], \\
& E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \leq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^+] \\
& = E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \leq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-]. \tag{17}
\end{aligned}$$

Now the left-hand side in the above equality equals zero, $E[1\{\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0) \leq 0\}(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^-] = 0$. Since by definition $(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))^- \geq 0$, it follows that $P(\tilde{Y}_{i2}(0) < \tilde{Y}_{i1}(0)) = 0$.

Statements (i) and (ii) together imply that $\tilde{Y}_{i2}(0) = \tilde{Y}_{i1}(0)$ w.p.1, which completes the proof. □

Proposition A.2 (Necessary condition for parallel trends for any distribution of unobservables). *Suppose that Assumptions A.1 and A.2 hold. Suppose further that $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}} \in \mathcal{F}$, where \mathcal{F} is a complete family of probability distributions satisfying $P(\tilde{Y}_{i1}(0) = \tilde{Y}_{i2}(0)) = 0$ and $P(g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}) = 1) \in (0, 1)$. If Assumption PT holds for any $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}} \in \mathcal{F}$, then $P(G_i = 1 | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = P(G_i = 1)$ a.s. for all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}} \in \mathcal{F}$.*

Proof. By Lemma B.2, Assumption PT is equivalent to $E[G_i(\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0))] = 0$, which in turn is equivalent to the following under Assumption A.2

$$E[\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}))] = 0, \tag{18}$$

where $\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}) = E[g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}) - E[G_i | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}]]$ and $\tilde{\xi}_t(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = \xi_t(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - E[Y_{it}(0)]$. The equivalence between $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$ and (18) follows by the law of iterated expectations and subtracting $E[G_i]E[\tilde{Y}_{i2}(0) - \tilde{Y}_{i1}(0)]$, noting that it equals zero by construction.

It follows that Assumption PT holds for all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}} \in \mathcal{F}$ is equivalent to

$$E[\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}))] = 0, \tag{19}$$

for all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}} \in \mathcal{F}$. By completeness of \mathcal{F} , the last equality implies (Lehmann and Romano, 2005, p.115) that

$$P(\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})) = 0) = 1 \quad \text{for all } F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}} \in \mathcal{F}. \tag{20}$$

Now note that the left-hand side of (20) can be simplified as follows,

$$\begin{aligned}
& P(\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = 0) \\
& = P(\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})), \tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})) \\
& \quad + P(\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})), \tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) \neq \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})) \\
& = P(\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})) | \tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) \neq \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})) \\
& = P(\bar{g}(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = 0) = 1, \tag{21}
\end{aligned}$$

where the penultimate equality follows since $P(\tilde{\xi}_2(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) \neq \tilde{\xi}_1(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})) = P(\tilde{Y}_{i2}(0) \neq \tilde{Y}_{i1}(0)) = 1$ by assumption. As a result,

$$P(E[g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}) | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}] = E[G_i]) = 1 \quad \text{for all } F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \eta_{i1}, \eta_{i2}} \in \mathcal{F}. \tag{22}$$

□

B Proofs

B.1 Proof of Lemma 3.1

Under Assumption SP, Assumption PT can be rewritten as

$$E[\varepsilon_{i2} - \varepsilon_{i1} | G_i = 1] = E[\varepsilon_{i2} - \varepsilon_{i1} | G_i = 0].$$

It follows that Assumption PT holds if and only if $E[\varepsilon_{i2} - \varepsilon_{i1} | G_i] = E[\varepsilon_{i2} - \varepsilon_{i1}] = 0$, where the last equality follows since $E[\varepsilon_{it}] = 0$ for $t = 1, 2$. It remains to show that (a) $E[\varepsilon_{i2} - \varepsilon_{i1} | G_i] = 0$ if and only if (b) $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$. Since $P(G_i = 1) \in (0, 1)$, the conclusion that (a) \Rightarrow (b) is immediate. Therefore, we are left to show that (b) \Rightarrow (a). Note that $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$ implies $E[\varepsilon_{i2} - \varepsilon_{i1} | G_i = 1] = 0$ since

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = E[\varepsilon_{i2} - \varepsilon_{i1} | G_i = 1]P(G_i = 1)$$

and $P(G_i = 1) \in (0, 1)$ by assumption. To show that $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$ implies $E[\varepsilon_{i2} - \varepsilon_{i1} | G_i = 0] = 0$, we subtract $E[\varepsilon_{i2} - \varepsilon_{i1}]$ from both sides of the former equality, multiply by -1 , which yields

$$E[(1 - G_i)(\varepsilon_{i2} - \varepsilon_{i1})] = 0$$

since $E[(\varepsilon_{i2} - \varepsilon_{i1})] = 0$ (Assumption SP). This implies $E[\varepsilon_{i2} - \varepsilon_{i1}|G_i = 0] = 0$ since $P(G_i = 0) \in (0, 1)$ and, by definition,

$$E[\varepsilon_{i2} - \varepsilon_{i1}|G_i = 0] = \frac{E[(1 - G_i)(\varepsilon_{i2} - \varepsilon_{i1})]}{P(G_i = 0)}.$$

□

B.2 Proof of Proposition 3.1

By Lemma 3.1, it suffices to show that $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$. To this end, first, note that, by definition, $\varepsilon_{i2} - \varepsilon_{i1} = (\varepsilon_{i2} - \varepsilon_{i1})^+ - (\varepsilon_{i2} - \varepsilon_{i1})^-$, where $(\varepsilon_{i2} - \varepsilon_{i1})^+ = |\varepsilon_{i2} - \varepsilon_{i1}|1\{\varepsilon_{i2} - \varepsilon_{i1} > 0\}$ and $(\varepsilon_{i2} - \varepsilon_{i1})^- = |\varepsilon_{i2} - \varepsilon_{i1}|1\{\varepsilon_{i2} - \varepsilon_{i1} \leq 0\}$. As a result, $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$ is equivalent to

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})^+] = E[G_i(\varepsilon_{i2} - \varepsilon_{i1})^-]. \quad (23)$$

If Assumption PT holds for all $g : \mathcal{A} \times \mathcal{E}^2 \mapsto \{0, 1\}$, then it holds for (i) $G_i = 1\{\alpha_i > c\}1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}$ and (ii) $G_i = 1\{\alpha_i > c\}1\{\varepsilon_{i2} - \varepsilon_{i1} \leq 0\}$, where c is chosen such that $P(\alpha_i > c) \in (0, 1)$. We consider both cases separately.

(i) Consider

$$\begin{aligned} E[1\{\alpha_i > c\}1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^+] &= E[1\{\alpha_i > c\}1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^-] \\ P(\alpha_i > c)E[1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^+] &= P(\alpha_i > c)E[1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^-] \\ E[1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^+] &= E[1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^-] \end{aligned} \quad (24)$$

The second equality follows by the independence of α_i and $(\varepsilon_{i1}, \varepsilon_{i2})$. The last equality follows by $P(\alpha_i > c) > 0$. Since the term on the right-hand side of the last equality equals zero, $E[1\{\varepsilon_{i2} - \varepsilon_{i1} \geq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^+] = E[(\varepsilon_{i2} - \varepsilon_{i1})^+] = 0$. Since by definition $(\varepsilon_{i2} - \varepsilon_{i1})^+ \geq 0$, it follows that $P(\varepsilon_{i2} > \varepsilon_{i1}) = 0$.

(ii) Similarly, consider

$$\begin{aligned} E[1\{\alpha_i > c\}1\{\varepsilon_{i2} - \varepsilon_{i1} \leq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^+] &= E[1\{\alpha_i > c\}1\{\varepsilon_{i2} - \varepsilon_{i1} \leq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^-] \\ E[1\{\varepsilon_{i2} - \varepsilon_{i1} \leq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^+] &= E[1\{\varepsilon_{i2} - \varepsilon_{i1} \leq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^-]. \end{aligned} \quad (25)$$

Now the left-hand side in the last equality equals zero, $E[1\{\varepsilon_{i2} - \varepsilon_{i1} \leq 0\}(\varepsilon_{i2} - \varepsilon_{i1})^-] = E[(\varepsilon_{i2} - \varepsilon_{i1})^-] = 0$. Since by definition $(\varepsilon_{i2} - \varepsilon_{i1})^- \geq 0$, it follows that $P(\varepsilon_{i2} < \varepsilon_{i1}) = 0$.

Statements (i) and (ii) together imply that $\varepsilon_{i1} = \varepsilon_{i2}$ w.p.1, which completes the proof. \square

B.3 Proof of Proposition 3.2

Recall that \mathcal{F} is a complete family of distributions satisfying $P(\varepsilon_{i1} = \varepsilon_{i2}) = 0$, $E[\varepsilon_{i1}] = E[\varepsilon_{i2}] = 0$, and $P(g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) = 1) \in (0, 1)$.

By Lemma 3.1, Assumption PT is equivalent to $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$, which in turn is equivalent to

$$E[\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1})] = 0, \quad (26)$$

where $\bar{g}(\varepsilon_{i1}, \varepsilon_{i2}) = E[g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) - E[g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})] | \varepsilon_{i1}, \varepsilon_{i2}]$. This follows by the law of iterated expectations and subtracting $E[G_i]E[\varepsilon_{i2} - \varepsilon_{i1}]$, noting that it equals zero by assumption.

It follows that Assumption PT holding for all $F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}$ is equivalent to

$$E[\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1})] = 0 \quad \text{for all } F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}. \quad (27)$$

By completeness of \mathcal{F} , the last equality implies (Lehmann and Romano, 2005, p.115) that

$$P(\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1}) = 0) = 1 \quad \text{for all } F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}. \quad (28)$$

Now note that the left-hand side of (28) can be simplified as follows,

$$\begin{aligned} & P(\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1}) = 0) \\ &= P(\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1}) = 0, \varepsilon_{i1} = \varepsilon_{i2}) + P(\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1}) = 0, \varepsilon_{i1} \neq \varepsilon_{i2}) \\ &= P(\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1}) = 0 | \varepsilon_{i1} \neq \varepsilon_{i2}) \\ &= P(\bar{g}(\varepsilon_{i1}, \varepsilon_{i2}) = 0) = 1, \end{aligned} \quad (29)$$

where the penultimate equality follows since $P(\varepsilon_{i1} \neq \varepsilon_{i2}) = 1$ by assumption. As a result, by the definition of $\bar{g}(\varepsilon_{i1}, \varepsilon_{i2})$, it follows that

$$P(E[g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}) | \varepsilon_{i1}, \varepsilon_{i2}] = E[G_i]) = 1 \quad \text{for all } F_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}} \in \mathcal{F}. \quad (30)$$

\square

B.4 Proof of Proposition 3.3

We prove the three statements separately. By Lemma 3.1, it suffices to show that $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$.

(i) First, consider Assumption SC1 with $(\varepsilon_{i1}, \varepsilon_{i2}) \perp \alpha_i$. Then

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = E[g(\alpha_i)(\varepsilon_{i2} - \varepsilon_{i1})] = E[g(\alpha_i)]E[\varepsilon_{i2} - \varepsilon_{i1}] = 0, \quad (31)$$

where the first equality follows from Assumption SC1.1.²¹ The second equality follows from Assumption SC1.2(i), and the last follows from $E[\varepsilon_{it}] = 0$ for $t = 1, 2$.

Second, consider Assumption SC1 with $\varepsilon_{i1}|\alpha_i \stackrel{d}{=} \varepsilon_{i2}|\alpha_i$, then

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = \int g(a)e_2 dF_{\alpha_i, \varepsilon_{i2}}(a, e_2) - \int g(a)e_1 dF_{\alpha_i, \varepsilon_{i1}}(a, e_1) = 0, \quad (32)$$

where the first equality follows from Assumption SC1.1 and the last equality follows from Assumption SC1.2(ii), which implies $F_{\alpha_i, \varepsilon_{i1}} = F_{\alpha_i, \varepsilon_{i2}}$. As a result, Assumption SC1 implies Assumption PT.

(ii) Consider

$$\begin{aligned} & E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] \\ &= \int \left(\int g(a, e_1, e_2)e_2 dF_{\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i}(e_1, e_2|a) - \int g(a, e_1, e_2)e_1 dF_{\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i}(e_1, e_2|a) \right) dF_{\alpha_i}(a) \\ &= \int \left(\int g(a, e_2, e_1)e_2 dF_{\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i}(e_2, e_1|a) - \int g(a, e_1, e_2)e_1 dF_{\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i}(e_1, e_2|a) \right) dF_{\alpha_i}(a) = 0, \end{aligned}$$

where the first equality follows from the definition of the selection mechanism under Assumption SC2.1. The second equality follows from the symmetry restrictions on $g(\cdot)$ and $F_{\varepsilon_{i1}, \varepsilon_{i2}|\alpha_i}$ in Assumption SC2, which imply that the difference in the conditional expectations in parenthesis equals zero. As a result, Assumption SC2 implies Assumption PT.

(iii) Consider

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = E[g(\alpha_i, \varepsilon_{i1})(\varepsilon_{i2} - \varepsilon_{i1})] = E[g(\alpha_i, \varepsilon_{i1})]E[\varepsilon_{i2} - \varepsilon_{i1}] = 0, \quad (33)$$

²¹Here and below, to avoid introducing additional notation, we write $g(a, e_1, e_2) = g(a)$ with a slight abuse of notation.

where the first equality follows from Assumption SC3.1 and the second equality follows from the independence of $(\varepsilon_{i2} - \varepsilon_{i1}) \perp (\alpha_i, \varepsilon_{i1})$ in Assumption SC3.2. As a result, Assumption SC3 implies Assumption PT. \square

B.5 Proof of Corollary 3.1

By Lemma 3.1, Assumption PT is equivalent to

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0. \quad (34)$$

By the law of iterated expectations,

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = E[E[G_i|\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}](\varepsilon_{i2} - \varepsilon_{i1})] \equiv E[g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\varepsilon_{i2} - \varepsilon_{i1})]. \quad (35)$$

Since the proof of Proposition 3.3 does not rely on $g(\cdot)$ being a binary variable, Assumptions SC1, SC2, and SC3 imply that the right-hand side of the last equality is zero by identical arguments to those in the proof of Proposition 3.3, which completes the proof. \square

B.6 Proof of Proposition 3.4

(i) The following holds for $(a, e_1, e_2, e'_1, e'_2) \in \mathcal{A} \times \mathcal{E}^4$, where $e_1 \neq e'_1$ and $e_2 \neq e'_2$,

$$\begin{aligned} g(a, e_1, e_2) &= \int \check{g}(a, e_1, e_2, t_1, t_2) dF_{\eta_{i1}, \eta_{i2}|\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}(t_1, t_2|a, e_1, e_2) \\ &= \int \check{g}(a, e'_1, e'_2, t_1, t_2) dF_{\eta_{i1}, \eta_{i2}|\alpha_i}(t_1, t_2|a) = g(a, e'_1, e'_2), \end{aligned} \quad (36)$$

where the penultimate equality follows by the definition of a trivial function and the conditional independence assumption imposed in (i).

(ii) The following holds for $(a, e_1, e_2) \in \mathcal{A} \times \mathcal{E}^2$

$$\begin{aligned} g(a, e_1, e_2) &= \int \check{g}(a, e_1, e_2, t_1, t_2) dF_{\eta_{i1}, \eta_{i2}|\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}(t_1, t_2|a, e_1, e_2) \\ &= \int \check{g}(a, e_2, e_1, t_1, t_2) dF_{\eta_{i1}, \eta_{i2}|\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}(t_1, t_2|a, e_2, e_1) = g(a, e_2, e_1), \end{aligned} \quad (37)$$

where the penultimate equality follows by the symmetry of $\check{g}(\cdot)$ and $F_{\eta_i|\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}$ in ε_{i1} and ε_{i2} in (ii).

(iii) The following holds for $(a, e_1, e_2, e'_2) \in \mathcal{A} \times \mathcal{E}^3$, $e_2 \neq e'_2$,

$$\begin{aligned} g(a, e_1, e_2) &= \int \check{g}(a, e_1, e_2, t_1, t_2) dF_{\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}, \varepsilon_{i2}}(t_1, t_2 | a, e_1, e_2) \\ &= \int \check{g}(a, e_1, e'_2, t_1, t_2) dF_{\eta_{i1}, \eta_{i2} | \alpha_i, \varepsilon_{i1}}(t_1, t_2 | a, e_1) = g(a, e_1, e'_2), \end{aligned} \quad (38)$$

where the penultimate equality follows from the definition of a trivial function and the conditional independence assumption in (iii). \square

B.7 Proof of Proposition 4.1

The proof proceeds in two steps. First, consider

$$\begin{aligned} E[G_i(Y_{i2}(0) - Y_{i1}(0))] &= E[G_i](\lambda_2 - \lambda_1) + E[G_i(\gamma_2(X_{i2}) - \gamma_1(X_{i1}))] + E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] \\ &= E[G_i](\lambda_2 - \lambda_1) + E[g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})(\gamma_2(X_{i2}) - \gamma_1(X_{i1}))] \\ &= E[G_i](\lambda_2 - \lambda_1) + E[G_i]E[\gamma_2(X_{i2}) - \gamma_1(X_{i1})] \\ &= E[G_i]E[Y_{i2}(0) - Y_{i1}(0)], \end{aligned} \quad (39)$$

where the second equality follows from $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})] = 0$ and Assumption SEL, the third equality follows from the independence of X_i and the unobservables, and the last one from the fact that, under Assumption SP-X, $E[Y_{i2}(0) - Y_{i1}(0)] = \lambda_2 - \lambda_1 + E[\gamma_2(X_{i2}) - \gamma_1(X_{i1})]$, since $E[\varepsilon_{i1}] = E[\varepsilon_{i2}] = 0$.

Equation (39) and $P(G_i = 1) > 0$ imply that

$$E[Y_{i2}(0) - Y_{i1}(0) | G_i = 1] = E[Y_{i2}(0) - Y_{i1}(0)]. \quad (40)$$

Note that

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

Imposing (40) yields

$$(1 - P(G_i = 1))E[Y_{i2}(0) - Y_{i1}(0)] = P(G_i = 0)E[Y_{i2}(0) - Y_{i1}(0) | G_i = 0], \quad (41)$$

which implies $E[Y_{i2}(0) - Y_{i1}(0) | G_i = 0] = E[Y_{i2}(0) - Y_{i1}(0)]$, since $P(G_i = 0) > 0$. This completes the proof. \square

B.8 Proof of Proposition 4.2

By Lemma B.1, it suffices to show that each assumption implies $E[G_i(\varepsilon_{i2} - \varepsilon_{i1})|X_i] = E[G_i|X_i]E[\varepsilon_{i2} - \varepsilon_{i1}|X_i]$ a.s.

(i) First, consider Assumption SC1-X with the conditional independence restriction, $(\varepsilon_{i1}, \varepsilon_{i2}) \perp \alpha_i | X_i$. Thus, it follows that, a.e. $(x_1, x_2) \in \mathcal{X}^2$,

$$\begin{aligned}
& E[G_i(\varepsilon_{i2} - \varepsilon_{i1})|X_i = (x_1, x_2)] \\
&= \int g(a, x_1, x_2)(e_2 - e_1)dF_{\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}|X_i}(a, e_1, e_2|x_1, x_2) \\
&= \int g(a, x_1, x_2)dF_{\alpha_i|X_i}(a|x_1, x_2) \int (e_2 - e_1)dF_{\varepsilon_{i1}, \varepsilon_{i2}|X_i}(e_1, e_2|x_1, x_2) \\
&= E[G_i|X_i = (x_1, x_2)]E[\varepsilon_{i2} - \varepsilon_{i1}|X_i = (x_1, x_2)], \tag{42}
\end{aligned}$$

where the first equality follows by Assumption SC1-X.1 and the second follows by Assumption SC1-X.2(i). The last equality follows by the definition of conditional expectations, which then implies Assumption PT-X holds.

Second, consider Assumption SC1-X with the conditional time homogeneity restriction, $\varepsilon_{i1}|\alpha_i, X_i \stackrel{d}{=} \varepsilon_{i2}|\alpha_i, X_i$. Thus, we have that, a.e. $(x_1, x_2) \in \mathcal{X}^2$,

$$\begin{aligned}
& E[G_i(\varepsilon_{i2} - \varepsilon_{i1})|X_i = (x_1, x_2)] \\
&= \int \left(\int g(a, x_1, x_2)e_2dF_{\varepsilon_{i2}|\alpha_i, X_i}(e_2|a, x_1, x_2) - \int g(a, x_1, x_2)e_1dF_{\varepsilon_{i1}|\alpha_i, X_i}(e_1|a, x_1, x_2) \right) dF_{\alpha_i|X_i}(a|x_1, x_2) \\
&= \int \left(\int g(a, x_1, x_2)e_2dF_{\varepsilon_{i1}|\alpha_i, X_i}(e_2|a, x_1, x_2) - \int g(a, x_1, x_2)e_1dF_{\varepsilon_{i1}|\alpha_i, X_i}(e_1|a, x_1, x_2) \right) dF_{\alpha_i|X_i}(a|x_1, x_2) \\
&= 0, \tag{43}
\end{aligned}$$

where the first equality follows from Assumption SC1-X.1 and the second follows from Assumption SC1-X.2(ii). The last equality is immediate and completes the proof.

(ii) The exchangeability restrictions in Assumption SC2-X imply the following:

$$\begin{aligned}
& E[g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})\varepsilon_{i1} | \alpha_i = a, X_i = (x_1, x_2)] \\
&= \int g(a, x_1, x_2, e_1, e_2) e_1 dF_{\varepsilon_{i1}, \varepsilon_{i2} | \alpha_i, X_i}(e_1, e_2 | a, x_1, x_2) \\
&= \int g(a, x_1, x_2, e_2, e_1) e_1 dF_{\varepsilon_{i1}, \varepsilon_{i2} | \alpha_i, X_i}(e_2, e_1 | a, x_1, x_2) \\
&= \int g(a, x_1, x_2, e_2, e_1) e_1 dF_{\varepsilon_{i1}, \varepsilon_{i2} | \alpha_i, X_i}(e_2, e_1 | a, x_1, x_2) \\
&= E[g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})\varepsilon_{i2} | \alpha_i = a, X_i = (x_1, x_2)], \tag{44}
\end{aligned}$$

a.e. $(a, x_1, x_2) \in \mathcal{A} \times \mathcal{X}^2$.

Integrating out $\alpha_i | X_i$ in the above yields the following a.e. equality:

$$\begin{aligned}
& \int E[g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})\varepsilon_{i1} | \alpha_i = a, X_{i1} = x_1, X_{i2} = x_2] dF_{\alpha_i | X_{i1}, X_{i2}}(a | x_1, x_2) \\
&= \int E[g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})\varepsilon_{i2} | \alpha_i = a, X_{i1} = x_1, X_{i2} = x_2] dF_{\alpha_i | X_{i1}, X_{i2}}(a | x_1, x_2). \tag{45}
\end{aligned}$$

As a result, by the law of iterated expectations we have that $E[G_i(\varepsilon_{i2} - \varepsilon_{i1}) | X_{i1}, X_{i2}] = 0$ almost surely. This completes the proof, since Assumptions SP-X and Assumption SC2-X.2 directly imply that $E[\varepsilon_{i2} - \varepsilon_{i1} | X_{i1}, X_{i2}] = 0$ a.s.

(iii) Consider

$$\begin{aligned}
& E[G_i(\varepsilon_{i2} - \varepsilon_{i1}) | X_i] \\
&= E[g(\alpha_i, X_{i1}, X_{i2}, \varepsilon_{i1})(\varepsilon_{i2} - \varepsilon_{i1}) | X_i] \\
&= E[G_i | X_i] E[\varepsilon_{i2} - \varepsilon_{i1} | X_i] \text{ a.s.}, \tag{46}
\end{aligned}$$

where the first and second a.s. equality follow from Assumption SC3-X.1 and SC3-X.2, respectively, which completes the proof. \square

B.9 Proof of Proposition 5.1

First, note that Lemma B.4 applies here by simply changing the conditioning set under Assumption SEL-NSP. As a result, Assumption PT-NSP under Assumption NSP and SEL-NSP holds iff

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

Here and below equalities involving conditional expectations are understood to hold a.s. We now proceed to show each statement separately.

(i): By Assumption NSP, SEL-NSP, and SC1-NSP.1, it follows that

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

where the first equality follows from Assumptions NSP, SEL-NSP and SC1-NSP.1. The second equality follows by applying the law of iterated expectations to the first term and the conditional independence imposed in Assumption SC1-NSP.3 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 SC1-NSP.2. The last equality follows from noting 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 law of iterated expectations.

(ii) By Assumption NSP and SEL-NSP, it follows that

$$\begin{aligned}
& E[G_i(Y_{i2}(0) - Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&= E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu) - \mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&\quad + E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu].
\end{aligned} \tag{47}$$

We first examine the first term on the RHS of the above equality. Note that by the symmetry

restrictions in Assumption SC2-NSP.1-2, it follows that a.e. $(a, x^\mu, x_1^\lambda, x_2^\lambda) \in \mathcal{A} \times \mathcal{X}_\mu \times \mathcal{X}_\lambda^2$

$$\begin{aligned}
& E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)\mu(X_{i1}^\mu, \alpha_i^\mu, \varepsilon_{i1}^\mu)|X_i^\lambda = (x_1^\lambda, x_2^\lambda), X_{i1}^\mu = X_{i2}^\mu = x^\mu, \alpha_i^\mu = a] \\
&= \int g(a, x^\mu, x^\mu, x_1^\lambda, x_2^\lambda, e_1, e_2)\mu(x^\mu, a, e_1)dF_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu|X_i^\lambda, X_{i1}^\mu=X_{i2}^\mu, \alpha_i^\mu}(e_1, e_2|(x_1^\lambda, x_2^\lambda), x^\mu, a) \\
&= \int g(a, x^\mu, x^\mu, x_1^\lambda, x_2^\lambda, e_2, e_1)\mu(x^\mu, a, e_1)dF_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu|X_i^\lambda, X_{i1}^\mu=X_{i2}^\mu, \alpha_i^\mu}(e_2, e_1|(x_1^\lambda, x_2^\lambda), x^\mu, a) \\
&= E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)\mu(X_{i2}^\mu, \alpha_i^\mu, \varepsilon_{i2}^\mu)|X_i^\lambda = (x_1^\lambda, x_2^\lambda), X_{i1}^\mu = X_{i2}^\mu = x^\mu, \alpha_i^\mu = a].
\end{aligned} \tag{48}$$

As a result, the first term in (47) equals zero by the law of iterated expectations and (48),

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

It follows that

$$\begin{aligned}
& E[G_i(Y_{i2}(0) - Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&= E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)(\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&= E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&= E[G_i|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu],
\end{aligned} \tag{49}$$

where the penultimate equality follows from the conditional independence assumption in Assumption SC2-NSP.3. 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 SC2-NSP.2 by Lemma B.3 and implies that $E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] = E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]$ as in the proof of (i). As a result, the above implies that Assumption PT-NSP holds.

(iii) By Assumptions NSP, SEL-NSP and SC3-NSP.1, it follows that

$$\begin{aligned}
& E[G_i(Y_{i2}(0) - Y_{i1}(0))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&= E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu)\Delta_{\mu,i}|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&\quad + E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu)(\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda))|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&= E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\Delta_{\mu,i}|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&\quad + E[g(\alpha_i^\mu, X_{i1}^\mu, X_{i2}^\mu, X_{i1}^\lambda, X_{i2}^\lambda, \varepsilon_{i1}^\mu)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[\lambda_2(X_{i2}^\lambda, \alpha_i^\lambda, \varepsilon_{i2}^\lambda) - \lambda_1(X_{i1}^\lambda, \alpha_i^\lambda, \varepsilon_{i1}^\lambda)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \\
&= E[G_i|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu]E[Y_{i2}(0) - Y_{i1}(0)|X_i^\lambda, X_{i1}^\mu = X_{i2}^\mu] \tag{50}
\end{aligned}$$

The penultimate equality follows from the conditional independence conditions in Assumption SC3-NSP.2-3, which jointly imply Assumption PT-NSP. This completes the proof. \square

B.10 Proof of Proposition 5.2

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

(i) To show the result, it suffices to show that (i.a) Assumptions SC1-NSP.1 and SC1-NSP.2 imply Assumption TH and (i.b) Assumptions SC1-NSP.1 and SC1-NSP.3 imply Assumption RE.

(i.a) Under Assumptions SC1-NSP.1 and SC1-NSP.2, $G_i = g(\alpha_i^\mu)$ is a degenerate random variable equaling either zero or one with probability one conditional on α_i^μ . As a result,

$$\begin{aligned}
F_{\varepsilon_{it}^\mu|G_i, \alpha_i^\mu}(e|g, a) &= \sum_{g=0,1} P(\varepsilon_{it}^\mu \leq e|G_i = g(a), \alpha_i^\mu = a)1\{g(a) = g\} \\
&= \sum_{g=0,1} P(\varepsilon_{it}^\mu \leq e|\alpha_i^\mu = a)1\{g(a) = g\} = \sum_{g=0,1} F_{\varepsilon_{it}^\mu|\alpha_i^\mu}(e|a)1\{g(a) = g\} \tag{51}
\end{aligned}$$

As a result, Assumption SC1-NSP.1 together with the time homogeneity of $F_{\varepsilon_{it}^\mu|\alpha_i^\mu}$ in Assumption SC1-NSP.2 is sufficient for the time homogeneity of $F_{\varepsilon_{it}^\mu|G_i, \alpha_i^\mu}(e|g, a)$, which yields Assumption TH.

(i.b) The statement (i.b) is immediate from noting that $G_i = g(\alpha_i^\mu)$ (Assumption SC1-NSP.1) and the independence condition in Assumption SC1-NSP.3 imply that $g(\alpha_i^\mu) \perp (\alpha_i^\lambda, \varepsilon_{i1}^\lambda, \varepsilon_{i2}^\lambda)$, which is equivalent to Assumption RE. This completes the proof of (i).

(ii) Here it suffices to show (ii.a) Assumptions SC2-NSP.1 and SC2-NSP.2 imply Assumption TH and (ii.b) Assumptions SC2-NSP.1 and SC2-NSP.3 imply Assumption RE.

(ii.a) Consider

$$F_{\varepsilon_{i1}^\mu, G_i | \alpha_i^\mu}(e_1, g|a) = F_{G_i | \varepsilon_{i1}^\mu, \alpha_i^\mu}(g|e_1, a) F_{\varepsilon_{i1}^\mu | \alpha_i^\mu}(e_1|a) \quad (52)$$

Assumption SC2-NSP.2 implies $F_{\varepsilon_{i1}^\mu | \alpha_i^\mu}(e|a) = F_{\varepsilon_{i2}^\mu | \alpha_i^\mu}(e|a)$ as well as $F_{\varepsilon_{i1}^\mu | \varepsilon_{i2}^\mu, \alpha_i^\mu}(e_1|e_2, a) = F_{\varepsilon_{i2}^\mu | \varepsilon_{i1}^\mu, \alpha_i^\mu}(e_1|e_2, a)$ by Lemma B.3, which implies

$$\begin{aligned} F_{G_i | \varepsilon_{i1}^\mu, \alpha_i^\mu}(g|e_1, a) &= \int 1\{g(a, e_1, e_2) \leq g\} dF_{\varepsilon_{i2}^\mu | \varepsilon_{i1}^\mu, \alpha_i^\mu}(e_2|e_1, a) \\ &= \int 1\{g(a, e_2, e_1) \leq g\} dF_{\varepsilon_{i1}^\mu | \varepsilon_{i2}^\mu, \alpha_i^\mu}(e_2|e_1, a) = F_{G_i | \varepsilon_{i2}^\mu, \alpha_i^\mu}(g|e_1, a) \end{aligned} \quad (53)$$

As a result,

$$\begin{aligned} F_{\varepsilon_{i1}^\mu, G_i | \alpha_i^\mu}(e_1, g|a) &= F_{G_i | \varepsilon_{i1}^\mu, \alpha_i^\mu}(g|e_1, a) F_{\varepsilon_{i1}^\mu | \alpha_i^\mu}(e_1|a) = F_{G_i | \varepsilon_{i2}^\mu, \alpha_i^\mu}(g|e_1, a) F_{\varepsilon_{i2}^\mu | \alpha_i^\mu}(e_1|a) \\ &= F_{\varepsilon_{i2}^\mu, G_i | \alpha_i^\mu}(e_1, g|a). \end{aligned} \quad (54)$$

This implies Assumption TH by the definition of a conditional distribution $F_{\varepsilon_{it}^\mu | G_i, \alpha_i^\mu}(e|g, a) = \frac{F_{\varepsilon_{it}^\mu, G_i | \alpha_i^\mu}(e, g|a)}{F_{G_i | \alpha_i^\mu}(g|a)}$, where $F_{G_i | \alpha_i^\mu}(g|a) > 0$ by assumption.

(ii.b) This statement follows in a straightforward manner from the definition of G_i in Assumption SC2-NSP.1 and the independence condition in Assumption SC2-NSP.3 which together imply Assumption RE. This completes the proof of (ii). \square

B.11 Supplementary lemmas

Lemma B.1. *Suppose SP-X and $P(G_i = 1 | X_i) \in (0, 1)$ a.s. hold. Assumption PT-X holds if and only if*

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1}) | X_i] = E[G_i | X_i] E[\varepsilon_{i2} - \varepsilon_{i1} | X_i] \text{ a.s.} \quad (55)$$

Proof. Since $P(G_i = 1 | X_i) \in (0, 1)$ a.s., Assumption PT-X holds iff

$$E[G_i(Y_{i2}(0) - Y_{i1}(0)) | X_i] = E[G_i | X_i] E[Y_{i2}(0) - Y_{i1}(0) | X_i] \text{ a.s.} \quad (56)$$

by arguments similar to Lemma 3.1 while conditioning on X_i . By Assumption SP-X, the left-hand side of the above simplifies to

$$\begin{aligned} &E[G_i(Y_{i2}(0) - Y_{i1}(0)) | X_i] \\ &= E[G_i | X_i](\lambda_2 - \lambda_1) + E[G_i | X_i](\gamma_2(X_i) - \gamma_1(X_i)) + E[G_i(\varepsilon_{i2} - \varepsilon_{i1}) | X_i] \text{ a.s.} \end{aligned}$$

As a result, Assumption PT-X holds iff

$$E[G_i(\varepsilon_{i2} - \varepsilon_{i1})|X_i] = E[G_i|X_i]E[\varepsilon_{i2} - \varepsilon_{i1}|X_i] \text{ a.s.} \quad (57)$$

□

Lemma B.2. *Suppose $P(G_i = 1) \in (0, 1)$ holds. Then, Assumption PT holds iff $E[G_i(Y_{i2}(0) - Y_{i1}(0))] = E[G_i]E[Y_{i2}(0) - Y_{i1}(0)]$.*

Proof. Assumption PT can be written as

$$E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] = E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0]. \quad (58)$$

As a result, Assumption PT holds iff $E[Y_{i2}(0) - Y_{i1}(0)|G_i] = E[Y_{i2}(0) - Y_{i1}(0)]$. It remains to show that (a) $E[Y_{i2}(0) - Y_{i1}(0)|G_i] = E[Y_{i2}(0) - Y_{i1}(0)]$ holds if and only if (b) $E[G_i(Y_{i2}(0) - Y_{i1}(0))] = E[G_i]E[Y_{i2}(0) - Y_{i1}(0)]$. The result that (a) \Rightarrow (b) is immediate. As for (b) \Rightarrow (a), $E[G_i(Y_{i2}(0) - Y_{i1}(0))] = E[G_i]E[Y_{i2}(0) - Y_{i1}(0)]$ implies $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] = E[Y_{i2}(0) - Y_{i1}(0)]$ by dividing the former equality by $E[G_i] = P(G_i = 1) \in (0, 1)$. It remains to show that $E[G_i(Y_{i2}(0) - Y_{i1}(0))] = E[G_i]E[Y_{i2}(0) - Y_{i1}(0)]$ implies $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0] = E[Y_{i2}(0) - Y_{i1}(0)]$. To do so, we subtract $E[Y_{i2}(0) - Y_{i1}(0)]$ from both sides of the previous equality and multiply by -1 .

$$E[(1 - G_i)(Y_{i2}(0) - Y_{i1}(0))] = E[1 - G_i]E[Y_{i2}(0) - Y_{i1}(0)] \quad (59)$$

Since $E[1 - G_i] = P(G_i = 0) \in (0, 1)$, the above implies that $E[Y_{i2}(0) - Y_{i1}(0)|G_i = 0] = E[Y_{i2}(0) - Y_{i1}(0)]$. This completes the proof. □

Lemma B.3. *Suppose that Assumption SC2-NSP.2 holds, then*

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

$$(ii) F_{\varepsilon_{i1}^\mu|\varepsilon_{i2}^\mu, \alpha_i^\mu}(e_1|e_2, a) = F_{\varepsilon_{i2}^\mu|\varepsilon_{i1}^\mu, \alpha_i^\mu}(e_1|e_2, a) \text{ a.e. } (a, e_1, e_2) \in \mathcal{A} \times \mathcal{E}^2.$$

Proof. (i) By the definition of the marginal distribution, Assumption SC2-NSP implies (i) by the following, a.e.

$$F_{\varepsilon_{i1}^\mu|\alpha_i^\mu}(e_1|a) = \lim_{e_2 \rightarrow \infty} F_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu|\alpha_i^\mu}(e_1, e_2|a) = \lim_{e_2 \rightarrow \infty} F_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu|\alpha_i^\mu}(e_2, e_1|a) = F_{\varepsilon_{i2}^\mu|\alpha_i^\mu}(e_1|a). \quad (60)$$

(ii) By the definition of the conditional distribution and (i) of this lemma, Assumption

SC2-NSP implies (ii) by the following

$$F_{\varepsilon_{i1}^\mu | \varepsilon_{i2}^\mu, \alpha_i^\mu}(e_1 | e_2, a) = \frac{F_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu | \alpha_i^\mu}(e_1, e_2 | a)}{F_{\varepsilon_{i2}^\mu | \alpha_i^\mu}(e_2 | a)} = \frac{F_{\varepsilon_{i1}^\mu, \varepsilon_{i2}^\mu | \alpha_i^\mu}(e_2, e_1 | a)}{F_{\varepsilon_{i1}^\mu | \alpha_i^\mu}(e_2 | a)} = F_{\varepsilon_{i2}^\mu | \varepsilon_{i1}^\mu, \alpha_i^\mu}(e_1 | e_2, a), \quad (61)$$

a.e. $(a, e_1, e_2) \in \mathcal{A} \times \mathcal{E} \times \mathcal{E}_a$, where \mathcal{E}_a is the support of $\varepsilon_{it}^\mu | \alpha_i^\mu$ for $t = 1, 2$. \square

Lemma B.4. *Suppose Assumptions NSP and $P(G_i = 1 | X_i) \in (0, 1)$ a.s. hold. Then, Assumption PT-X holds iff*

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

Proof. We first note that Assumption PT-X holds iff

$$E[Y_{i2}(0) - Y_{i1}(0) | G_i, X_i] = E[Y_{i2}(0) - Y_{i1}(0) | X_i] \text{ a.s.} \quad (62)$$

It remains to show that (a) $E[Y_{i2}(0) - Y_{i1}(0) | G_i, X_i] = E[Y_{i2}(0) - Y_{i1}(0) | X_i]$ a.s. iff (b) $E[G_i(Y_{i2}(0) - Y_{i1}(0)) | X_i] = E[G_i | X_i] E[Y_{i2}(0) - Y_{i1}(0) | X_i]$ a.s. The result that (a) \Rightarrow (b) follows immediately by the law of iterated expectations. As for (b) \Rightarrow (a),

$$E[G_i(Y_{i2}(0) - Y_{i1}(0)) | X_i] = E[G_i | X_i] E[Y_{i2}(0) - Y_{i1}(0) | X_i] \text{ a.s.} \quad (63)$$

implies that

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

since $P(G_i = 1 | X_i) \in (0, 1)$ a.s. Now subtracting $E[Y_{i2}(0) - Y_{i1}(0) | X_i]$ from both sides of (63) and multiplying by -1 yields

$$E[(1 - G_i)(Y_{i2}(0) - Y_{i1}(0)) | X_i] = E[(1 - G_i) | X_i] E[Y_{i2}(0) - Y_{i1}(0) | X_i] \text{ a.s.} \quad (64)$$

Since $P(G_i = 0 | X_i) \in (0, 1)$ a.s., the above implies

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

This completes the proof. \square