# Supplemental Appendix for Difference-in-Differences Designs: A Practitioner's Guide

Andrew Baker* Brantly Callaway† Scott Cunningham‡
Andrew Goodman-Bacon§ Pedro H. C. Sant'Anna¶

December 9, 2025

## A Some additional DiD-related procedures

This section discusses some important DiD-related topics that we did not cover in our main text. These discussions are short by design, and we focus on providing the main ideas related to challenges and solutions specific to the problem. We abstract from weights and use $\mathbb{E}[\cdot|\cdot]$ to denote (conditional) expectations.

### A.1 Setups with treatment turning on and off

Our main text focuses on setups where treatment remains in place from the period it begins until the end of the sample period, but in practice, some treatments turn on and off over time. This is the setting tackled by de Chaisemartin and D'Haultfoeuille (2020, 2023a), Imai, Kim and Wang (2023), and Liu, Wang and Xu (2024).

To tackle this problem from first principles, we need to augment the potential outcomes to reflect the richer notion of treatment *sequences*. Following Robins (1986), let $Y_{i,t}(\mathbf{d})$ denote the potential outcome for unit $i$ at time $t$ if this unit received the $T$-dimensional treatment sequence $\mathbf{d} \in \{0,1\}^T$. For simplicity, let's say that $T = 3$ and that no unit is treated in the first period. In this case, we have four treatment sequences (or histories), which define four potential outcomes for each unit: $Y_{i,t}(0,0,0), Y_{i,t}(0,0,1), Y_{i,t}(0,1,0)$ and $Y_{i,t}(0,1,1)$. We then define treatment groups by treatment sequences: $G = \mathbf{d}_0 \equiv (0,0,0)$ (never-treated), $G = \mathbf{d}_1 \equiv (0,0,1)$ (treated in the third

*University of California, Berkeley
†University of Georgia
‡Baylor University
§Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis
¶Emory University

period), $G = \mathbf{d}_2 \equiv (0,1,1)$ (treated in the second and third period), and $G = \mathbf{d}_3 \equiv (0,1,0)$ (treated only in the second period). In general, we would have as many groups as we have different (realized) treatment sequences. Recall that in a staggered timing design with an absorbing treatment, treatment timing fully characterizes a treatment sequence.

Once potential outcomes and groups are well-defined, one can move to parameters of interest. We proceed similarly to the staggered treatment setup in Section 4 and consider group-and-time specific ATTs as building blocks, except that groups are now based on more complex treatment sequences. Let $\mathbf{0}$ denote a T-dimensional vector of zeros. One intuitive building block parameter on which to base a DiD analysis is

$$ATT(\mathbf{d}, t) = \mathbb{E}[Y_t(\mathbf{d}) - Y_t(\mathbf{0})|G = \mathbf{d}],$$

the average treatment effect at time period $t$ of being exposed to treatment sequence $\mathbf{d}$ instead of never being exposed to treatment, among units that received treatment sequence $\mathbf{d}$.[1]

Next, one needs to establish identification for the parameters, and propose appropriate estimators and inference procedures. Following similar arguments to those in Section 4, a DiD approach to this problem would involve imposing a parallel trends assumption (potentially conditional on covariates) and a no-anticipation assumption to establish that the $ATT(\mathbf{d}, t)$'s are identified. If each treatment group is sufficiently large, one could proceed in a similar fashion as the staggered setup, comparing average outcome paths for a given sequence with the average outcome path for never-treated (or not-yet-treated) units. One can also aggregate these different $ATT(\mathbf{d}, t)$ to form different summary parameters.

In practice, however, it is often the case that the number of treatment groups is large and each group is small. This essentially creates a "curse of dimensionality" problem: there are too many building block parameters defined for too-small groups to be estimated reliably. In such cases, additional assumptions that limit treatment effect dynamics (or how past treatments affect future outcomes) are often imposed, and different aggregated summary parameters are usually targeted. We provide a brief overview of several different solutions that have been proposed to address this issue.

de Chaisemartin and D'Haultfoeuille (2020) impose a "no-carryover" assumption that implies that past treatments do not affect future outcomes; which is to say that treatment effects in a given period last only during that period. With such an assumption (in addition to parallel trends and a no-anticipation assumption), they propose DiD estimators for an instantaneous average treatment effects parameter by comparing currently treated units with untreated units. Imai et al. (2023) adopt a similar approach, though they impose a limited-carryover assumption where treatments may last for $\ell$ periods (with $\ell$ specified by the researcher). They then propose estimators for an average treatment effect of switching into treatment in period $t$ among units that experience the

[1] One could also adopt alternative building blocks not discussed here, such as the average effect of treatment lasting one period longer or a treatment spell of a given length beginning one period later.

policy change in period $t$, and share the same treatment history over the previous $k$ periods; see Liu et al. (2024) for a related procedure. Finally, de Chaisemartin and D'Haultfoeuille (2023a) avoid making assumptions related to carryover effects and extend the DiD framework in de Chaisemartin and D'Haultfoeuille (2020) to allow for treatment effect dynamics. The way they proceed is to first "staggerize" treatment sequences according to first-time of treatment exposure, compute a staggered DiD procedure for this "intention-to-treat" type parameter, and normalize them by a DiD estimate based on the number of treated periods. A potential challenge with de Chaisemartin and D'Haultfoeuille's (2023a) approach is the interpretability of their proposed summary parameter, though we should acknowledge that this is a complex setup.

One important takeaway is that comparing these DiD procedures for treatments to turn on and off may be challenging, as they target different causal parameters of interest, and practitioners should be aware of the different assumptions and limitations. We refer the reader to de Chaisemartin and D'Haultfoeuille (2023b) and Liu et al. (2024) for additional discussions on these types of DiD estimators.

## A.2 DiD setups with continuous or multi-valued treatments

Our paper focuses on binary treatments, but many treatments take multiple values or are even continuous. A number of recent papers have studied this particular type of treatment design. These include Callaway, Goodman-Bacon and Sant'Anna (2021, 2024); de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare (2024a); and de Chaisemartin, D'Haultfuille and Vazquez-Bare (2024b). Here we focus on a two-period setting in which no unit is treated in period one and some units receive a treatment with varying intensities (or doses) in period two. Most of the key results that distinguish multi-valued from binary treatments are evident with two periods (Callaway et al., 2024).

We now need to define potential outcomes that reflect varying treatment intensity. We denote $Y_{i,t}(0,d)$ as potential outcomes for unit $i$ in period $t$ if they are untreated in period one and receive treatment dosage $d$ in period two. As we focus on setups where all units are untreated in period one, we simplify notation and index potential outcomes by treatment intensity in period two; that is, $Y_{it}(d) = Y_{i,t}(0,d)$. An important feature is that $d$ is not restricted to $\{0,1\}$ and can take on richer treatment intensities instead. We denote the treatment dosage for unit $i$ as $D_i$ in period two and stress that in this context, our notion of the treatment group is tied to units' treatment dosage: groups are defined by their treatment dosage in period 2.

A multi-valued treatment defines several different types of causal parameters that may be of interest. For instance, dose-specific average treatment effect parameters such as

$$ATT(d|d') = \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(0)|D = d'] \quad \text{and} \quad ATE(d) = \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(0)],$$

reflect the average effect of dose $d$ relative to no treatment. Here $ATT(d|d')$ is the average treatment

effect for units that experienced dose $d'$; when $d' = d$, it is the $ATT$ among units that received dose $d$. On the right side, $ATE(d)$ is defined analogously, except that it is the effect on the overall population. Of course, one can also aggregate these dose-specific parameters to form more precisely estimable summary quantities; see, e.g., Callaway et al. (2021).

The two treatment effect parameters above provide average treatment effects in levels, and so one reason why they vary could be because $d$ itself varies. To account for the differences in $d$, one may be interested in "per-dosage" effects:

$$ATT_{\text{pd}}(d|d') = \frac{ATT(d|d')}{d} \quad \text{and} \quad ATE_{\text{pd}}(d) = \frac{ATE(d)}{d}.$$

One can also aggregate these parameters across dosages to analyze $\mathbb{E}[ATT_{\text{pd}}(D|D)|D > 0]$, an average treatment effect among treated (or, more generally, among switchers). One can also consider weighted averages of these to learn about $\mathbb{E}[ATT(D|D)|D > 0]/\mathbb{E}[D|D > 0]$; see de Chaisemartin et al. (2024a) for a general discussion about such target parameters.

Finally, researchers are often interested in the causal effect of a marginal increment in the dose. This notion is the average causal response (ACR), similar to Angrist and Imbens (1995), defined as follows (when the dose is absolutely continuous):

$$ACRT(d|d') = \frac{\partial ATT(l|d')}{\partial l}\bigg|_{l=d} = \frac{\partial\, \mathbb{E}[Y_{t=2}(l)|D = d']}{\partial l}\bigg|_{l=d} \text{ and } ACR(d) = \frac{\partial ATE(d)}{\partial d} = \frac{\partial\, \mathbb{E}[Y_{t=2}(d)]}{\partial d}.$$

Here $ACRT(d|d)$ equals the derivative of the average potential outcome in period two for units that received dose $d$ evaluated at $d$—this is equivalent to the derivative of $ATT(l|d)$ with respect to $l$, evaluated at $l = d$. We can interpret $ACR(d)$ analogously.[2]

The relevant questions pertain to (a) what assumptions are needed to impose to identify these parameters, (b) how to estimate and make inferences about these parameters of interest once identification is established, (c) how to summarize treatment effect heterogeneity across doses to generate interpretable aggregated causal parameters, and (d) whether traditional regression specifications based on TWFE recover a sensible and easy-to-understand causal parameter of interest. These questions are addressed in detail by Callaway et al. (2021) and de Chaisemartin et al. (2024a).

Callaway et al. (2021) highlight how, when no units are treated in period one, identification and estimation of $ATT(d|d)$'s (or their functionals) follows the binary case. They propose flexible nonparametric estimators for the $ATT(d|d)$ curve—the relationship between outcome changes (minus the average change for untreated units) and the dose $d$, making it possible to visualize and make inference about treatment effect heterogeneity across dosages. They also propose estimators that aggregate across dosage values and can be more precisely estimated. The identification of causal response parameters or ATE-type parameters, however, requires a stronger version of

[2]For discrete treatments, ACR's are defined in a similar way but with a slightly different notation to accommodate the discreteness of $d$: $ACRT(d_j|d_k) = \mathbb{E}[Y_{t=2}(d_j) - Y_{t=2}(d_{j-1})|D = d_k]$, and $ACR(d_j) = \mathbb{E}[Y_{t=2}(d_j) - Y_{t=2}(d_{j-1})]$.

parallel trends that holds for potential outcomes at non-zero treatment doses. Under these strong parallel trends and no anticipation assumptions, they discuss estimation and inference procedures for the ACR curves and their summary measures.[3]

de Chaisemartin et al. (2024a) consider the setup where units are already exposed to different levels of treatment in period one. They discuss how one can identify causal quantities that generalize $ATT_{\text{pd}}(d|d')$ to this more complex setup when (a) a sizable number of units do not change treatment dosage over time (stayers), and (ii) there is no-carryover from past treatment to future outcomes. They propose estimation and inference procedures for aggregated parameters akin to $\mathbb{E}[ATT_{\text{pd}}(D|D)|D>0]$ and $\mathbb{E}[ATT(D|D)|D>0]/\mathbb{E}[D|D>0]$.

Lastly, these papers target different causal parameters, put more emphasis on different DiD designs, and, therefore, should be viewed as complements rather than substitutes. In our view, DiD with continuous treatment is another area in which more methodological research is warranted. See Callaway et al. (2021) and de Chaisemartin et al. (2024a) for a more thorough discussion of many other cases.

## A.3 Triple differences

The causal interpretation of DiD estimates depends on the plausibility of their identification assumptions, which involve a no-anticipation and a parallel trends condition. In some applications, however, these assumptions may not hold—for example, when the trends of average untreated outcomes among men and women vary across treatment groups. In these cases, a common empirical practice is to attempt to model these violations of parallel trends directly or to conduct sensitivity analysis (Freyaldenhoven, Hansen, Pérez-Pérez and Shapiro, 2024; Rambachan and Roth, 2023). In some specific treatment designs in which treatment is rolled out to different units or groups (e.g., states), but is targeted to a specific subset (partition) of the population (e.g., women), it is possible to relax DiD-type parallel trends so that partition-specific and group-specific violations of parallel trends are allowed. Such setups are often referred to as "triple differences" (DDD). Since its introduction by Gruber (1994), DDD has become very popular among empirical researchers—see Olden and Møen (2022) for documentation. In this section, we provide a brief overview of the target parameters and identifying assumptions in DDD. We also highlight that, contrary to conventional wisdom, DDD procedures cannot generally be expressed as the difference between two DiD, especially when parallel trends assumptions only hold after conditioning on covariates or when treatment adoption is staggered. This discussion borrows heavily from Ortiz-Villavicencio and Sant'Anna (2025).

[3]Interestingly, they also show that commonly used TWFE regression specifications are too rigid to lead to easy-to-interpret causal parameters of interest. In fact, they show that one can provide several different decompositions of the TWFE treatment coefficient depending on the specific causal parameter being used as a building block for the analysis, though every decomposition considered by them has some issues related to negative-weighting, additional "bias" terms, or non-interpretable weights that can distort inference. They emphasize that all this can be easily resolved by adopting the forward-engineering approach.

We start our analysis by discussing potential outcomes and treatment design. As we focus on binary treatments (with potential staggered adoption), the potential outcome is the same as discussed in the main text, with $Y_{i,t}(g)$ denoting the potential outcome for unit $i$ in time $t$ if first exposed to treatment in period $g$. In DDD setups, a unit $i$ is *exposed* to treatment in period $t$ if (i) it belongs to a group (e.g., state) that enabled treatment in period $g$ and $t$ is a post-treatment period, $t \geqslant g$, and (ii) it belongs to the subset of the population that qualifies (or is *eligible*) for treatment (e.g., women). Let $S \in \mathcal{S} \subseteq \{2, ..., T\} \cup \{\infty\}$ denote the time each group (e.g., state) enables the policy/treatment, with the notion that $S = \infty$ if the policy is not enabled in the observed time frame. We also denote the partition of the population that (eventually) qualifies for the treatment by $Q$ with $Q_i = 1$ if unit $i$ is (eventually) eligible for treatment and $Q_i = 0$ otherwise. With these notations, we can define the treatment groups $G_i$ according to the first time a unit $i$ is *exposed* to treatment; that is, $G_i = S_i$ if $Q_i = 1$ and $G_i = \infty$ if $Q_i = 0$.[4]

Similar to standard DiD designs, DDD is interested in the $ATT(g,t)$-type parameters discussed in Section 3.1. Given the particular structure of the DDD problem, we can write $ATT(g,t)$'s as

$$ATT(g,t) \equiv \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|G_i = g] = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|S_i = g, Q_i = 1],$$

to stress that it measures the average treatment effect at time period $t$ of first being exposed to treatment in period $g$ versus not being exposed to treatment, among units that are actually exposed to treatment in period $g$, i.e., units that are in groups that the policy was first enabled in period $g$ and that qualify for treatment. One can also analyze aggregations of these $ATT(g,t)$ parameters to form causal summary parameters that can be more precisely estimated and highlight treatment effect heterogeneity in some specific directions. This would follow the exact same steps as we discussed in Section 5, once again highlighting the importance of our forward-engineering approach.

Identifying these causal parameters involves a no-anticipation assumption and a (conditional) parallel trends assumption. Assumption Assumption 4 can be recycled here, as DDD has the same empirical content as DiD when it comes to no-anticipation. The parallel trends assumption, though, needs to be adjusted as an empirical appeal of DDD is that it can identify ATT parameters even when Assumption Assumption 3 or the other PT variations discussed in Section 4 do not hold. Here, we consider a variation of Assumption Assumption 3 that holds only after conditioning on covariates and allows for some partition-specific and group-specific non-parallel trends.

**Assumption DDD-PT-GT-all** (DDD-Parallel Trends for every period and group)**.** For every group $s$ and $s'$ and time periods $t$, with probability one,

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s, Q = 1, X\right] \quad - \quad \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s, Q = 0, X\right]$$
$$=$$

[4]Note that when all units are eligible for treatment, we have $G_i = S_i$, which gets us back to a (staggered) DiD setup.

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s', Q = 1, X\right] \quad - \quad \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s', Q = 0, X\right].$$

When there are only two periods, $t = 1, 2$, and two groups, $S \in \{2, \infty\}$, and covariates play no role in terms of identification—that is, Assumption DDD-PT-GT-all holds without $X$ (or, equivalently, with $X = 1$ for all units)—Olden and Møen (2022) show that one can identify $ATT(2, 2)$ as the difference of two DiD estimands:

$$\begin{aligned} ATT(2,2) \quad = \quad & \mathbb{E}\left[Y_{t=1} - Y_{t=1}|S = 2, Q = 1\right] - \mathbb{E}\left[[Y_{t=1} - Y_{t=1}|S = 2, Q = 0\right] \\ & - \big(\mathbb{E}\left[Y_{t=1} - Y_{t=1}|S = \infty, Q = 1\right] - \mathbb{E}\left[[Y_{t=1} - Y_{t=1}|S = \infty, Q = 0\right]\big). \end{aligned}$$

Estimation and inference would be straightforward, as one could use the analogy principle or a two-way fixed effects regression with triple interactions—see Olden and Møen (2022) for details.

Ortiz-Villavicencio and Sant'Anna (2025) show that DDD estimands cannot be written as the difference of two DiD estimands when covariates are important for identification, or when treatment adoption is staggered over time and one wants to use not-yet-treated units as a comparison group (as is commonly done in DiD setups). They show how ignoring these considerations and proceeding as if DDD were indeed just a difference of two DiDs can lead to severely biased estimates for the $ATT(g, t)$'s. Ortiz-Villavicencio and Sant'Anna (2025) also show how one can avoid these issues by adopting a forward-engineering approach to the DDD problem. They propose regression-adjusted, inverse probability weighting, and doubly robust estimators for DDD setups that can reliably recover $ATT(g, t)$ and their associated summary parameters under mild assumptions. The paper discusses using multiple comparison groups to generate more precise estimates than simply using a single comparison group. Relatedly, Strezhnev (2023) discusses several limitations of common two-way fixed effects regression specifications commonly used for DDD analysis.

Sometimes researchers use the term "triple differences" to mean different things and often use different identification assumptions to estimate these different quantities. Caron (2025) discusses using a triple difference strategy to estimate treatment effect heterogeneity. We recommend that practitioners be transparent about target parameters, research designs, and identification assumptions to allow the research community to understand the goals and the differences between DDD procedures.

## A.4 Distributional DiD procedures

Our paper focuses on learning about *average* treatment effects in various DiD setups. However, approaches that embrace heterogeneity can also target quantities that describe heterogeneity other than average treatment effect parameters. In some settings, researchers may want more information about the distributional impacts of treatment participation. For instance, if a policymaker faces two different labor market programs with very similar average effects on earnings, they may prefer the one that potentially has a higher impact on the lower tail of the income distribution. Difference-in-Differences-type strategies can also be used to identify, estimate, and make inferences about

various distributional features of the outcome of interest. This area has received a substantial amount of methodological consideration by econometricians in recent years; see Athey and Imbens (2006), Bonhomme and Sauder (2011), Callaway, Li and Oka (2018), Callaway and Li (2019), Roth and Sant'Anna (2023), Ghanem, Kédagni and Mourifié (2023), Fernández-Val, Meier, van Vuuren and Vella (2024), and references therein. For some empirical literature using distributional DiD procedures, see Meyer, Viscusi and Durbin (1995), Finkelstein and McKnight (2008), and Cengiz, Dube, Lindner and Zipperer (2019), among many others.

An analysis of distributional quantities does not require different potential outcomes notation from Section 4; it just targets functionals of the potential outcome distributions other than their means. The first thing to notice is that there are several types of distributional causal parameters in the treated group that one may care about. The unique feature of them is that they are all functionals of $F_{Y_t(g)|G=g}(y) = \mathbb{P}(Y_t(g) \leqslant y|G = g)$ and $F_{Y_t(\infty)|G=g}(y) \equiv \mathbb{P}(Y_t(\infty) \leqslant y|G = g)$. Examples of such functionals include distributional treatment effects in time period $t$ among units first treated in period $g$ (denominated in probability units),

$$DTT(y|g,t) = F_{Y_t(g)|G=g}(y) - F_{Y_t(\infty)|G=g}(y),$$

quantile treatment effects in time period $t$ among units first treated in period $g$ (denominated in outcome units),

$$QTT(\tau|g,t) = F^{-1}_{Y_t(g)|G=g}(\tau) - F^{-1}_{Y_t(\infty)|G=g}(\tau),$$

where $F^{-1}_{Y_t(g)|G=g}(\tau) = inf\{y : F_{Y_t(g)|G=g}(y) \geqslant \tau\}$ denotes the $\tau$-quantile of $Y_t(g)$ among units in group $G = g$, and $F^{-1}_{Y_t(\infty)|G=g}(\tau)$ is defined analogously. Other functionals related to inequality measures can also be obtained; see Firpo and Pinto (2016) for a discussion on this topic.

To make inferences about these different causal parameters, one needs to identify $F_{Y_t(\infty)|G=g}(y)$ and $F_{Y_t(g)|G=g}(y)$. Identification of $F_{Y_t(g)|G=g}(y)$ is usually non-controversial, as we can use data from units in group $G = g$ to learn about the distribution of $Y_t(g)$. The main challenge is related to how to learn the counterfactual distribution $F_{Y_t(\infty)|G=g}(y)$ from the data. This is where different DiD-type procedures differ, as each paper in this literature relies on different and often non-nested identification assumptions that, if true, identify $F_{Y_t(\infty)|G=g}(y)$. Given the space constraints, we do not provide explicit and detailed discussion about how these different DiD-related distributional procedures function. However, all distributional DiD estimators share our forward-engineering approach; they clearly state their identification assumptions and target parameters and then provide estimators that recover well-defined causal quantities. We also note that most distributional DiD methodological papers focus on two-period and two-group setups. However, it is straightforward to build similar arguments to those in Section 4 to extend the designs to more general settings, which is again another benefit of the forward-engineering approach to causal inference.

We close this section by noting that there exist other types of distributional parameters of interest related to the distribution *of* the treatment effects in period $t$ among the units in group $g$,

$\mathbb{P}(Y_t(g) - Y_t(\infty) \leqslant y | G = g)$. In general, such causal quantities cannot be point identified, as discussed in Heckman, Smith and Clements (1997), Fan and Yu (2012) and Callaway (2021). However, often one can still partially identify such policy-relevant parameters under different restrictions. We refer the reader to Callaway (2021) for a more detailed discussion of this topic.

## A.5 Repeated cross-sections and unbalanced panel data

An appealing feature of DiD procedures is that, although helpful, a balanced panel is not a requirement for DiD analyses, which can also be deployed with repeated cross-sectional data or unbalanced panels. Indeed, as discussed in Section 3 and made explicit in equation (4), the $2 \times 2$ building block in unconditional DiD analyses involves only averages that are group and time specific, and does not require the same unit to be observed in all periods. As discussed in Callaway and Sant'Anna (2021), the same applies to unconditional staggered adoption setups, and one need not enforce a balanced panel even within each subset of the data used to estimate the $ATT(g,t)$ building blocks. One caveat is that the interpretation of the parameter of interest may change, which we discuss more below.

When covariates are available and play an important role in the plausibility of the identification assumptions, the differences between DiD with a balanced panel and repeated cross-sections (or unbalanced panel) are subtle, can be practically important, and are often not discussed in methodological papers. The gist of the problem relates to potential compositional changes over time. Most DiD papers that rigorously discuss repeated cross-section setups, including Abadie (2005), Sant'Anna and Zhao (2020), and Callaway and Sant'Anna (2021), rule out compositional changes by assuming that the joint distribution of covariates and treatment groups is invariant over time, a stationarity-type assumption. However, this may not be warranted in empirical applications, and erroneously imposing this additional assumption can lead to biases (Hong, 2013; Sant'Anna and Xu, 2023). On the other hand, when this stationarity assumption is justified and correctly used, the gains in power when conducting inference for DiD parameters can be noticeable (Sant'Anna and Xu, 2023). In what follows, we use the $2 \times 2$ setup to explain how compositional changes can complicate the analysis and why ruling it out leads to a gain in precision.

To see how issues related to compositional changes affect the analysis, let us first assume that there are no compositional changes and that the stationarity assumption is valid. In this case, the average treatment effect on the treated in period two (post-treatment) can be written as

$$\begin{aligned} ATT(2) &\equiv \mathbb{E}[Y_{i,t=2}(1)|D_i = 1] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1] \\ &= \mathbb{E}[Y_{i,t=2}(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1, T_{i,t=2} = 1] \\ &= \mathbb{E}[Y_i(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_i(0)|D_i = 1, T_{i,t=2} = 1], \end{aligned} \tag{A.1}$$

where $T_{i,t}$ is an indicator if unit $i$ is observed in period $t$, $Y_i(d) = T_{i,t=2}\ Y_{i,t=2}(d) + T_{i,t=1}\ Y_{i,t=1}(d)$ is the potential outcome for unit $i$, $D_i = 1\{G_i = 2\}$ is a treatment group dummy that equals one

if a unit is first treated in period two and zero if it is untreated in both periods. We also set $X_i$ to be a vector of (pre-treatment) covariates. Note that even here, we already use the stationarity condition that the joint distribution of $(D_i, X_i)$ is invariant to $T_{i,t=2}$ to move from the first to the second line and establish A.1.

To identify $ATT(2)$ it is often constructive to first establish the identification of its conditional-on-covariates analog; that is, the conditional ATT in period two among units with covariates $X_i$, $ATT_{X_i}(2)$. This is exactly how we proceeded in Section 3.2. Under the stationarity condition, and similarly to A.1, we can express this quantity as[5]

$$\begin{aligned} ATT_{X_i}(2) &\equiv \mathbb{E}[Y_{i,t=2}(1)|D_i = 1, X_i] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1, X_i] \\ &= \mathbb{E}[Y_i(1)|D_i = 1, X_i, T_{i,t=2} = 1] - \mathbb{E}[Y_i(0)|D_i = 1, X_i, T_{i,t=2} = 1]. \end{aligned} \tag{A.2}$$

Next, we have to establish the identification of this quantity. As expected, we will again use conditional parallel trends, no-anticipation, and overlap assumptions. The no-anticipation condition used here is the same as the one in the main text. The conditional parallel trends and overlap assumptions need to be modified, as we now work with multiple partitions of the data depending on treatment status and the period a unit is observed. In this sense, we modify Assumptions Assumption 2 and Assumption 1 to the following related, but different, assumptions. These modifications are warranted regardless of whether compositional changes are present; this step is instead tied to data structure.[6]

**Assumption CPT-RCS** ($2 \times 2$ Conditional Parallel Trends with repeated cross-sections)**.** We assume that, with probability one,

$$\begin{aligned} \mathbb{E}[Y_{i,t=2}(0)|X_i, D_i = 1, T_{i,t=2} = 1] \quad &- \quad \mathbb{E}_\omega[Y_{i,t=1}(0)|X_i, D_i = 1, T_{i,t=1} = 1] \\ &= \\ \mathbb{E}[Y_{i,t=2}(0)|X_i, D_i = 0, T_{i,t=2} = 1] \quad &- \quad \mathbb{E}[Y_{i,t=1}(0)|X_i, D_i = 0, T_{i,t=1} = 1]. \end{aligned} \tag{A.3}$$

**Assumption SO-RCS** (Strong overlap with repeated cross-sections)**.** For some $\epsilon > 0$ and every $(d, s) \in \{0, 1\} \times \{0, 1\}$, $\epsilon < P[D_i = d, T_{i,t=2} = s|X_i] < 1 - \epsilon$.

With this modification, we can now show that when Assumptions Assumption 4, CPT-RCS, and Assumption 1 hold, the conditional ATT parameter $ATT_{X_i}(2)$ is identified by[7]

$$\begin{aligned} ATT_{X_i}(2) =&(\mathbb{E}[Y_i|D_i = 1, T_{i,t=2} = 1, X_i] - \mathbb{E}[Y_i|D_i = 1, T_{i,t=1} = 1, X_i]) \\ &- (\mathbb{E}[Y_i|D_i = 0, T_{i,t=2} = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, T_{i,t=1} = 1, X_i]). \end{aligned} \tag{A.4}$$

[5]To guarantee that all the conditional expectations in A.2 are well-defined, we need an overlap condition that guarantees that $P(T_{i,t=2} = 1, D_i = 1|X_i) > 0$. We discuss this below.

[6]In setups where we rule out compositional changes and impose the stationarity condition that the joint distribution of $(D_i, X_i)$ is invariant to $T_{i,t=2}$, we may not need to modify Assumption Assumption 2. We do it here for transparency purposes.

[7]We also require the assumption that the pooled repeated cross-section data $\{Y_i, D_i, X_i, T_{i,t=2}, T_{i,t=1}\}_{i=1}^n$ is *iid*, though this is fairly standard and uncontroversial; see, for instance, Abadie (2005) and Sant'Anna and Zhao (2020, Assumption 1). We maintain this condition as an assumption throughout this section.

This step provides a methodological justification to estimate the $ATT_{X_i}(2)$'s using four conditional expectations that use only the available data. In addition, it highlights that, under the stationarity assumption, once we learn the $ATT_{X_i}(2)$'s, we can aggregate them using the covariate distribution of treated units *available from both time periods* to get the $ATT(2)$. More formally, $ATT(2)$ is identified by

$$\begin{aligned} ATT(2) &= \mathbb{E}[ATT_{X_i}(2)|D_i = 1] \\ &= \mathbb{E}[ATT_{X_i}(2)|D_i = 1, T_{i,t=2} = 1]P(T_{i,t=2} = 1|D_i = 1) \\ &\quad + \mathbb{E}[ATT_{X_i}(2)|D_i = 1, T_{i,t=1} = 1]P(T_{i,t=1} = 1|D_i = 1), \end{aligned}$$

where $ATT_{X_i}(2)$ is given by A.4. This is the second point at which the stationarity assumption and the absence of compositional changes are necessary: under these conditions, covariates from treated units across the entire dataset can be used to identify $ATT(2)$. The fact that you can pool data across all periods to learn about $ATT(2)$ translates to gains in power, as formally discussed by Sant'Anna and Zhao (2020) and Sant'Anna and Xu (2023). The third place where the stationarity condition affects the analysis is in the characterization of how "the most precise" (regular and asymptotically linear) estimator for the $ATT(2)$ should look. This point relates to the semi-parametric efficiency bound and the construction of efficient (and doubly robust) estimators. As these points are slightly more technical, we refer the readers to Sant'Anna and Zhao (2020) and Sant'Anna and Xu (2023) for more details.

Overall, when group composition does not change over time, one can pool information across the entire dataset, which has an impact on the definition of target parameters and leads to more precise inference procedures. But what happens when this condition fails? How does this affect the analysis?

First, this matters for the definition of the treatment effect of interest. In setups where the sampling varies across periods, we do not have a single notion of $ATT(2)$. Instead, we need to accommodate the fact that the $ATT(2)$ may vary across units sampled from different periods. Thus, when we do not rule out compositional changes, we must be explicit about the treated subpopulation that we are interested in. It is common to focus on the average treatment effect in period two among treated units *that are also sampled in period two*, that is,[8]

$$\begin{aligned} ATT(2|T_{t=2} = 1) &\equiv \mathbb{E}[Y_{i,t=2}(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1, T_{i,t=2} = 1] \\ &= \mathbb{E}[Y_i(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_i(0)|D_i = 1, T_{i,t=2} = 1]. \end{aligned} \tag{A.5}$$

Although A.5 has the same statistical estimand as A.1—that is, the formulas on the right-hand side of the equation coincide—it has a very different interpretation. It is the $ATT(2)$ among units sampled in period 2, and is not an "overall" $ATT(2)$. One may think this difference is

[8] One may also be interested in the ATT in period two among treated units that are sampled in period one. The arguments required to establish (point) identification of this parameter differ from those we use here. A main challenge is that we do not observe $\mathbb{E}[Y_{i,t=2}(1)|D_i = 1, T_{i,t=1} = 1]$, and the parallel trends assumption we leverage does not involve treated potential outcomes.

merely cosmetic, but as discussed below, this has implications for constructing estimands when covariates are important for identification. Where covariates do not play an important role, it is simply a matter of changing the interpretation of your reported estimates (which also applies to unconditional staggered setups, to be clear).

When covariates do play an important role, however, and when Assumptions CPT-RCS, Assumption 4 and SO-RCS hold, the conditional ATT parameter among units sampled in period two, $ATT_{X_i}(2|T_{t=2})$ is identified by

$$\begin{aligned} ATT_{X_i}(2|T_{t=2}) &= (\mathbb{E}[Y_i|D_i=1,T_{i,t=2}=1,X_i]-\mathbb{E}[Y_i|D_i=1,T_{i,t=1}=1,X_i]) \\ &\quad -(\mathbb{E}[Y_i|D_i=0,T_{i,t=2}=1,X_i]-\mathbb{E}[Y_i|D_i=0,T_{i,t=1}=1,X_i]), \end{aligned} \tag{A.6}$$

which, in turn, implies that $ATT(2|T_{t=2}=1)$ is identified by

$$ATT(2|T_{t=2}=1)=\mathbb{E}[ATT_{X_i}(2|T_{t=2})|D_i=1,T_{i,t=2}=1]. \tag{A.7}$$

Several remarks are worth making. First, the statistical estimand in A.6 is the same as when one rules out compositional changes as in A.4, suggesting that, once again, what changes in this step is the interpretation. However, these interpretative issues have direct consequences on the appropriate method for aggregating across covariate values. As clearly stated in A.7, in the presence of potential compositional changes, one is not allowed to pool information across periods to identify (and also estimate and make inference) about $ATT(2|T_{t=2}=1)$. As discussed in Sant'Anna and Xu (2023), ignoring these issues and pooling data from all periods in the presence of compositional changes leads to a bias that is important to be aware of. We refer the reader to Sant'Anna and Xu (2023) for a discussion related to unbalanced panels and also on a discussion about doubly robust and semiparametric efficient DiD estimators under compositional changes. We are not aware of any papers that formally extend the discussion in Sant'Anna and Xu (2023) to staggered DiD designs. Still, this extension is surely possible by following our forward-engineering approach to DiD.

We close this section by highlighting that, in practice, it is possible to test for compositional changes by comparing the estimates from estimators that impose it and those that do not. Sant'Anna and Xu (2023) discuss Hausman-type tests in the two-period setting, though one can extend those to more general setups. We also highlight that when it comes to DiD setups with staggered adoption, some equivalence results discussed in Section 4 no longer hold with repeated cross-sections or unbalanced panel data. For instance, the Sun and Abraham (2021) regression-based strategy to estimate $ATT(g,t)$'s using (5.11) no longer coincides with Callaway and Sant'Anna's (2021) estimators using the analog of (5.3):

$$\widehat{ATT}_{\text{never}}(g,t)=(\overline{Y}_{G=g,t}-\overline{Y}_{G=g,t=g-1})-(\overline{Y}_{G=\infty,t}-\overline{Y}_{G=\infty,t=g-1}),$$

where $\overline{Y}_{G=a,t=s}$ is the sample mean of $Y$ among units that belong in group $G=a$ and are observed in period $t=s$. In fact, it is unclear exactly what estimand is being recovered when one uses

(5.11) with an unbalanced panel. If one replaces unit fixed effects with treatment group dummies in (5.11), such equivalence is restored, though we suspect that many practitioners do not use this alternative specification. In general, we caution against extrapolating from a well-motivated regression specification that was studied under one specific setup to another related but inherited different framework. This practice has led to many issues in DiD, which can be fully avoided by adopting the forward-engineering approach discussed in this paper.

# References


**Abadie, Alberto**, "Semiparametric Difference-in-Difference Estimators," *The Review of Economic Studies*, 2005, *72*, 1–19.

**Angrist, Joshua D. and Guido W. Imbens**, "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 1995, *90* (430), 431–442.

**Athey, Susan and Guido W. Imbens**, "Identification and Inference in Nonlinear Difference in Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

**Bonhomme, Stéphane and Ulrich Sauder**, "Recovering Distributions in Difference-in-Differences Models: a Comparison of Selective and Comprehensive Schooling," *Review of Economics and Statistics*, 2011, *93* (May), 479–494.

**Callaway, Brantly**, "Bounds on distributional treatment effect parameters using panel data with an application on job displacement," *Journal of Econometrics*, 2021, *222* (2), 861–881.

_ **and Pedro HC Sant'Anna**, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

_ **and Tong Li**, "Quantile Treatment Effects in Difference in Differences Models with Panel Data," *Quantitative Economics*, 2019, *10* (4), 1579–1618.

_ **, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna**, "Difference-in-Differences with a Continuous Treatment," *arXiv:2107.02637 [econ]*, 2021.

_ **,** _ **, and** _ , "Event Studies with a Continuous Treatment," *AEA Papers and Proceedings*, May 2024, *114*, 601–605.

_ **, Tong Li, and Tatsushi Oka**, "Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with Only Two Time Periods," *Journal of Econometrics*, 2018, *206* (2), 395–413.

**Caron, Laura**, "Triple Difference Designs with Heterogeneous Treatment Effects," *arXiv:2502.19620*, 2025.

**Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer**, "The Effect of Minimum Wages on Low-Wage Jobs," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1405–1454.

**de Chaisemartin, Clément and Xavier D'Haultfoeuille**, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

_ **and** _ , "Difference-in-Differences Estimators of Intertemporal Treatment Effects," 2023. Working Paper.

_ **and** _ , "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey," *Econometrics Journal*, 2023, *Forthcoming.*

_ **,** _ **, Félix Pasquier, and Gonzalo Vazquez-Bare**, "Difference-in-Differences for Continuous Treatments and Instruments with Stayers," *arXiv:2201.06898*, 2024.

_ **, Xavier D'Haultfuille, and Gonzalo Vazquez-Bare**, "Difference-in-Difference Estimators with Continuous Treatments and No Stayers," *AEA Papers and Proceedings*, May 2024, *114*, 610–613.

**Fan, Yanqin and Zhentao Yu**, "Partial Identification of Distributional and Quantile Treatment Effects in Difference-in-Differences Models," *Economics Letters*, 2012, *115* (3), 511–515.

**Fernández-Val, Iván, Jonas Meier, Aico van Vuuren, and Francis Vella**, "Distribution Regression Difference-In-Differences," *arXiv:2409.00123*, 2024.

**Finkelstein, Amy and Robin McKnight**, "What Did Medicare Do? The Initial Impact of Medicare on Mortality and Out-of-Pocket Medical Spending," *Journal of Public Economics*, 2008, *92* (7), 1644–1668.

**Firpo, Sergio and Cristine Pinto**, "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures," *Journal of Applied Econometrics*, 2016, *31* (3), 457–486.

**Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez-Pérez, and Jesse M. Shapiro**, "Visualization, identification, and estimation in the linear panel event-study design," in "Advances in Economics and Econometrics: Twelfth World Congress" 2024. Forthcoming.

**Ghanem, Dalia, Désiré Kédagni, and Ismael Mourifié**, "Evaluating the Impact of Regulatory Policies on Social Welfare in Difference-in-Difference Settings," *arXiv:2306.04494*, 2023.

**Gruber, Jonathan**, "The Incidence of Mandated Maternity Benefits," *American Economic Review*, June 1994, *84* (3), 622–641.

**Heckman, James J., Jeffrey Smith, and Nancy Clements**, "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 1997, *64* (4), 487–535.

**Hong, Seung-Hyun**, "Measuring the effect of Napster on recorded music sales: difference-in-differences estimates under compositional changes," *Journal of Applied Econometrics*, 2013, *28* (2), 297–324.

**Imai, Kosuke, In Song Kim, and Erik H. Wang**, "Matching Methods for Causal Inference with Time-Series Cross-Sectional Data," *American Journal of Political Science*, 2023, *67* (3), 587–605.

**Liu, Licheng, Ye Wang, and Yiqing Xu**, "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data," *American Journal of Political Science*, 2024, *68* (1), 160–176.

**Meyer, Bruce, W. Kip Viscusi, and David Durbin**, "Workers Compensation and Injury Duration: Evidence from a Natural Experiment," *The American Economic Review*, 1995, *85* (3), 322–340.

**Olden, Andreas and Jarle Møen**, "The triple difference estimator," *The Econometrics Journal*, 2022, *25* (3), 531–553.

**Ortiz-Villavicencio, Marcelo and Pedro H. C. Sant'Anna**, "Better Understanding Triple Differences Estimators," 2025. Working paper.

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

**Robins, James**, "A New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 1986, *7*, 1393–1512.

**Roth, Jonathan and Pedro H. C. Sant'Anna**, "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica*, 2023, *91* (2), 737–747.

**Sant'Anna, Pedro H. C. and Jun Zhao**, "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics*, 2020, *Forthcoming.*

_ **and Qi Xu**, "Difference-in-Differences with Compositional Changes," *arXiv:2304.14256*, 2023.

**Strezhnev, Anton**, "Decomposing Triple-Differences Regression under Staggered Adoption," *arXiv:2307.02735*, 2023.

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