The role of parallel trends in event study settings:
An application to environmental economics[*]

Michelle Marcus   Pedro H. C. Sant'Anna

Vanderbilt University   Vanderbilt University

September 3, 2020

Difference-in-Differences (DID) research designs usually rely on variation of treatment timing such that, after making an appropriate parallel trends assumption, one can identify, estimate, and make inference about causal effects. In practice, however, different DID procedures rely on different parallel trends assumptions (PTA), and recover different causal parameters. In this paper, we focus on staggered DID (also referred as event-studies) and discuss the role played by the PTA in terms of identification and estimation of causal parameters. We document a "robustness" vs. "efficiency" trade-off in terms of the strength of the underlying PTA, and argue that practitioners should be explicit about these trade-offs whenever using DID procedures. We propose new DID estimators that reflect these trade-offs and derived their large sample properties. We illustrate the practical relevance of these results by assessing whether the transition from federal to state management of the Clean Water Act affects compliance rates.

1 Introduction

Researchers and policy makers are often interested in evaluating the causal effect of a given treatment/intervention on an outcome of interest. When data from randomized control trials related to the causal question of interest are not available, researchers often rely on "natural experiments" and make use of difference-in-differences (DID) methods to estimate the effect of a given policy. The canonical DID method presumes the existence of two groups, the treated and the comparison group, two time periods, pre-treatment and post-treatment periods, such that the comparison group is not treated in either time period, and the treated group is only treated at the post-treatment period. Then, one estimates the average treatment effect among those treated units by comparing the average difference in pre and post-treatment outcomes of two groups, or, equivalently, by using a two-way fixed effects regression model with a group and a time fixed effect; see, e.g., Section 2 of Lechner (2010) for details about the history of DID procedures.

It is worth stressing that the causal interpretation of the two groups, two time periods (henceforth 2×2) DID procedure relies on a so-called parallel trend assumption (PTA): in the absence of the treatment, the average outcome for the treated and comparison groups would have evolved in parallel. Such an assumption is well-understood, see e.g. Chapter 5 of Angrist and Pischke (2009), Chapter 10 of Cunningham (2018), and Section 2 of Sant'Anna and Zhao (2020). Importantly, it restricts the average counterfactual outcome for the treated units *at the post-treatment period* had they not been subject to the treatment, but it does not directly impose restrictions on the outcome in pre-treatment periods. In addition, it is worth mentioning that the PTA is untestable in this 2×2 setup, see, e.g., Chapter 10 of Cunningham (2018) and Section 4 of Callaway and Sant'Anna (2020).

Although most of the aforementioned points are well-understood in the 2×2 setup, in many DID applications, however, there are more than two time periods, and units can be treated in different points in time, which leads to multiple treatment groups as well. This many periods, many groups DID setup is substantially more challenging than the canonical 2×2 one. For instance, Sun and Abraham (2020) (henceforth S&A), Callaway and Sant'Anna (2020) (henceforth C&S), de Chaisemartin and D'Haultfœuille (2020) (henceforth dC&D), and Goodman-Bacon (2019), study DID procedures with multiple periods and multiple groups, and each of these papers rely on *different* types of parallel trends assumptions and/or propose *different* estimators for *different* causal parameters of interest. This is in sharp contrast with the 2×2 setup, where there is only one type of PTA[1] and the

---

[1]In the 2×2 DID setup, the only variation in PTA one observes is whether it holds unconditionally, or only after conditioning on a vector of observed characteristics, see, e.g., Heckman et al. (1998), Abadie (2005), and Sant'Anna and Zhao (2020). This is not the type of variation of the PTA we are referring to.

main parameter of interest is the average treatment effect among the treated units. It is also worth stressing that in this more complex DID setup treatment effect estimates based on two-way fixed effects (TWFE) regression models can only be interpreted as weighted averages of causal effects, and, perhaps even more problematic, that some of these weights can be negative, see, e.g., S&A, Borusyak and Jaravel (2017), dC&D, and Goodman-Bacon (2019); see also Laporte and Windmeijer (2005), Wooldridge (2005), Chernozhukov et al. (2013), and Gibbons et al. (2018) for earlier related results based on (one-way) fixed-effect estimators. As so, policy evaluations based on TWFE linear regression models can be misleading, especially when treatment effects are dynamic.

In this paper, we attempt to clarify the role played by different parallel trends assumptions in DID setups with many time periods and many treatment groups. Towards this goal, we revisit S&A, C&S, and dC&D, and discuss their main differences in terms of parallel trends assumptions and parameters of interest — we focus on these three papers because their primary concern is how one can identify causal parameters of interest without imposing restrictions beyond a PTA[2]. We exclusively focus on DID settings with staggered adoption designs and binary treatments. By doing so, we can compare the PTA and parameters of interest discussed in S&A, C&S, and dC&D in a more direct manner.

We show that the PTA invoked by S&A and dC&D (i) not only restricts counterfactual trends after the treatment, but also imposes parallel pre-treatment trends, and (ii) imposes that every *individual* group that is not-yet treated by time t can be used as a valid comparison group for those earlier-treated units, at time t. C&S, on the other hand, considers two different PTAs, one that relies on using "never-treated" units as a comparison group, and one that uses not-yet-treated units as valid comparison groups for the earlier-treated units. Interestingly, both PTAs considered by C&S are, at least technically speaking, weaker than the PTA invoked by S&A and dC&D, as they either do not restrict pre-treatment trends, or, when they do, these restrictions are potentially less demanding. Although these PTAs differ in their "strength", we show that they all can be used to recover the same variety of average treatment effects measures.

Overall, we argue that, in practice, one should be explicit about the type of PTA invoked in the DID analysis. On top of adding transparency and objectivity to the analysis, see, e.g., Rubin (2007, 2008), we stress that the choice of the parallel trends assumption can also help in selecting the most appropriate estimator for a given parameter of interest. For instance, in situations where one is comfortable with a "stronger" PTA, we show that one can exploit

---

[2]See also Athey and Imbens (2018), Goodman-Bacon (2019), Arkhangelsky et al. (2018), Borusyak and Jaravel (2017), Ferman and Pinto (2019), and Rambachan and Roth (2019) for other recent contributions to the DID literature.

the fact that such PTAs lead to *overidentification*, and then use the generalized method of moments (GMM) framework to form more efficient treatment effect estimators than those currently available in the literature; see Proposition 4.1. Another consequence of adopting the GMM framework is that it is relatively straightforward to test for the credibility of a "stronger" parallel trends assumption by conducting a classical Hansen-Sargan J-test. To the best of our knowledge, this paper is the first to make this simple, but important observation.

In many other situations, however, we expect that researchers will not be a priori comfortable with a "stronger" version of the PTA, as it may impose more restrictions on the data than those strictly required for identification of treatment effect parameters. Indeed, when the number of groups and time periods is moderate, the number of restrictions implied by the "stronger" PTA can be close to the number of observations available in the data. In such cases, it may be reasonable to favor "weaker" versions of the PTA. When a sufficiently large "never-treated" group is available, researchers can use the easy-to-implement nonparametric DID estimators based on sample means proposed by C&S. When an appropriate "never-treated" group is unavailable, we show that one can rely on an alternative "weaker" PTA and use a simple plug-in DID estimator that differs from the ones considered by S&A, C&S, and dC&D. We show that this new DID estimator is consistent and asymptotically normal, and also describe a bootstrap procedure to conduct inference that is robust against multiple-testing problems. Interestingly, this new DID estimator does not rely on restricting pre-treatment trends, and, at the same time, exploits data from all available groups in the given application. On the other hand, both this newly proposed DID estimator and the one proposed by C&S are, in general, less efficient than the GMM estimator, which relies on more stringent assumptions. To the best of our knowledge, we are the first to document this "robustness" versus "efficiency" trade-off in terms of the strength of the underlying PTA invoked in DID setups.

We illustrate the practical relevance of the aforementioned observations by revisiting Grooms (2015). We examine the effect of the transition from federal to state management of the Clean Water Act (CWA) on violation rates. Similarly to Grooms (2015), we find that the transition from federal to state control has little to no effect on violation rates — this result is robust across different parallel trends assumptions and different causal parameters of interest.

Next, like Grooms (2015), we also analyze whether states with a long prevalence of corruption see a large decrease in the violation rate after authorization relative to states without corruption. Grooms (2015) uses a dynamic TWFE (event-study) linear regression model, and finds strong evidence that violation rates decreased more in more corrupt states than in less corrupt states after the transition to state control. However, given that Grooms

(2015) focuses exclusively on TWFE-type estimators, it is not clear what kind of PTA is actually being made in the analysis. Here, we show how it can be beneficial to separate the analysis into two steps: (i) identification and the relevance of the PTA, and (ii) data analysis and estimation procedures. By proceeding in this manner, we find that the conclusion that violation rates dropped more in more corrupt states than less corrupt states depends on the type of PTA imposed. For instance, when one assumes that, in the absence of treatment, the counterfactual outcome trends differ depending on whether a state has a long prevalence of corruption or not ("corruption-specific trends"), we find essentially no evidence that the treatment effects vary depending on whether a state is more or less corrupt. On the other hand, if one assumes an alternative PTA such that one can use averages of both corrupt and non-corrupt states as valid comparison groups, we find evidence that more corrupt states see a larger decrease in the violation rate after authorization than less corrupt states, just like the original findings of Grooms (2015). As "corruption" is not randomly assigned, we believe that allowing for corruption-specific trends is the most natural identification set up in this context. These conflicting findings highlight the importance of explicitly stating the underlying PTA invoked in the exercise.

The rest of this paper is organized as follows. In Section 2, we present the general framework, compare the different PTAs using a stylized example, and describe the different parameters of interest considered by S&A, C&S, and dC&D. In Section 3, we discuss the testability of the PTAs and the practical considerations a researcher might take into account when choosing a PTA and a DID estimator. Section 4 describes how one can use generalized method of moments (GMM) framework to form more efficient treatment effect estimators when the chosen PTA leads to overidentification. Section 5 presents a new easy-to-compute DID estimator based on an alternative "weaker" PTA than what has been previously seen in the literature. Finally, Section 6 presents the empirical application, and we conclude in Section 7. Proofs and additional results are available at the Web Appendix, at https://pedrohcgs.github.io/files/Marcus_SantAnna_2020_webAppendix.pdf.

2   Difference-in-differences with multiple time periods

2.1   Framework

We first introduce the notation we use throughout the paper, which resembles that adopted by C&S. We consider the case with $\mathcal{T}$ periods and denote a particular time period by t where $t = 1, \ldots, \mathcal{T}$. In the canonical DID setup, $\mathcal{T} = 2$ and no one is treated in period 1. Let D_t be a binary variable equal to one if a unit is treated in period t and equal to zero otherwise.

Also, define G_g to be a dummy variable that is equal to one if a unit is first treated in period g, and define C as a dummy variable that is equal to one for units who are not treated in any period $\mathcal{T}$. For each unit, exactly one of the G_g or C is equal to one. Finally, let $Y_t(1)$ and $Y_t(0)$ be the potential outcomes at time t with and without treatment, respectively. The observed outcome in each period can be expressed as $Y_t = D_t Y_t(1) + (1 - D_t) Y_t(0)$.

Henceforth, we refer to "groups" as the group associated with the time a unit is first treated. Throughout the paper, we maintain the following assumptions.

Assumption 2.1 (Sampling). $\{Y_{i1}, Y_{i2}, \ldots Y_{i\mathcal{T}}, D_{i1}, D_{i2}, \ldots, D_{i\mathcal{T}}\}_{i=1}^{n}$ *is independent and identically distributed (iid).*

Assumption 2.2 (Staggered treatment design). *For $t = 2, \ldots, \mathcal{T}$,*
$$D_{t-1} = 1 \text{ implies that } D_t = 1$$

Assumption 2.3 (No Anticipation). *For all $t = 1, \ldots, \mathcal{T}$, $g = 2, \ldots, \mathcal{T}$ such that $t < g$,*
$$\mathbb{E}\left[Y_{it} \middle| G_g = 1\right] = \mathbb{E}\left[Y_{it}(0) \middle| G_g = 1\right]$$

Assumption 2.4 (Overlap). *$P(G_1 = 1) = 0$ and, for some $\epsilon > 0$, and all $g = 2, \ldots, \mathcal{T}$, $P(G_g = 1) > \epsilon$.*

Assumption 2.1 implies that we are considering the case of panel data. The discussions related to the case where only repeated cross-section data are available follows similar arguments and is omitted to avoid repetition. Assumption 2.1 does not restrict the temporal dependence across outcomes, though it relies on "large n, fixed t" panel data. Assumption 2.1 also rules out covariates; we only impose this simplification to allow for a more direct comparison between the proposals of S&A, C&S, and dC&D; we refer the reader to C&S for a detailed discussion about flexibly accommodating covariates into the DID analysis.

Assumption 2.2 imposes that treatment is "irreversible", i.e., once a unit is treated at time $t - 1$, it is "forever" treated. This assumption is usually referred to as staggered treatment adoption in the DID literature. We interpret this assumption as if units that experience treatment are forever affected by this experience, and do not "forget" about it.[3] We emphasize that, by imposing Assumption 2.2, we are able to directly compare the DID contributions of S&A, C&S, and dC&D.

Assumption 2.3 implies that there is no anticipatory response to treatment for those units that are eventually treated. This assumption is standard in the DID literature, though many

---

[3]When treatment can "turn on" and later "turn off", one usually needs to augment the potential outcome notation to analyze the effect of a given treatment path, see, e.g. Han (2019). When one is interested only in an average of the instantaneous effects of the policy among all units that switch treatments, one can bypass some of this complications by imposing a "no carryover assumption", see, e.g., dC&H.

times it is only implicitly imposed. Such an assumption can be violated if units foresee that they will be treated and change their behavior before the treatment takes place.[4]

Assumption 2.4 imposes that no unit is treated in the first time period, and that a new set of units are treated in each time period with a strictly positive probability. If there is an "always treated" group, i.e., units that are already treated in the first time period, we drop those observations because neither the data nor parallel trends assumptions for $Y_t(0)$ provide information information to identify the average treatment effect for this group.[5] We assume that new sets of units are treated in each time period only for notation convenience. Also, note that Assumption 2.4 accommodates, but does not require, that there is a "never treated" group available.

Next, we revisit S&A, C&S, and dC&D, paying particular attention to their PTA, underlying parameter of interest, and how one can estimate these parameters. When presenting these results, we impose Assumptions 2.1 - 2.4, which may result in slight changes of notation when compared to their original statements. In terms of notation, we follow C&S and, whenever possible, attempt to express the different parameters of interest in terms of functionals of "group-time average treatment effects", i.e., the average treatment effect at time t, for those units first treated at time g,[6]

$$
\begin{aligned}
ATT\left(g,t\right) & \equiv \; \mathbb{E}\left[Y_{it}\left(1\right)\middle| G_g = 1\right] - \mathbb{E}\left[Y_{it}\left(0\right)\middle| G_g = 1\right]. \qquad (2.1) \\
& = \; \alpha_{g,t}\left(1\right) - \alpha_{g,t}\left(0\right)
\end{aligned}
$$

To convey the discussion in an easy-to-understand manner, we consider a simple, stylized example. Assume that we observe Y_{it} for a sample of units $i = 1, \ldots, n$ in four time periods, $t = 1, 2, 3, 4$. Some units are first treated at time 3 ($G_{i3} = 1$), others at time 4, ($G_{i4} = 1$), and the remaining units are not treated in the entire observation window ($C_i = 1$). Once a unit i is treated at time g, it remains treated for all time periods $t \geq g$. Let $W = (Y_1, Y_2, Y_3, Y_4, G_3, G_4, C)'$, and assume that we observe a random sample $\{W_i\}_{i=1}^n$ of W.

---

[4]When the researcher is worried about anticipatory effects, this can be circumvented by simply redefining g to denote the period in which anticipatory effects begin. However, this may require strengthening other assumptions; see C&S for a discussion.

[5]In this paper, we attempt to only make parallel trends assumption about the evolution of $Y_t(0)$, and remain agnostic about the trends for $Y_t(1)$. When one is willing to impose parallel trends for $Y_t(1)$, too, then we can leverage the existence of an "always treated" group to form alternative parameters of interest, though such an assumption further restricts treatment effect heterogeneity. We leave a detailed discussion about this case for future research.

[6]S&A refers to $ATT\left(g,t\right)$ as cohort-specific average treatment effects on the treated, though they express it in terms of event-time $t - g$, i.e., the time elapsed since treatment started.

2.2 The different parallel trends assumptions

In this subsection, we present the three different parallel trends assumptions considered by S&A, C&S, and dC&D. We start by describing each PTA conceptually and then make use of the stylized example to highlight the key differences between these assumptions.

We first present the PTA assumption invoked by S&A and dC&D (see, Assumption 1 in S&A and Assumption 5 in dC&D).

Assumption 2.5 (Parallel trends assumption across all time periods and all groups). *For all $t = 2, \ldots, \mathcal{T}$, all $g = 2, \ldots, \mathcal{T}$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0)\right],$$

where the first equality holds only when there exists a "never-treated" group.

Assumption 2.5 states that, in the absence of treatment, the expectation of the outcome of interest follows the same path in all groups and in all time periods available in the data. Although fairly intuitive, such an assumption imposes important restrictions on the data (when combined with Assumptions 2.1 - 2.4). In particular, Assumption 2.5 imposes a parallel pre-trends condition across all treatment groups, and, as a consequence, allows one to use any *individual* group that has not yet been treated by time t (units with $G_s = 1$, $s > t$) as a valid comparison group for those units already treated by time t.

To visualize these restrictions, let us consider our stylized example. Under Assumptions 2.1 - 2.4, the PTA 2.5 can be written as the following seven moment conditions:

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | C = 1] + \mathbb{E}[Y_2 | G_3 = 1], \quad (2.2)$$

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | G_4 = 1] + \mathbb{E}[Y_2 | G_3 = 1], \quad (2.3)$$

$$\alpha_{3,4}(0) = \mathbb{E}[Y_4 - Y_3 | C = 1] + \alpha_{3,3}(0), \quad (2.4)$$

$$\alpha_{4,4}(0) = \mathbb{E}[Y_4 - Y_3 | C = 1] + \mathbb{E}[Y_3 | G_4 = 1], \quad (2.5)$$

$$\mathbb{E}[Y_2 - Y_1 | G_3 = 1] = \mathbb{E}[Y_2 - Y_1 | C = 1], \quad (2.6)$$

$$\mathbb{E}[Y_2 - Y_1 | G_3 = 1] = \mathbb{E}[Y_2 - Y_1 | G_4 = 1] \quad (2.7)$$

$$\mathbb{E}[Y_3 - Y_2 | G_4 = 1] = \mathbb{E}[Y_3 - Y_2 | C = 1]. \quad (2.8)$$

These moment restrictions highlight all restrictions PTA 2.5 impose on the data. First, the moment conditions (2.2) and (2.3) formalize the notion that the evolution of the outcome for the "never-treated" and "late-treated" units can be used to identify $\alpha_{3,3}(0)$, which in turn, would allow one to identify $ATT(3,3)$.[7] An empirically important implication of this

---

[7]Recall that, for $g \geq t$, $ATT(g,t) = \alpha_{g,t}(1) - \alpha_{g,t}(0) = \mathbb{E}[Y_t | G_g = 1] - \mathbb{E}[Y_t(0) | G_g = 1]$. Thus, given that $\mathbb{E}[Y_t | G_g = 1]$ is estimable from the data, one only needs to identify $\alpha_{g,t}(0)$ in order to recover $ATT(g,t)$ from the data.

observation is that *any linear combination* of $\mathbb{E}\left[Y_3 - Y_2 | C = 1\right]$ and $\mathbb{E}\left[Y_3 - Y_2 | G_4 = 1\right]$ can be used to impute $\alpha_{3,3}(0)$. Given that the "never-treated" units are the only units that have not yet experienced treatment at time $t = 4$, they form the only group that can be used to recover $\alpha_{3,4}(0)$ and $\alpha_{4,4}(0)$— this notion is formalized by the moment restrictions (2.4) and (2.5). Finally, (2.6)-(2.8) impose a parallel "pre-trends" condition, i.e., that the evolution of the outcome before treatment occurs is the same across all groups. Note that the moment condition (2.8) is a linear combination of the moment conditions (2.2) and (2.3), so (2.8) is redundant in the aforementioned system of equations. Nonetheless, this observation allows one to conclude that, by assuming that both never-treated units and the units that are not-yet-treated at time $t = 3$ can be used as valid comparison groups for the units first treated at time $t = 3$, one is imposing the parallel pre-trends condition (2.8). Of course, the reverse argument is also true, highlighting that parallel pre-trends across groups may have important identification content; we further discuss this in Section 3.

Given that we now have a better understanding of the PTA invoked by S&A and dC&D, we turn our attention to the PTA invoked by C&S. In fact, C&S consider two different PTA depending on whether a "never treated" group is available or not (see Assumptions 4 and 5 in C&S).

Assumption 2.6 (Parallel trends assumption based on "never treated" units). *For all* $g, t = 2, \ldots, \mathcal{T}$, $g = 2, \ldots, \mathcal{T}$, *such that* $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1\right].$$

Assumption 2.7 (Parallel trends assumption based on "not-yet treated" units). *For all* $g, s, t = 2, \ldots, \mathcal{T}$, *such that* $t \geq g$, $s \geq t$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | D_s = 0\right].$$

The difference between the parallel trends assumptions 2.6 and 2.7 is that the former uses the "never treated" units as a fixed comparison group, whereas the latter allows one to use averages of different groups of units that are not-yet treated by time t as a comparison group. At first sight, it may not be clear whether the PTAs 2.6 and 2.7 also restrict pre-trends as the PTA 2.5 does.

In order to compare these PTAs, it is illustrative to focus our attention to the stylized example where we again pre-impose Assumptions 2.1 - 2.4. In this context, the PTA 2.6 imposes the following three moment restrictions:

$$
\begin{align}
\alpha_{3,3}(0) &= \mathbb{E}\left[Y_3 - Y_2 | C = 1\right] + \mathbb{E}\left[Y_2 | G_3 = 1\right], \tag{2.9}\\
\alpha_{3,4}(0) &= \mathbb{E}\left[Y_4 - Y_3 | C = 1\right] + \alpha_{3,3}(0), \tag{2.10}\\
\alpha_{4,4}(0) &= \mathbb{E}\left[Y_4 - Y_3 | C = 1\right] + \mathbb{E}\left[Y_3 | G_4 = 1\right]. \tag{2.11}
\end{align}
$$

9

As is evident from (2.9)-(2.11), the PTA 2.6 does not restrict pre-trends across groups, and does not presume that "later treated" units can be used as a valid comparison group for "early treated" units. Although the moment conditions (2.9), (2.10), and (2.11) are respectively the same as (2.2), (2.4), and (2.5), it does not impose the moment restrictions (2.3), (2.6), and (2.7) imposed by the PTA 2.5. Therefore, one can reasonably argue that the PTA 2.6 is "weaker" then the PTA 2.5.

Next, we describe the PTA 2.7, which, in the context of our stylized example, imposes the following moment restrictions:

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2|D_3 = 0] + \mathbb{E}[Y_2|G_3 = 1], \tag{2.12}$$

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2|D_4 = 0] + \mathbb{E}[Y_2|G_3 = 1], \tag{2.13}$$

$$\alpha_{3,4}(0) = \mathbb{E}[Y_4 - Y_3|D_4 = 0] + \alpha_{3,3}(0), \tag{2.14}$$

$$\alpha_{4,4}(0) = \mathbb{E}[Y_4 - Y_3|D_4 = 0] + \mathbb{E}[Y_3|G_4 = 1]. \tag{2.15}$$

where $D_{it} = 1$ if a unit i is treated by time t, and equal to zero otherwise. In the context of our stylized example, $D_4 = 0$ if and only if $C = 1$, implying that (2.13)-(2.15) are equivalent to (2.9)-(2.11), respectively. Thus, from this simple observation, we can conclude that the PTA 2.7 is "stronger" then PTA 2.6, as the latter does not involve the moment restriction (2.12).

To compare the PTA 2.7 with the PTA 2.5, we need to understand the implications of adding the moment restriction (2.12) to the other moment restrictions implied by PTA 2.6. Note that when we combine (2.12) with (2.13), we have that $\mathbb{E}[Y_3 - Y_2|D_3 = 0] = \mathbb{E}[Y_3 - Y_2|D_4 = 0]$, which in our example is the same as $\mathbb{E}[Y_3 - Y_2|D_3 = 0] = \mathbb{E}[Y_3 - Y_2|C = 1]$. Given that $D_3 = 0$ if and only if either $G_4 = 1$ or $C = 1$, it follows that

$$\mathbb{E}[Y_3 - Y_2|D_3 = 0] = \mathbb{E}[Y_3 - Y_2|C = 1]$$

$$\Longleftrightarrow$$

$$\mathbb{E}[Y_3 - Y_2|G_4 = 1] = \mathbb{E}[Y_3 - Y_2|C = 1].$$

Thus, by exploiting this simple but subtle observation, we can conclude that the moment restrictions implied by the PTA 2.7, (2.12)-(2.15), are *equivalent* (2.2)-(2.5), a subset of the moment restrictions implied by the PTA 2.5. Importantly, and in contrast with the PTA 2.6, the PTA 2.7 does rule out non-parallel pre-trends for *some* groups and pre-treatment periods,[8] though, technically, it is still weaker than the PTA 2.5 as the latter completely rules out any type of non-parallel pre-trends.

---

[8]The PTA 2.7 does not restrict pre-trends involving time periods before the first unit is treated, and does not restrict pre-trends for the earliest treatment group.

In summary, from the discussion presented above, one can conclude that the PTA 2.6 does not restrict pre-trends and is weaker than the PTA 2.5 and 2.7, though it requires the existence of a "never treated" group. In addition, the PTA 2.7 is arguably weaker then the PTA 2.5, as the latter restricts all pre-trends in all pre-treatment periods, while the former does not restrict pre-trends involving time periods before the first unit is treated. This can be practically relevant in applications where data are available on many time periods before the first group of units is treated.

2.3   Parameters of interest

In this subsection, we discuss the different parameters of interest that may arise when one deviates from the canonical 2×2 DID setting. Before presenting the parameters of interest considered by S&A, C&S, and dC&D, it is worth stressing the potential pitfalls associated with the commonly used TWFE regression specifications.

2.3.1   Pitfalls of TWFE regression specifications

As Borusyak and Jaravel (2017), dC&D and Goodman-Bacon (2019) point out, one of the most popular specifications in this many periods and many groups DID setting is the following TWFE regression specification,[9]

$$Y_{it} = \alpha_g + \alpha_t + \beta_{fe}D_{it} + u_{it}, \tag{2.16}$$

where α_g and α_t are group and time fixed effects, respectively, and u_{it} is an idiosyncratic error term. Although practitioners often consider β_{fe} to be a main parameter of interest, these aforementioned papers show that, when treatment effects are allowed to be heterogeneous across groups and time periods, β_{fe} can only be interpreted as a weighted average of treatment effects, and, perhaps even more problematic, some of these weights can be negative; see also Laporte and Windmeijer (2005), Wooldridge (2005), Chernozhukov et al. (2013), and Gibbons et al. (2018) for earlier related results based on (one-way) fixed-effect estimators. As so, interpreting estimates of β_{fe} as sensible causal summary parameters can lead to misleading conclusions about the policy effectiveness.

Moreover, the negative (and non-intuitive) weighting problem is not specific to (2.16). dC&D show that it also applies to the first-difference specification. In addition, S&A show the non-convex weighting problem extends to many variations of the dynamic TWFE re-

---

[9]All the results remain the same if one replaces α_g with α_i, an unit-specific fixed effect. We prefer to include α_g as it closely resemble the canonical DID regression specification.

gression specification

$$Y_{it} = \alpha_i + \alpha_t + \sum_{e=-K}^{-1} \beta_e 1\{t - G_i + 1 = e\} + \sum_{e=1}^{L} \beta_e 1\{t - G_i + 1 = e\} + v_{it}, \qquad (2.17)$$

where G_i is the time a unit i is first treated (equal to infinity if unit i is "never-treated"), $t - G_i + 1$ is the "event time" , i.e., the number of time periods a unit has been treated, $1\{t - G_i + 1 = e\}$ is an indicator for unit i being treated for e time periods. Taken together, these results suggest that the common practice of attaching sensible causal interpretation to the coefficients of TWFE regression models is not, in general, warranted.

2.3.2 More sensible treatment effect parameters and their estimators

Given the potential pitfalls associated with traditional estimation procedures, S&A, C&S and dC&D propose different estimators for different treatment effect parameters. In this subsection, we review these procedures and highlight their differences.

Given that traditional estimation procedures do not recover easy to interpret causal parameters without further restricting treatment effect heterogeneity, S&A, C&S and dC&D propose different estimators for different treatment effect parameters. In this subsection, we review these procedures.

dC&D focuses on an instantaneous treatment effect measure across all "ever treated" groups. More precisely, dC&D is mainly interested in estimating

$$\delta^S \equiv \mathbb{E}\left[\frac{\sum_{i=1}^{n} \sum_{t=2}^{\mathcal{T}} G_{ig} \cdot (Y_{it}(1) - Y_{it}(0))}{\sum_{i=1}^{n} \sum_{t=2}^{\mathcal{T}} G_{it}}\right], \qquad (2.18)$$

the average of the treatment effect at the time when a group starts receiving the treatment, across all groups that become treated at some point (see Section 4 of dC&D).[10]

dC&D also proposes an easy-to-implement estimator for δ^S. To better understand their estimator, let

$$\widehat{ATT}_{ny}(g,t) = \frac{n^{-1}\sum_{i=1}^{n} G_{ig}(Y_{it} - Y_{ig-1})}{n^{-1}\sum_{i=1}^{n} G_{it}} - \frac{n^{-1}\sum_{i=1}^{n}(1-D_{it})(1-G_{ig})(Y_{it} - Y_{ig-1})}{n^{-1}\sum_{i=1}^{n}(1-D_{it})(1-G_{ig})} \qquad (2.19)$$

be a DID estimator for (2.1) that uses not-yet treated units by time t as a comparison group for treatment group g, at time t. Consider the estimator for the probability of a unit being in group g given that it is among the units that are treated for at least $e = t - g + 1$ periods given by

$$\widehat{w}(g;e) \equiv \widehat{P}(G_g = 1|\text{Treated for} \geq \text{e periods}) = \frac{N_{g \cap \geq e}}{N_{\geq e}}, \qquad (2.20)$$

---

[10]This is the case with staggered treatment adoption. In more general treatment adoption setups, the parameter of interest considered by dC&H differs from δ^S as defined above.

where $N_{g \cap \geq e}$ denotes the number of observations in group g among those units that have been treated for at least e periods, and $N_{\geq e}$ is the number of units who have been treated for at least e periods. dC&D then show that, under the PTA 2.5 and some additional regularity conditions,

$$\widehat{\delta}^S = \sum_{g=2}^{\mathcal{T}} \widehat{P} \left(G_g = 1 | \text{Treated for} \geq 1 \text{ period} \right) \cdot \widehat{ATT}_{ny} \left(g, g \right), \qquad (2.21)$$

is an unbiased estimator of δ^S, and, as (effective) sample size grows, $\widehat{\delta}^S$ is also consistent and asymptotically normal.

From the discussion above, it is evident that (2.21) is a well-defined estimator for the easy-to-interpret causal parameter of interest δ^S as defined in (2.18). On the other hand, $\widehat{\delta}^S$ is, by design, only suitable to summarize instantaneous treatment effects. Hence, when one is interested in treatment effect dynamics, one needs to consider alternative causal parameters of interest.

A particular way of considering more general parameters of interest that are able to capture richer sources of treatment effect heterogeneity is to follow C&S, and center the analysis on the average treatment effect at time t, for those units first treated at time g, $ATT(g,t)$ as defined in (2.1). By doing so, one can highlight different sources of treatment effect heterogeneity. For instance, one can look at how the $ATT(g,t)$ for particular group g evolves over time, which would allow one to study group-specific treatment effect dynamics. Alternatively, one can form different weighted averages of the $ATT(g,t)$ that are able to summarize overall treatment effects. Examples of these weighted averages include (i) a "simple" average of the $ATT(g,t)$,

$$ATT^{simple} = \frac{\sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} P(G = g) \cdot ATT(g,t)}{\sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} P(G = g)}, \qquad (2.22)$$

(ii) the "event-study-type" causal parameter

$$\delta^{es}(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{t - g + 1 = e\} P \left(G_g = 1 | \text{Treated for} \geq e \text{ periods} \right) ATT(g,t), \quad (2.23)$$

which provides the average treatment effect for units that have been treated for e periods, and (iii) the average of $\delta^{es}(e)$ over all possible (positive) values of e,

$$\delta^{e,avg} = \frac{1}{\mathcal{T} - 1} \sum_{e=1}^{\mathcal{T}-1} \delta^{es}(e). \qquad (2.24)$$

Note that all these weighted averages of the $ATT(g,t)$ are easy-to-interpret, less "data hungry" than the disaggregated $ATT(g,t)$, and can be use to summarize short, medium and long run effects of a given policy. In fact, one can show that (2.18) is equal to $\delta^{es}(e)$ with $e = 1$.

The key challenge in estimating all these causal parameters of interest is to show that one can indeed nonparametrically point-identify all $ATT(g,t)$'s with $t \geq g$. C&S shows that one can bypass such challenges by imposing either the PTA 2.6 or 2.7, though each assumption leads to a different estimand. More precisely, C&S shows that, for $t \geq g$, $ATT(g,t)$ is nonparametrically identified by

$$ATT_{never}(g,t) = \mathbb{E}\left[Y_t - Y_{g-1} | G_g = 1\right] - \mathbb{E}\left[Y_t - Y_{g-1} | C = 1\right], \qquad (2.25)$$

$$ATT_{ny}(g,t) = \mathbb{E}\left[Y_t - Y_{g-1} | G_g = 1\right] - \mathbb{E}\left[Y_t - Y_{g-1} | D_t = 0, G_g = 0\right], \qquad (2.26)$$

when one respectively imposes either the PTA 2.6 or 2.7. These quantities can be straightforwardly estimated by

$$\widehat{ATT}_{never}(g,t) = \frac{n^{-1}\sum_{i=1}^{n} G_{ig}(Y_{it} - Y_{ig-1})}{n^{-1}\sum_{i=1}^{n} G_{ig}} - \frac{n^{-1}\sum_{i=1}^{n} C_i(Y_{it} - Y_{ig-1})}{n^{-1}\sum_{i=1}^{n} C_i} \qquad (2.27)$$

and by $\widehat{ATT}_{ny}(g,t)$ as defined in (2.19).

With either $\widehat{ATT}_{never}(g,t)$ or $\widehat{ATT}_{ny}(g,t)$ on hand, one can then form the more aggregated parameters by replacing $\widehat{ATT}(g,t)$ by either of these estimators, and by replacing the weights by their natural (plug-in) estimators. For instance, depending on whether one imposes the parallel trends assumption 2.6 or 2.7, one can naturally estimate $\delta^{es}(e)$ by

$$\widehat{\delta}_{never}^{es}(e) = \sum_{g=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}} 1\{t - g + 1 = e\} \widehat{w}(g;e)\widehat{ATT}_{never}(g,t),$$

$$\widehat{\delta}_{ny}^{es}(e) = \sum_{g=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}} 1\{t - g + 1 = e\} \widehat{w}(g;e)\widehat{ATT}_{ny}(g,t),$$

respectively, where $e \geq 1$ and $\widehat{w}(g;e)$ is defined as in (2.20). The aggregated estimators for ATT^{simple} and for $\delta^{e,avg}$ are formed analogously.

C&S derive the large sample properties of all these aforementioned estimators and propose bootstrap procedures to construct simultaneous confidence bands for these treatment effect measures. They emphasize the practical importance of using simultaneous inference procedures when one estimates multiple parameters of interest (e.g., when one estimate $\delta^{es}(e)$ for multiple e's), as failing to account for multiple testing usually lead to misleading inference.

We conclude this subsection by noting that S&A is mainly interested in recovering the event-study-type parameter $\delta^{es}(e)$. More precisely, S&A propose the following interaction-weighted estimator for $\delta^{es}(e)$ (see Section 4 of S&A). In the first-step, they use the linear two-way fixed effects specification that interacts relative time indicators with treatment group indicator:

$$Y_{it} = \lambda_i + \lambda_t + \sum_{g=2}^{\mathcal{T}-1}\sum_{e\neq 0} \delta_{ge} \cdot G_{ig} 1\{t - G_i + 1 = e\} + v_{it} \qquad (2.28)$$

14

on observations from $t = 1, \ldots, \mathcal{T} - 1$, where the last time period $\mathcal{T}$ is dropped in order to accommodate the case where there is no "never treated" group; if there is a never-treated group available, dropping data from time period $\mathcal{T}$ is unnecessary. S&A shows that, under the PTA 2.5, the estimator $\widehat{\delta}_{ge}$ is consistent for $ATT(g,t)$, $t - g + 1 = e$. Here, it is important to emphasize that, when a "never treated" group is not available, only the units treated at the last time period are used as comparison units when computing $\widehat{\delta}_{ge}$, which differs from the C&S proposed estimator (2.19) that uses not-yet treated units as comparison units. When a "never treated" group is available, though, the interaction-weighted estimator $\widehat{\delta}_{ge}$ is equivalent to $\widehat{ATT}_{never}(g,t)$ once one maps event-time to calendar time (or vice-versa).

Armed with $\widehat{\delta}_{ge}$, S&A then propose to estimate $\delta^{es}(e)$ by

$$\widehat{\delta}^{es}_{S\&A}(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\left\{t - g + 1 = e\right\} \widehat{w}(g;e)\widehat{\delta}_{gt},$$

where $\widehat{w}(g;e)$ is defined as in (2.20). S&A establish the large sample properties of $\widehat{\delta}^{es}_{S\&A}(e)$ and provide valid pointwise inference procedures for $\delta^{es}(e)$.

Remark 2.1. *Many times one wishes to check for the existence of non-parallel pre-trends as a way to assess the credibility of the DID setup. We note that one can use $\delta^{es}_{never}(e)$ and/or $\delta^{es}_{ny}(e)$ with $e < 0$ as estimators of pre-trends, though, when e is negative one must replace the estimated weights $\widehat{w}(g;e)$ as defined as in (2.20) with*

$$\widehat{w}(g;e-) \equiv \widehat{P}(G_g = 1 | at\ least\ |e|\ pre\text{-}treatment\ periods\ available) = \frac{N_{g \cap \geq e-}}{N_{\geq e-}},$$

where $N_{g \cap \geq e-}$ denotes the number of observations in group g among those units that have least $|e|$ pre-treatment time periods of data available, and $N_{\geq e-}$ is the number of units who have least $|e|$ pre-treatment time periods of data available. Importantly, these event-study estimators avoid the pitfalls associated with using the dynamic TWFE to assess the credibility of parallel trends; see S&A for a detailed discussion of this important issue.

3   Not all parallel trends assumptions are made equal

From the discussion in the previous section, it is clear that in DID designs with staggered treatment adoption one can make different parallel trends assumptions, and can estimate different parameters of interest. The discussion in Section 2 also indicates that, despite their peculiarities, these different parameters of interest can be estimated using weighted averages of estimators for the $ATT(g,t)'s$. From this observation, one can reasonably argue that identifying (and estimating) the $ATT(g,t)'s$ from the data is the most challenging step of the analysis, and that the different PTAs help researchers to overcome it. Once this is done,

constructing event-study-type estimators, for instance, becomes straightforward.

In practice, however, researchers must choose and justify the use of a given PTA. In this section, we aim to highlight some practical consequences of adopting different versions of the PTA. In order to simplify the discussion, we exploit the stylized example introduced in Section 2 whenever possible, and we implicitly impose Assumptions 2.1-2.4.

3.1   Are these parallel trends testable?

Practitioners routinely use estimates of pre-treatment event-study coefficients to assess the credibility of an underlying PTA.[11] Can these tests for parallel pre-treatment trends be interpreted as *direct* tests for the validity of the underlying PTA, or should these tests be interpreted as "placebo/falsification" type of tests? With the help of the stylized example, we show that the answer depends on the chosen PTA.

Let us first consider the case of Assumption 2.5. As it is evident from (2.2)-(2.8), the PTA 2.5 imposes six linearly independent moment restrictions to recover three counterfactual parameters, $\alpha_{3,3}(0)$, $\alpha_{3,4}(0)$ and $\alpha_{4,4}(0)$. That is, imposing the PTA 2.5 leads to an overidentified system of equations and, consequently, we can directly test for the validity of the PTA 2.5. Indeed, it is easy to see that the PTA 2.5 implies parallel pre-treatment trends across every group, see e.g., (2.6)-(2.8), and such restrictions can be directly assessed from the data, for instance, by testing if pre-treatment event-study-type estimates of (2.23) are all equal to zero. Thus, under the PTA 2.5, non-zero pre-treatment event-study estimates should be interpreted as direct evidence against the identifying assumptions.[12]

A somewhat similar conclusion is also reached when one relies on the PTA 2.7: (2.12)-(2.15) suggest four linearly independent moment restrictions to recover three counterfactual parameters, which also leads to an overidentified system of equations. Recalling that (2.12)-(2.15) is equivalent to (2.2)-(2.5), we can then see from (2.2)-(2.3) that the PTA 2.7 also imposes parallel pre-trends among "never treated" ($C = 1$) and the "later treated"($G_4 = 1$) from time $t = 2$ to $t = 3$, but does not restrict pre-trends of these two groups from $t = 1$ to $t = 2$, nor the pre-trends of the early treated group ($G_3 = 1$). Moving from the stylized example to the general case, we have that PTA 2.7 imposes parallel pre-treatment

---

[11] As stressed by S&A, one should not use pre-treatment coefficients from TWFE event-study-type regressions to assess the credibility of the PTA as these coefficients can be contaminated with post-treatment effects. Using estimates of (2.23) with $e < 0$, on the other hand, does not suffer from these pitfalls. Please refer to S&A for a detailed discussion about these issues.

[12] If Assumption 2.3 is violated, though, it is possible that violations from Assumption 2.3 "offset" violations of Assumption 2.5 and the test based on (2.6)-(2.8) would not capture such violations. This should always be taken into account. In addition, failing to reject these tests should not be interpret as evidence in favor of the identifying assumptions, as it may be the case that the test lacks power to detect some non-trivial deviations from the null.

trends from $t = g_{min} - 1$ to $t = \mathcal{T}$ for all groups except the first-treated group (who is treated at time g_{min}). Interestingly, because estimators of (2.23) with $e < 0$ exploit *all* pre-treatment trends (including the one for group $G_3 = 1$ in our stylized example), non-zero pre-treatment estimates can not, at least strictly speaking, be interpreted as direct tests of the identifying assumptions, but rather as placebo-type tests. Nonetheless, one can easily bypass this limitation by constructing alternative tests for the identifying assumptions. For instance, in the context of our stylized example, one can directly test if $\mathbb{E}\left[Y_3 - Y_2 | C = 1\right] = \mathbb{E}\left[Y_3 - Y_2 | G_4 = 1\right]$ using a standard t-test. Rejecting the null hypothesis would provide direct evidence against the identifying assumptions.

Finally, note that the conclusion is very different when one imposes the PTA 2.6: (2.9)-(2.11) suggest we have a just-identified system of equations, implying that the PTA 2.6 can not be directly tested. Indeed, as we dicussed in Section 2.2, PTA 2.6 does not restrict pre-treatment trends, and, therefore, event-study estimates for pre-treatment periods provide, at best, placebo-type evidence against the PTA 2.6.

The discussion above highlights that whether tests for parallel pre-treatment trends provide direct or indirect evidence against the invoked identifying assumptions crucially depends on the invoked PTA. This is very different from the case where treatment adoptions does not vary across time. In that case, tests for non-parallel pre-treatment trends *always* provide only indirect evidence against the adopted design.

3.2 Can the parallel trends assumption guide us on choosing a DID estimator?

In this section, we discuss some potential trade-offs one may face when adopting different PTAs, as they can lead to different DID estimators.

3.2.1 What if we impose the PTA 2.6?

The PTA 2.6 is the weakest PTA among the three we have considered so far as it does not impose any restriction on pre-treatment trends across groups. Given that this PTA leads to a just-identified system of equations, in situations where researchers are not willing to impose additional restrictions on the data, $\widehat{ATT}_{never}\left(g, t\right)$ as defined in (2.27) is the *only* suitable estimator for the $ATT\left(g, t\right)$ and their different functionals such as the event-study parameters (2.23).

Of course, in order to rely on the PTA 2.6 and use the DID estimator (2.27), we must have a set of units that do not experience treatment in the time-window we want to analyze. When such a group of units is available but its relative size is small, inference procedures

17

based on (2.27) may not be as precise as one wishes. However, it is important to stress that this potential "loss of efficiency" is a direct consequence of not exploiting restrictions on pre-treatment trends across groups.

In practice, we foresee researchers taking into account this "robustness" versus "efficiency" trade-off when deciding if the PTA 2.6 is the most suitable for the given application. In situations where there is a "reasonably large" number of units that cannot be treated because of some application-specific institutional detail, we expect that the gains in robustness should dominate the potential gains in efficiency associated with using other PTAs and DID estimators. The same holds true if researchers are not comfortable with *a priori* ruling out non-parallel pre-trends. In these cases, we foresee researchers favoring the PTA 2.6 and the DID estimator $\widehat{ATT}_{never}(g, t)$ over the other alternatives.

3.2.2 What if we impose either the PTA 2.5 or the PTA 2.7?

In many situations, a "never-treated" group is not available, implying that the PTA 2.6 does not provide any identifying restriction that can be used to estimate the $ATT(g, t)$'s. In other cases, the "never-treated" group may be "too small" to be of practical use, and/or researchers may be *a priori* comfortable restricting pre-treatment trends and using those "not-yet-treated" units as valid comparison groups for those "earlier-treated". In such cases, researchers can then choose between the PTA 2.5 and the PTA 2.7. In both cases, though, they can use the the DID estimator $\widehat{ATT}_{ny}(g, t)$, as defined in (2.19), to study policy effectiveness.

Although (2.19) can be used under either PTA, we still recommend researchers to explicitly specify which PTA they are making for at least three reasons. First, being explicit about the identifying assumption adds transparency to the analysis, which is always desirable. Second, the interpretation of pre-tests based on event-study-type estimates can vary depending on the assumptions, as we discusses in Section 3.1. Third, the choice of the PTA has an important impact on what other estimators you could use instead of (2.19). This is particularly important because (2.19) does not fully exploit all the restrictions imposed by either PTA. Being aware of the exact PTA invoked allows researchers to adopt an alternative estimation procedure that fully exploits all these moment restrictions, resulting in estimators that are more efficient than (2.19). Here, we stress that the gains in efficiency will vary depending on the underlying PTA used, as the PTA 2.5 imposes more restrictions on the data than PTA 2.7.

Before describing how one can exploit these additional moment restrictions to form a more efficient DID estimator, it is worth describing situations where researchers may favor either the PTA 2.5 or the PTA 2.7. Recall that the main difference between PTA 2.5 and PTA

2.7 is that the former imposes parallel pre-treatment trends across all groups and all time periods, whereas the latter only restricts pre-treatment trends since the time the first group is treated. These differences can be meaningful in applications where data on multiple time periods before the first group of units is treated are available, and the economic environment in these "early-periods" were potentially different from the "later-periods". In these cases, the outcome of the different groups may evolve in a non-parallel manner during "early-periods" periods, perhaps because the groups were exposed to different shocks, but these non-parallel trends become less of a concern as time pass by. In such cases, we expect researchers to favor the PTA 2.7 over the PTA 2.5. In other situations, though, researchers may prefer to impose the PTA 2.5, allowing them to enjoy some potential gains in efficiency if they use estimators that exploit the additional restrictions imposed by the PTA 2.5 when compared to the PTA 2.7. Again, the "robustness" versus "efficiency" trade-off should be taken into account when deciding which PTA is more appropriate for the specific application.

4   Using GMM to estimate DID parameters

As we described in Section 3.2, in situations where researchers are comfortable with imposing either the PTA 2.5 or the PTA 2.7, the DID estimator (2.19) is not efficient, as it does not fully exploit all the restrictions implied by these PTAs. In this section, we describe how one can exploit all the restrictions implied by the identifying assumptions to form efficient DID estimators by casting the problem into the familiar GMM framework (Hansen, 1982). In what follows, we provide a step-by-step description of how one can form these efficient GMM DID estimators. To avoid repetition, we focus on the case where researchers impose the PTA 2.7; the implementation based on the PTA 2.5 is completely analogous.

The key to implement the GMM is to list *all* moment restrictions we are imposing to recover the $ATT(g,t)$'s, which involves not only the moment restrictions implied by the PTA 2.7, but also the *observational restrictions* that, for all $t \geq g$, $\alpha_{g,t}(1) \equiv \mathbb{E}[Y_t(1)|G_g = 1] = \mathbb{E}[Y_t|G_g = 1]$, $\alpha_g^{prop} \equiv \mathbb{E}[G_g]$, $\alpha_C^{prop} \equiv \mathbb{E}[C]$. We can then use these "augmented" moment restrictions (consisting of observational restriction and all the moment restrictions implied by the PTA) to efficiently estimate all the unknown parameters involved in our problem by following Hansen (1982).

To gain more intuition on how to implement the efficient GMM, we turn of attention to our stylized example. In this case, the unknown parameters consist of $\alpha \equiv (\alpha_{3,3}(1), \alpha_{3,3}(0), \alpha_{3,4}(1), \alpha_{3,4}(0), \alpha_{4,4}(1), \alpha_{4,4}(0), \alpha_C^{prop}, \alpha_3^{prop}, \alpha_4^{prop})'$, which can be efficiently estimated by

$$\widehat{\alpha}^{gmm} = \arg\min_{\alpha \in \Theta} \bar{g}_\alpha(W)' \widehat{\Sigma}_{\bar{\alpha},gmm}^{-1} \bar{g}_\alpha(W), \tag{4.1}$$

where $\bar{g}_\alpha(W)$ is the sample average of the augmented moment conditions, $n^{-1}\sum_{i=1}^n g_\alpha(W_i)$, with $g_a(W_i)$ combining all (linearly independent) moment conditions[13],and

$$\widehat{\Sigma}_{\check{\alpha},gmm} = \frac{1}{n}\sum_{i=1}^n g_{\check{\alpha}}(W_i) g_{\check{\alpha}}(W_i)',$$

$\check{\alpha}$ being a preliminary consistent estimator for α, say the minimizer of (4.1) with $\widehat{\Sigma}_{\check{\alpha},gmm}$ replaced by the identity matrix.

With $\widehat{\alpha}^{gmm}$, one can then efficiently estimate the parameters of interest: $ATT(3,3)$, $ATT(3,4)$ and $ATT(4,4)$ by

$$\widehat{ATT}_{gmm}\begin{pmatrix} 3,3 \\ 3,4 \\ 4,4 \end{pmatrix} = \begin{pmatrix} \widehat{\alpha}_{3,3}^{gmm}(1) - \widehat{\alpha}_{3,3}^{gmm}(0) \\ \widehat{\alpha}_{3,4}^{gmm}(1) - \widehat{\alpha}_{3,4}^{gmm}(0) \\ \widehat{\alpha}_{4,4}^{gmm}(1) - \widehat{\alpha}_{4,4}^{gmm}(0) \end{pmatrix}. \tag{4.2}$$

In what follows, we establish the asymptotic properties of $\widehat{\alpha}^{gmm}$. The asymptotic properties of $\widehat{ATT}_{gmm}(g,t)$ follow directly from the delta method. Let $\Delta Y_t = Y_t - Y_{t-1}$. Define $\Sigma_{\alpha,gmm}$ as the probability limit of $\widehat{\Sigma}_{\check{\alpha},gmm}$, and

$$\Psi \;=\; \mathbb{E}\left[\frac{\partial g_\alpha(W)}{\partial \alpha}\right]$$

Let the vector of scores associated with the efficient GMM estimator be defined as

$$\phi_\alpha^{gmm}(W_i) = -\left(\Psi'\Sigma_{\alpha,gmm}^{-1}\Psi\right)^{-1}\Psi'\Sigma_{\alpha,gmm}^{-1}\cdot g_\alpha(W_i).$$

Proposition 4.1. *Assume that all random variables have finite second moments, $\Sigma_{\alpha,gmm}$ is positive definite, and that Assumptions 2.1-2.4 hold. Then, when the parallel trends assumption 2.7 holds, we have that:*

(i) As $n \to \infty$,

$$\begin{aligned} \sqrt{n}\left(\widehat{\alpha}^{gmm} - \alpha\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_\alpha^{gmm}(W_i) + o_p(1) \\ &\xrightarrow{d} N\left(0, \left(\Psi'\Sigma_{\alpha,gmm}^{-1}\Psi\right)^{-1}\right). \end{aligned}$$

(ii) The GMM estimator $\widehat{\alpha}^{gmm}$ is semiparametrically efficient.

Proof. Proof is presented in the Web Appendix C. $\qquad\square$

Proposition 4.1 has important practical implications, which we illustrate in the context of our stylized example. First, and perhaps most important, it implies that

$$\sqrt{n}\left(\widehat{ATT}_{gmm} - ATT\right)\begin{pmatrix} 3,3 \\ 3,4 \\ 4,4 \end{pmatrix} \xrightarrow{d} N(0,\Omega)$$

---

[13]See Web Appendix A for the details about $g_a(W_i)$.

with $\Omega = A \left(\Psi' \Sigma_{\alpha,gmm}^{-1} \Psi \right)^{-1} A'$, A is a "selection matrix"[14] and that Ω is equal to the semi-parametric efficiency bound for $ATT(g,t)$ under the PTA 2.7. As so, $\widehat{ATT}_{gmm}$ exploits all available information in the data to estimate the $ATT(g,t)$'s, which, in general, translates to tighter confidence intervals. In fact, under the PTA 2.7, the GMM DID estimator $\widehat{ATT}_{gmm}(g,t)$ is, in general, more efficient than $\widehat{ATT}_{ny}(g,t)$ or $\widehat{ATT}_{never}(g,t)$ as defined in (2.19), and (2.27), respectively, or those based on the "interaction-weighted" regression (2.28). This is a main advantage of the GMM DID estimator when compared to the other available estimators.

A second implication of Proposition 4.1 is that, given that we have an *overidentified* system of equations, one can directly use the Sargan-Hansen J-test as a test for the validity of the PTA 2.7. More precisely, under the null hypothesis that the PTA 2.7 is true,

$$ J = n \cdot \left(\bar{g}_{\widehat{\alpha}^{gmm}}(W)' \; \widehat{\Sigma}_{\widehat{\alpha}^{gmm},gmm}^{-1} \; \bar{g}_{\widehat{\alpha}^{gmm}}(W) \right) \xrightarrow{d} \chi_{10-9}^2 \text{ as } n \to \infty. $$

If the PTA 2.7 holds, any deviation of J from zero should be within the range of sampling error, whereas if the PTA 2.7 is violated, J should be "large." Thus, the Sargan-Hanen J-test can be useful for detecting violations of the PTA 2.7.

At this moment, one may wonder about the situations where one may favor the GMM DID estimator (4.2) over the simpler DID estimator (2.19). Given that (4.2) is more efficient than and as-robust-as (2.19), the only obstacle we see for its widespread adoption is its implementation: whenever the number of treatment groups and/or time periods available are large, the number of moment conditions needed to be considered into the efficient GMM can be fairly large — in our application, for example, where we have 16 treatment groups and 33 time periods, the GMM involves 780 moments with *195 overidentification restrictions*, whereas sample size (state-year pairs) is equal to 759. In such cases, we expect researchers to favor the simpler, but inefficient DID estimator (2.19). However, when the number of groups and/or time periods is moderate such that implementation of the efficient GMM is not challenging, we would recommend using it.

5   A simpler and more robust DID estimator

As highlighted in the previous section, the main attractive feature of using the GMM estimation procedure is that it leads to efficient estimators that fully exploit all the available information compatible with the underlying identifying assumptions. On the other hand, the implementation of such GMM DID estimator is not always straightforward.

In this section, we describe an alternative DID estimator for the $ATT(g,t)$. Although

---

[14]See Web Appendix A for its formal definition.

the estimator is usually less efficient than the GMM, it is (i) easy to compute, (ii) does not require the existence of a "never-treated" group, (iii) exploits more data than (2.19), and (iv) does not explicitly restrict pre-trends as it relies on a "weaker" PTA described below.

Assumption 5.1 ("Weaker" Parallel trends assumption based on "not-yet treated" units). *For all $g, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid D_t = 0\right].$$

The PTA 5.1 imposes that the evolution of the outcome at time t among those units that have not yet experienced treatment by time t can help us identify the $ATT(g, t)$'s. Unlike PTA 2.7, it does not impose that *every individual* not-yet-treated group can be used as a comparison group, which, in turn, suggests that the $ATT(g, t)$, $t \geq g$ are nonparametrically *just-identified*. We formalize this result in the next proposition. Let $\Delta Y_t = Y_t - Y_{t-1}$ denote the first-difference of Y_t.

Proposition 5.1. *Assume that Assumptions 2.1-2.4 hold. Then, when the parallel trends assumption 5.1 holds, it follows that, for $2 \leq g \leq t \leq \mathcal{T}$, $ATT(g, t) = ATT_{ny+}(g, t)$, where*

$$ATT_{ny+}(g, t) \equiv \mathbb{E}\left[Y_t - Y_{g-1} \mid G_g = 1\right] - \left(\sum_{s=g}^{t} \mathbb{E}\left[\Delta Y_s \mid D_s = 0, G_g = 0\right]\right). \quad (5.1)$$

Proof. Proof is presented in the Web Appendix C. □

To better grasp how the PTA 5.1 allows us to use $ATT_{ny+}(g, t)$ as an estimand for the $ATT(g, t)$, $t \geq g$, it is illustrative to go back to our stylized example. In this specific context, we have that, under Assumptions 2.1-2.4, the PTA 5.1 is equivalent to the following restrictions:

$$\alpha_{3,3}(0) = \mathbb{E}\left[Y_3 - Y_2 \mid D_3 = 0\right] + \mathbb{E}\left[Y_2 \mid G_3 = 1\right], \quad (5.2)$$

$$\alpha_{3,4}(0) = \mathbb{E}\left[Y_4 - Y_3 \mid D_4 = 0\right] + \alpha_{3,3}(0), \quad (5.3)$$

$$\alpha_{4,4}(0) = \mathbb{E}\left[Y_4 - Y_3 \mid D_4 = 0\right] + \mathbb{E}\left[Y_3 \mid G_4 = 1\right]. \quad (5.4)$$

By listing these restrictions we can now see how we get $ATT_{ny+}(g, t)$, as defined in (5.1). First, from (5.2) and (5.4), it follows that, when $g = t$, $\alpha_{g,t}(0)$ can be explicitly written as functionals of observable data (and not potential outcomes). As so, $ATT(g, t)$ is identified by (5.1). Interestingly, in this case with $g = t$, (5.1) reduces to $ATT_{ny}(g, t)$, as defined in (2.26). When one moves away from the "instantaneous average treatment effects", though, these two estimands differ. Indeed, by exploiting the moment restrictions (5.2) and (5.3), we can see that the $ATT(3, 4)$ is nonparametrically identified by

$$ATT_{ny+}(3, 4) = \mathbb{E}\left[Y_4 - Y_2 \mid G_3 = 1\right] - \left(\mathbb{E}\left[Y_3 - Y_2 \mid D_3 = 0\right] + \mathbb{E}\left[Y_4 - Y_3 \mid D_4 = 0\right]\right),$$

which, of course, is a special case of (5.1). Here, we stress that $ATT_{ny+}(3,4)$ uses data from all groups, $G_3 = 1$, $G_4 = 1$, and $C = 1$, whereas $ATT_{ny}(3,4)$ only uses data from $G_3 = 1$ and $C = 1$. Hence, one may expect that estimators based on (5.1) to be more precise than (2.19) because they utilize more data. Furthermore, because the PTA 5.1 does not restrict pre-trends, i.e., it does not impose that $\mathbb{E}[Y_3 - Y_2|D_3 = 0] = \mathbb{E}[Y_3 - Y_2|D_4 = 0]$ as implied by (2.12) and (2.13), one can also expect additional gains in "robustness" by exploiting (5.1) instead of (2.26).

Next, we discuss how one can exploit Proposition 5.1 to estimate the $ATT(g,t)$s. Here, the most natural way to proceed is to use the sample analogue of (5.1):

$$\widehat{ATT}_{ny+}(g,t) = \frac{n^{-1}\sum_{i=1}^{n} G_{ig}(Y_{it} - Y_{ig-1})}{n^{-1}\sum_{i=1}^{n} G_{it}} - \sum_{s=g}^{t}\left(\frac{n^{-1}\sum_{i=1}^{n}(1 - D_{is})(1 - G_{ig})\Delta Y_{is}}{n^{-1}\sum_{i=1}^{n}(1 - D_{is})(1 - G_{ig})}\right).$$
(5.5)

Note that (5.5) is very easy to compute as it only involves combinations of sample means. Next, we show that these DID estimators also enjoy good asymptotic properties. More precisely, we prove they are $\sqrt{n}$-consistent and establish their *joint* asymptotic distribution. Before we present the results, we need to introduce some additional notation. For each (g,t)-pair, let $\phi_{ny+}(W_i; g, t)$ be the influence function of $\widehat{ATT}_{ny+}(g,t)$,

$$\phi_{ny+}(W_i; g, t) = \left(\frac{G_{ig}}{\mathbb{E}[G_g]}\left((Y_{it} - Y_{ig-1}) - \frac{\mathbb{E}[G_g \cdot (Y_t - Y_{g-1})]}{\mathbb{E}[G_g]}\right)\right.$$
$$\left. - \sum_{s=g}^{t}\frac{(1 - D_{is})(1 - G_{ig})}{\mathbb{E}[(1 - D_s)(1 - G_g)]}\left(\Delta Y_{is} - \frac{\mathbb{E}[(1 - D_s)(1 - G_g) \cdot \Delta Y_s]}{\mathbb{E}[(1 - D_s)(1 - G_g)]}\right)\right).$$

Finally, let $\widehat{ATT}_{ny+}(t \geq g)$ and $ATT(t \geq g)$ denote the vector of $\widehat{ATT}_{ny+}(g,t)$ and $ATT(g,t)$, respectively, for all $g, t = 2, \ldots, \mathcal{T}$ with $t \geq g$. Analogously, let $\Phi_{ny+}(W_i; t \geq g)$ denote the collection of $\phi_{ny+}(W_i; g, t)$ across all periods t and groups g such that $t \geq g$. [15]

Proposition 5.2. *Assume that Assumptions 2.1-2.4 and Assumption 5.1 hold. Then, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{ATT}_{ny+} - ATT\right)(g,t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{ny+}(W_i; g, t) + o_p(1).$$
(5.6)

Furthermore,
$$\sqrt{n}\left(\widehat{ATT}_{ny+}(t \geq g) - ATT(t \geq g)\right) \xrightarrow{d} N(0, V),$$
(5.7)

with $V = \mathbb{E}\left(\Phi_{ny+}(W; t \geq g)\Phi_{ny+}(W; t \geq g)'\right)$.

Proof. Proof is presented in the Web Appendix C. □

---

[15] We restrict our attention to $t \geq g$ just because these are the post-treatment periods, which presumably are the periods of main interest for the analysis. However, our results naturally extend to the case where one consider all possible g, t's, with the caveat that $ATT_{ny+}(g,t)$ may differ from $ATT(g,t)$ for $t < g$, as the PTA 5.1 does not explicit restrict pre-trends.

Proposition 5.2 provides the influence function for estimating the vector of group-time average treatment effects, as well as its limiting distribution. Interestingly, Proposition 5.2 suggests two potential paths to conduct inference for the $ATT(g, t)$'s. The first and perhaps more standard approach is to use the analogy principle and directly estimate V, which leads directly to standard errors and pointwise confidence intervals. However, it is worth stressing that when one is interested in making inference about *multiple $ATT(g, t)$'s*, inference procedures based on this standard approach such as those based on traditional t-tests and/or individual confidence intervals are usually inappropriate as they do not account for the fact that one is (implicitly) conducting *multiple hypotheses testing*. As a direct consequence, significant treatment effects may emerge simply by chance, even when all $ATT(g, t)$'s are equal to zero, see, e.g., Romano and Wolf (2005), Anderson (2008) and section 8 of Romano et al. (2010).

An alternative path to conduct asymptotically valid inference for multiple parameters of interest that is robust against the multiple-testing problem is to leverage the asymptotic linear representation (5.6) to construct computationally-simple bootstrapped *simultaneous* confidence intervals for multiple $ATT(g, t)$. The idea of this bootstrap procedure is fairly simple, and each bootstrap iteration simply amounts to "perturbing" the asymptotic linear representation of the $\widehat{ATT}_{ny+}(g, t)$'s by a random weight V, and it does not require re-estimating the $ATT(g, t)$'s at each bootstrap draw. In Web Appendix B, we provide a step-by-step description of how one can implement such a procedure.

Remark 5.1. *It is worth stressing that the $ATT_{ny+}(g, t)$ estimand defined in (5.1) is only suitable for post-treatment periods, i.e., for $t \geq g$. Hence, in contrast to (2.25) and (2.26), we can not fix the estimand to analyze pre-treatment periods $t < g$. To address this issue, we suggest using the estimand*

$$ATT_{ny+}^{pre}(g, t) \equiv \mathbb{E}\left[\Delta Y_t \,|\, G_g = 1\right] - \mathbb{E}\left[\Delta Y_t \,|\, D_t = 0, G_g = 0\right], \quad for \ \ t < g, \qquad (5.8)$$

which should be equal to zero under Assumptions 2.1-2.4 and a stronger version of the PTA 5.1 that holds for both pre and post-treatment periods (and not only for post-treatment periods as PTA 5.1). One can then use estimates of (5.8) to provide indirect evidence for the PTA 5.1, as PTA 5.1 cannot be directly tested. We stress that (5.8) should not be directly compared with (2.25) and (2.26) for $t < g$ as (5.8) measures "local deviations" (from time $t - 1$ to t) of a zero pre-treatment trends conditions, whereas (2.25) and (2.26) capture "cumulative deviations" (from time t until $g - 1$) of zero pre-treatment trends conditions.

Remark 5.2. *It is straightforward to build on $\widehat{ATT}_{ny+}(g, t)$ to construct event-study estimators for $\delta^{es}(e)$ as defined in (2.23). Following the same steps described in Section 2.3.2,*

a natural estimator for $\delta^{es}(e)$ is

$$\widehat{\delta}^{es}_{ny+}(e) = \sum_{g=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}} 1\{t-g+1=e\}\, \widehat{w}(g;e)\widehat{ATT}_{ny+}(g,t),$$

where the weights $\widehat{w}(g;e)$ are as defined in (2.20). By building on Theorem 5.2 and the fact that the weights admit an asymptotic linear representation, it is easy to show that $\widehat{\delta}^{es}_{ny+}(e)$ is consistent and asymptotically normal; see, e.g. C&S.

We conclude this section by highlighting situations where we foresee (5.5) being favored over the other available DID estimators. First, we envision researchers favoring (5.5) over (2.19) in situations where they are not comfortable explicitly restricting pre-trends and/or when they want to use data from all groups to estimate the $ATT(g,t)$'s. This can be particularly relevant when one wants to conduct cluster-robust inference when only a moderate number of clusters is available. We also expect researchers to favor (5.5) over the efficient GMM estimator when implementation is challenging. In this case, we expect that (5.5)'s "easiness-to-use" would dominate the potential efficiency gains of using GMM. Finally, we expect researcher to favor (5.5) over (2.27) when a "never-treated" group is relatively small, though we stress that these two estimators rely on non-nested PTAs.

Remark 5.3. *Given that different estimators (and PTAs) have different implications for robustness and efficiency, it may be tempting to "specific-to-general" specification search: start the analysis considering estimators that rely on "stronger" assumptions and then test the validity of these assumptions; in case one does not reject them, one stop and uses the "more efficient" estimators, but in case one rejects the invoked PTA, one then chooses a "more robust" but "less efficient" DID estimator. Although fairly intuitive, the aforementioned strategy is dangerous and should not be used in practice since this specification search is based on multiple-testing procedure, and, as so, inference procedures that treat the "final" estimator (or the "winner") as "true" can be severely distorted, see, e.g. Roth (2020) for detailed discussion of this issue. Hence, we argue that researchers should select the PTA taking into account the "robustness" versus "efficiency" trade-off, and that these considerations should be done based on external, context-specific information, and not on pre-tests.*

6 The effect of enforcing the Clean Water Act

To illustrate the inherent trade-offs described above, we replicate Katherine Grooms' (2015) analysis of the transition from federal to state management of the Clean Water Act (CWA). Environmental policy mandated at the federal level is often implemented at the state level. Yet, there exists variation in the level of enforcement across states. Grooms (2015) exploits

the staggered timing of the transfer from federal to state monitoring and enforcement of the CWA. Using TWFE specifications akin to (2.16) and (2.17), she finds that state-level prevalence of corruption plays an important role in the enforcement and compliance of environmental regulation after transitioning to state control.

We begin by describing the data, and then we discuss the practical relevance of the key assumptions and specifications we use for the analysis given our context. Finally, we show the baseline and corruption-specific results for both the TWFE specification and the new DID estimators that rely on the different PTAs discussed above. Finally, we discuss the implications of the findings and the importance of choosing an appropriate PTA.

6.1 Data

We follow the data construction from Grooms (2015) as closely as possible. Table D.1 in Web Appendix D replicates key summary statistics from Grooms (2015) and provides additional detail on data sources and construction. As described further in Web Appendix D, we follow Grooms (2015) to construct a measure of the fraction of total facilities with at least one inspection, violation, or enforcement action in a state and year.

The timing of state authorization is distributed fairly evenly throughout our sample period, with the exception that 27 states received authorization prior to the sample period, between 1973-1975.[16] Given that neither the data nor the parallel trends assumptions for $Y_t(0)$ provide information to identify the average treatment effect for these "always treated" states, these states are dropped from the analysis. Figure 6.1 highlights the year that each of the remaining 23 states started treatment, i.e., the year in which the state was authorized to administer individual NPDES permits. The bottom four states are what we call the "never-treated" units, i.e., the states that remain unauthorized to administer individual NPDES permits through the entire sample period.[17] Figure 6.1 also allows one to visualize which states form each treatment group (those states whose colors turn to dark blue in the same year), and who the "not-yet-treated" states are at any point in time (those units that are colored light-blue in a given year).

Finally, we follow Grooms (2015) in defining states with above median federal public corruption convictions across all years as "corrupt" states.[18] Figure 6.2 shows corrupt states

---

[16] Figure D.1 in the Web Appendix D shows the distribution of the timing of state authorization across years. As many states receive authorization for the first four phases in the same year, we define the year of authorization as the year in which the state was authorized to perform the first phase of the program, administering individual NPDES permits.

[17] As of 2008, four states remained unauthorized to administer individual NPDES. Idaho received authorization in 2018, outside of the sample period used here to be consistent with Grooms (2015).

[18] See Web Appendix D for additional detail.

Figure 6.1: Timing of State Authorization
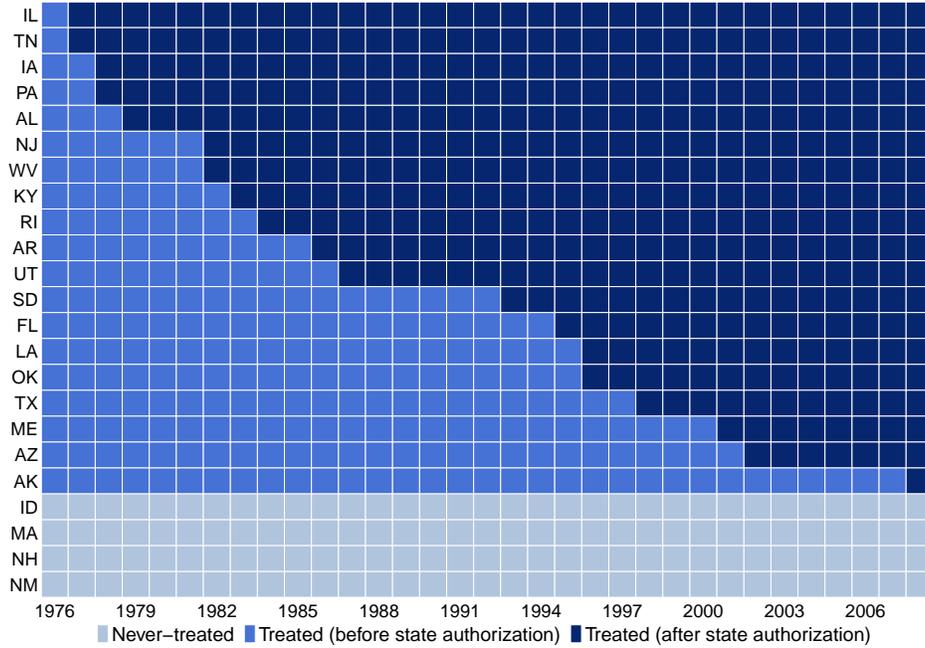
Notes: Shows the timing treatment adoption, where treatment is defined as the year in which the state was authorized to administer individual NPDES permits.

in red and non-corrupt states in blue. "Always-treated" states are shown in grey. Based on this measure, "corrupt" states are mostly from the mid-Atlantic and southern regions, while "non-corrupt" states tend to appear in New England and the west.

6.2   Specifications and assumptions

Like Grooms (2015), the starting point of our exercise is to examine the impact of authorization on compliance outcomes — for the sake of brevity, we focus on violation rates, though results for inspection rate and enforcement rate are available upon request.[19]

Since we are particularly interested in treatment effect dynamics, we estimate event-study-type parameters using four different procedures. First, we replicate the dynamic TWFE specification from Grooms (2015). The exact specification we use is the following:

$$Y_{it} = \lambda_i + \lambda_t + \sum_{e=-30, e \neq 0}^{32} \beta_e 1\{t - G_i + 1 = e\} + v_{it}, \tag{6.1}$$

which includes 30 treatment lead indicators (all the indicators associated with β_e with $e < 0$) and 32 treatment lag indicators (all the indicators associated with β_e with $e > 0$). We follow Borusyak and Jaravel (2017) and omit the treatment lead indicators associated with $e = 0$

---

[19]Overall, we find essentially zero effect on these other outcomes, regardless of the PTA and model specification used. This is in line with the results in Grooms (2015).

Figure 6.2: Corrupt and non-corrupt states
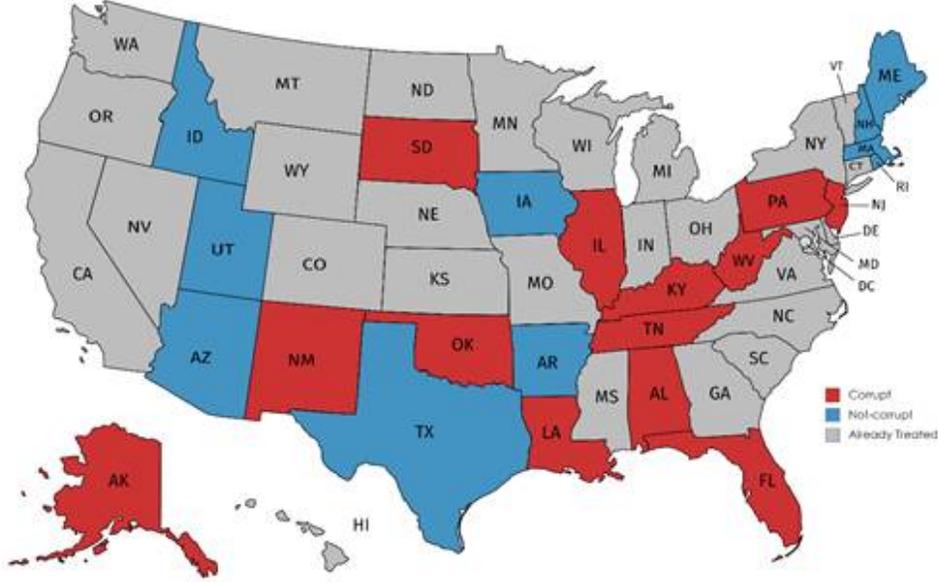
Notes: Corrupt states, shown in red, are those above the median of average convictions per capita across all years. Non-corrupt states are shown in blue. Grey states are "already treated" prior to the sample window and are not included in the analysis.

and with $e = -31$. Like Grooms (2015), our specifications are weighted by total facilities in a state, and all standard errors are clustered at the state level.

Second, we make specific PTAs and use the new estimators described previously. Because our empirical application includes a set of "never-treated" states, we estimate event-study-type parameters based on the PTA (2.6) and use $\widehat{\delta}^{es}_{never}(e)$ as an estimator for $\delta^{es}(e)$. We also leverage the PTA (2.7) and use $\widehat{\delta}^{es}_{ny}(e)$ as an estimator for $\delta^{es}(e)$. Finally, we employ the PTA (5.1) and use $\widehat{\delta}^{es}_{ny+}(e)$ as an estimator for $\delta^{es}(e)$. We do not use the event-study estimates based on GMM framework discussed in Section 4, since, in our specific application, the GMM associated with the PTA 2.5 involves 780 moments with *195 overidentification restrictions*, whereas sample size (state-year pairs) is equal to 759.

Next, we analyze whether the effect of state authorization on violation rates vary depending on whether a state has a long prevalence of corruption. To do so, we follow Grooms (2015) and consider the following TWFE specification

$$Y_{it} = \alpha_i + \alpha_t + \sum_{e=-30, e\neq 0}^{32} \beta_e 1\{t - G_i + 1 = e\} + \sum_{e=-30, e\neq 0}^{32} \beta_e^c (1\{t - G_i + 1 = e\} \times Corrupt_i) + v_{it},$$

(6.2)

where the β_e^c's are considered to be a measure of how treatment effects vary depending on whether a state is "corrupt" or not: positive (negative) point estimates suggest that the violation rates increased (decreased) more in corrupt states than in non-corrupt states.

At this stage, two important questions arise. First, what type of parallel trends assumption is actually being invoked to justify attaching a causal interpretation to the β_e^c's in (6.2)? Second, is (6.2) susceptible to the potential pitfalls discussed in Section 2.3.1? Answering these questions is inherently hard, as TWFE is a model specification and not a "research design." An alternative, and perhaps more constructive way of approaching this problem is to construct event-study-type estimators that explicitly rely on a particular PTA, and that, by design, avoid the potential lack of a clear interpretation associated with the TWFE specification (6.2). We follow this latter path.

With respect to the PTA, there are two natural variants of each of the PTAs 2.6, 2.7, and 5.1 that one can invoke to highlight treatment effect heterogeneity with respect to whether a state is corrupt or not. These variants differ from each other depending on whether or not one allows for different counterfactual trends between corrupt and non-corrupt states. One may be concerned, for example, that, in the absence of treatment, the evolution of the violation rate could differ between corrupt and non-corrupt states. In this case, one would prefer a "weaker" assumption to allow for corruption-specific trends. In the context of our application, "corruption" is not randomly assigned and the geographic clustering of corrupt and non-corrupt states may lead us to prefer a PTA that permits corruption-specific trends if we think, for example, there may be regional trends that differ across these states. We believe this is the most natural identification set up, as this is how one would proceed if one were to separately estimate counterfactuals for corrupt states and non-corrupt states, and only later compare their difference. We formalize these six different PTAs below.

Assumption 6.1 (Parallel trends assumption based on "never treated" units, with corruption-specific trends). *For $c = 0, 1$, and all $g, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| G_g = 1, Corr = c\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| C = 1, Corr = c\right].$$

Assumption 6.2 (Parallel trends assumption based on "not-yet treated" units, with corruption-specific trends). *For $c = 0, 1$, and all $g, s, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$, $s \geq t$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| G_g = 1, Corr = c\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| D_s = 0, Corr = c\right].$$

Assumption 6.3 ("Weaker" Parallel trends assumption based on "not-yet treated" units, with corruption-specific trends). *For $c = 0, 1$, and all $g, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| G_g = 1, Corr = c\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| D_t = 0, Corr = c\right].$$

Assumption 6.4 (Parallel trends assumption based on "never treated" units, without corruption-specific trends). *For $c = 0, 1$, and all $g, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| G_g = 1, Corr = c\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| C = 1\right].$$

Assumption 6.5 (Parallel trends assumption based on "not-yet treated" units, without corruption-specific trends). *For $c = 0, 1$, and all $g, s, t = 2, \dots, \mathcal{T}$, such that $t \geq g$, $s \geq t$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0)\middle| G_g = 1, Corr = c\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0)\middle| D_s = 0\right].$$

Assumption 6.6 ("Weaker" Parallel trends assumption based on "not-yet treated" units, without corruption-specific trends). *For $c = 0, 1$, and all $g, t = 2, \dots, \mathcal{T}$, such that $t \geq g$,*

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0)\middle| G_g = 1, Corr = c\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0)\middle| D_t = 0\right].$$

The difference between these PTAs depends on whether one uses the "never-treated", some "not-yet-treated", or all "not-yet-treated" as valid comparisons group, and whether one only uses states with the same corruption status (corrupt or non-corrupt) as valid comparison groups. Assumptions 6.1-6.3) do not assume that the evolution of the violation rate is the same between corrupt and non-corrupt states. These three assumptions are the analogues of Assumptions 2.6, 2.7 and 5.1 when one restricts attention to the subset of units with corruption status equal to c. Assumptions 6.4-6.6, on the other hand, assume that, in the absence of treatment, the evolution of the violation rate is the same for corrupt and non-corrupt states, i.e., it rules out corruption-specific trends. As so, one may argue that the Assumptions 6.1-6.3 are "weaker" than Assumptions 6.4-6.6.

Next, one can easily leverage any of these PTAs to identify and estimate sensible treatment effect parameters by following the same steps described in Section 2.3.2. The first step toward this goal is to show that the $ATT(g,t)$'s for the units with corruption status equal to c, $c = 0, 1$, defined by

$$ATT(g, t; c) \equiv \mathbb{E}\left[Y_{it}(1) - Y_{it}(0)\middle| G_g = 1, Corr = c\right],$$

are nonparametrically point-identified for all $t \geq g$. However, given the results in Theorem 1 of C&S and the discussion in Sections 2.3.2 and 5, this is a straightforward task. Indeed, one can easily show that for all $t \geq g$, the $ATT(g,t)$'s are nonparametrically identified by

$$
\begin{aligned}
ATT^1_{never}(g, t; c) &= \mathbb{E}\left[Y_t - Y_{g-1}\middle| G_g = 1, Corr = c\right] - \mathbb{E}\left[Y_t - Y_{g-1}\middle| C = 1, Corr = c\right], &(6.3)\\
ATT^1_{ny}(g, t; c) &= \mathbb{E}\left[Y_t - Y_{g-1}\middle| G_g = 1, Corr = c\right] - \mathbb{E}\left[Y_t - Y_{g-1}\middle| D_t = 0, Corr = c\right] &(6.4)\\
ATT^1_{ny+}(g, t; c) &= \mathbb{E}\left[Y_t - Y_{g-1}\middle| G_g = 1, Corr = c\right] - \left(\sum_{s=g}^{t}\mathbb{E}\left[\Delta Y_s\middle| D_s = 0, Corr = c\right]\right) &(6.5)\\
ATT^2_{never}(g, t; c) &= \mathbb{E}\left[Y_t - Y_{g-1}\middle| G_g = 1, Corr = c\right] - \mathbb{E}\left[Y_t - Y_{g-1}\middle| C = 1\right], &(6.6)\\
ATT^2_{ny}(g, t; c) &= \mathbb{E}\left[Y_t - Y_{g-1}\middle| G_g = 1, Corr = c\right] - \mathbb{E}\left[Y_t - Y_{g-1}\middle| D_t = 0\right], &(6.7)\\
ATT^2_{ny+}(g, t; c) &= \mathbb{E}\left[Y_t - Y_{g-1}\middle| G_g = 1, Corr = c\right] - \left(\sum_{s=g}^{t}\mathbb{E}\left[\Delta Y_s\middle| D_s = 0\right]\right), &(6.8)
\end{aligned}
$$

depending on whether one respectively imposes one of the PTA 6.1-6.6.[20] These results are analogous to (2.25), (2.26) and (5.1). Likewise, all the aforementioned quantities can be estimated using the analogy principle, i.e., by replacing population expectation with sample expectations.

Armed with these estimators, one can form different summary measures for the overall treatment effect following the same steps described in Section 2.3.2. To explicitly show how one can form event-study-type estimators, let $ATT_{generic}(g,t;c)$ be a generic notation for $ATT^1_{never}(g,t;c)$, $ATT^1_{ny}(g,t;c)$, $ATT^1_{ny+}(g,t;c)$, $ATT^2_{never}(g,t;c)$, $ATT^2_{ny}(g,t;c)$, and $ATT^2_{ny+}(g,t;c)$ and denote its plug-in estimator by $\widehat{ATT}_{generic}(g,t;c)$. Then, one can estimate the average treatment effect for units with corruption status equal to c that have been treated for e periods by

$$\widehat{\delta}^{es}_{generic}(e;c) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\left\{t-g+1=e\right\} \widehat{w}(g;e,c)\,\widehat{ATT}_{generic}(g,t;c), \qquad (6.9)$$

where the weights are given by

$$\widehat{w}(g;e,c) \equiv \widehat{P}\left(G_g=1|\text{Treated for} \geq \text{e periods, Corr} = c\right) = \frac{N_{g\cap \geq e\cap c}}{N_{\geq e\cap c}},$$

$N_{g\cap \geq e\cap c}$ denotes the number of observations in group g among those units with corrupt status c that have been treated for at least e periods, and $N_{\geq e}$ is the number of units with corrupt status c who have been treated for at least e periods.[21] Given that our main goal is to compare the evolution of treatment effects between corrupt and non-corrupt states, we can simply compute the difference between $\widehat{\delta}^{es}_{generic}(e;1)$ and $\widehat{\delta}^{es}_{generic}(e;0)$. Denote this (generic) estimator by $\widehat{\delta}^{es}_{generic}(e;1-0)$, i.e.,

$$\widehat{\delta}^{es}_{generic}(e;1-0) = \widehat{\delta}^{es}_{generic}(e;1) - \widehat{\delta}^{es}_{generic}(e;0). \qquad (6.10)$$

Here, we stress that, regardless of which of the six different estimators for (6.10) one adopts, they are all directly and explicitly tied to a given PTA, and, by design, they bypass the potential pitfalls associated with the TWFE specification.

In addition to the event-study estimates, we further aggregate these treatment effect curves into scalar, easy to interpret parameters. Toward this end, we report the plug-in estimators for the following two aggregated treatment effect parameters proposed by C&S,

$$ATT^{simple,1-0} = ATT^{simple;1} - ATT^{simple;0}, \qquad (6.11)$$

---

[20] Although the PTAs 6.4-6.6 lead to overidentification, for the sake of simplicity we do not fully exploit all these restrictions when proposing the aforementioned estimands.

[21] When $e < 0$, we replace $\widehat{w}(g;e,c)$ in (6.9) with $\widehat{w}(g;e-,c)$, where $\widehat{w}(g;e-,c)$ is analogous to $\widehat{w}(g;e-)$ as defined in Remark 2.1.

where, for $c = 0, 1$ $ATT^{simple;c}$ is defined analogously to (2.22), i.e.,

$$ATT^{simple,c} = \frac{\sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} P(G = g, Corr = c) \cdot ATT(g,t;c)}{\sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} P(G = g, Corr = c)},$$

and the average of $\delta^{es}(e; 1 - 0)$ over all possible (positive) values of e,

$$\delta^{e,avg,1-0} = \delta^{e,avg;1} - \delta^{e,avg;0}. \tag{6.12}$$

where, for $c = 0, 1$, $\delta^{e,avg;c}$ is defined analogously to (2.24), i.e.,

$$\delta^{e,avg;c} = \frac{1}{\mathcal{T} - 1} \sum_{e=1}^{\mathcal{T}-1} \delta^{es}(e; c).$$

We report estimators for these functionals that rely on the PTA 6.1-6.6, respectively. For the sake of comparison, we also report the OLS estimate of β_{fe}^{c} associated with the following TWFE specification,

$$Y_{it} = \alpha_g + \alpha_t + \beta_{fe} D_{it} + \beta_{fe}^{c} D_{it} \times Corrupt_i + u_{it}, \tag{6.13}$$

though these estimates are also subject to pitfalls briefly described in Section 2.3.1.

6.3   Results

6.3.1   Baseline results

Figure 6.3 displays the results based on the TWFE specification (6.1), and those based on the event-study estimators $\widehat{\delta}_{never}^{es}(e)$, $\widehat{\delta}_{ny}^{es}(e)$, and $\widehat{\delta}_{ny+}^{es}(e)$. We report the point-estimates associated with 20 treatment leads and 20 treatment lags (red line), their associated 90% point-wise confidence intervals (dark-shaded area), and 90% *simultaneous* confidence intervals (light-shaded area) — we do not report simultaneous confidence intervals for the TWFE specification as these are usually not reported by practitioners who adopt such specifications. It is important to emphasize that, in each of the panels in Figure 6.3, we have 40 different estimates, one for each considered e. Point-wise inference procedures proceed "as if" one were conducting a single hypothesis test, and report standard confidence interval for each e. Failing to account for the fact that one is performing 40 *different* hypotheses tests may lead to significant treatment effects and/or pre-trends that emerge simply by chance. Simultaneous confidence intervals, on the other hand, account for this multiple testing problem, and asymptotically cover the entire event-study curve with probability 1 - α, where α is the significance level. As so, simultaneous confidence intervals are suitable to analyze global properties of the event-study curve, such as monotonicity and presence of statistically non-zero effects. In practice, one simply has to replace the commonly used critical value (say, 1.645 for a 90% confidence interval) with the one simulated via a bootstrap procedure akin
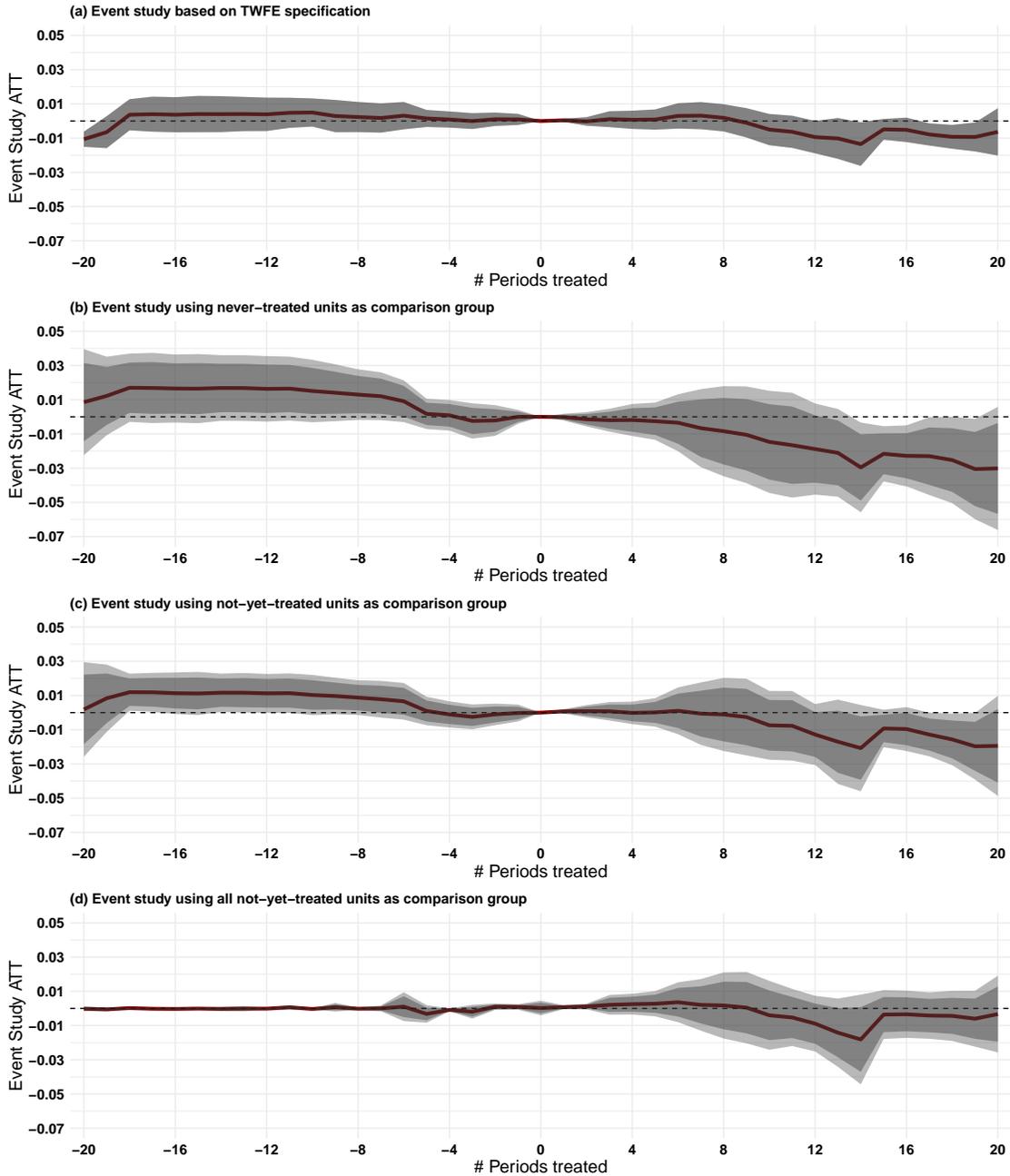
32

to Algorithm B.1; see Section 4 of C&S for additional details.

The results shown in Figure 6.3 suggest that, regardless of the PTA and the estimator used, there is little to no evidence that the transition to state control decreased violation rates. Despite the similarity in terms of conclusions, we find that comparing the results from each specification highlights some interesting practical features. For instance, the point estimates associated with the TWFE specification (Panel (a)) and with the estimator that uses the "all not-yet-treated" states as a comparison group (Panel (d)) are close to each other, whereas using "never-treated" states as a comparison group (Panel (b)) suggests a slightly stronger long-run effect. Furthermore, when using "all not-yet-treated" states as a comparison group (Panel (d)), the (simultaneous) confidence interval is tighter, suggesting, as we discussed in Section 5, that it makes more efficient use of the available data. In terms of interpreting the pre-treatment coefficients, the pre-trend point estimates when using the "not-yet-treated" comparison group in Panel (c) are closer to zero than when one uses the "never-treated" states as a comparison group in Panel (b). It is also very noticeable that pre-treatment trends in Panel (d) are very precisely estimated zeros. However, it is important to recall from Remark 5.1 that these pre-treatment coefficients should not be directly compared to the other pre-treatment trends as they measure "local deviations" of zero pre-treatment trends rather "cumulative deviations" of pre-treatment trends.

Although these estimators lead to similar conclusions, they are not (a priori) "made equal". As highlighted by S&A and discussed in Section 2.3.1, the β_e's associated with the TWFE specification (6.1) are not guaranteed to have a clear causal interpretation, *even when one invokes the PTA 2.5*, which, in our application, imposes 195 overidentifying restrictions on the evolution of violation rates across states. The estimators in Panels (b), (c), and (d), on the other hand, are designed to bypass the potential pitfalls of the TWFE specification, and rely on clearly stated parallel trends assumptions (Assumptions 2.6, 2.7, and 5.1, respectively).

As discussed in Section 2.3.2, there are multiple sensible measures that one can use to summarize the overall effect of state authorization on violation rates across all treated states. For instance, one can use the "simple" average of all the $ATT(g,t)$ where $t \geq g$, ATT^{simple}, as defined in (2.22), or the average of the event-study-type estimands $\delta^{es}(e)$ over the positive values of e, $\delta^{e,avg}$, as defined in 2.24. Table 6.1 shows the estimates of these parameters when one adopts Assumption 2.6 (Column (1)), Assumption 2.7 (Column (2)), or Assumption 5.1 (Column (3)). For the sake of comparison we also report the OLS estimate of β_{fe} (Column (4)). Standard errors, clustered at the state level, are reported in parenthesis and 90% confidence intervals are reported in brackets. Essentially, all these summary measures indicates that state authorization has close to zero effect on violation

Figure 6.3: Event-study analysis of violation rate: baseline results

(a) Event study based on TWFE specification

(b) Event study using never–treated units as comparison group

(c) Event study using not–yet–treated units as comparison group

(d) Event study using all not–yet–treated units as comparison group

Notes: Red line displays the point estimate, dark-shaded area the 90% pointwise confidence interval, and the light-shaded are the 90% simultaneous confidence band. Panel (a) displays the ordinary least squares (OLS) estimates of the β_e associated with the two-way fixed-effects linear regression specification (6.1); Panel (b) displays the results based on (2.23) that uses (2.27) as an estimator for $ATT(g,t)$; Panel (c) displays the results based on (2.23) that uses (2.19) as an estimator for $ATT(g,t)$; Panel (d) displays the results based on (2.23) that uses (5.5) as an estimator for $ATT(g,t)$. All standard errors are clustered at the state level, though the standard errors in Panel (a) are based on analytical results, whereas those in Panel (b)-(d) are based on the multiplicative bootstrap procedure discussed in Algorithm B.1) and in C&S (we use 1,000 bootstrap draws). The critical value for the simultaneous confidence bands is computed using Algorithm B.1 (which is akin to the one proposed by C&S).

Table 6.1: Effect of authorization on violation rate: Baseline specifications.

Summary measures	Never-treated (1)	Not-yet-treated (2)	All Not-yet-treated (3)	TWFE (4)
ATT^{simple}	-0.017	-0.010	-0.003	—
	(0.009)	(0.009)	(0.006)	—
	[-0.032, 0.001]	[-0.024, 0.004]	[-0.014, 0.008]	—
$\delta^{e,avg}$	-0.015	-0.008	-0.003	—
	(0.007)	(0.006)	(0.004)	—
	[-0.027, -0.002]	[-0.017, 0.002]	[-0.010, 0.004]	—
TWFE	—	—	—	-0.003
	—	—	—	(0.010)
	—	—	—	[-0.019, 0.013]

Notes: The point estimates, cluster-robust standard errors (in parenthesis), and 90% confidence interval (in brackets) for the effect of state authorization on violation rates. ATT^{simple} is as defined in (2.22) and denotes the weighted average of all post-treatment $ATT(g,t)'s$. $\delta^{e,avg}$ is as defined in (2.24) and denotes the time-average of all event-study parameters $\delta^{es}(e)$, $e > 0$. TWFE refers to the ordinary least square estimates of β_{fe} in the TWFE linear regression specification (2.16), which is invariant to the comparison group being used. Column (1) display the results that uses (2.27) as an estimator for $ATT(g,t)$, column (2) displays the results that uses (2.19) as an estimator for $ATT(g,t)$, and column (3) displays the results that uses (5.5) as an estimator for $ATT(g,t)$ Column (4) displays the result using the TWFE regression specification. Standard errors are clustered at the state level, and, with the exception of the TWFE summary measure, are computed using the multiplicative bootstrap procedure described in Algorithm B.1, which is akin to the one proposed by C&S. We use 1,000 bootstrap draws.

rates, which is in line with the findings from Grooms (2015).
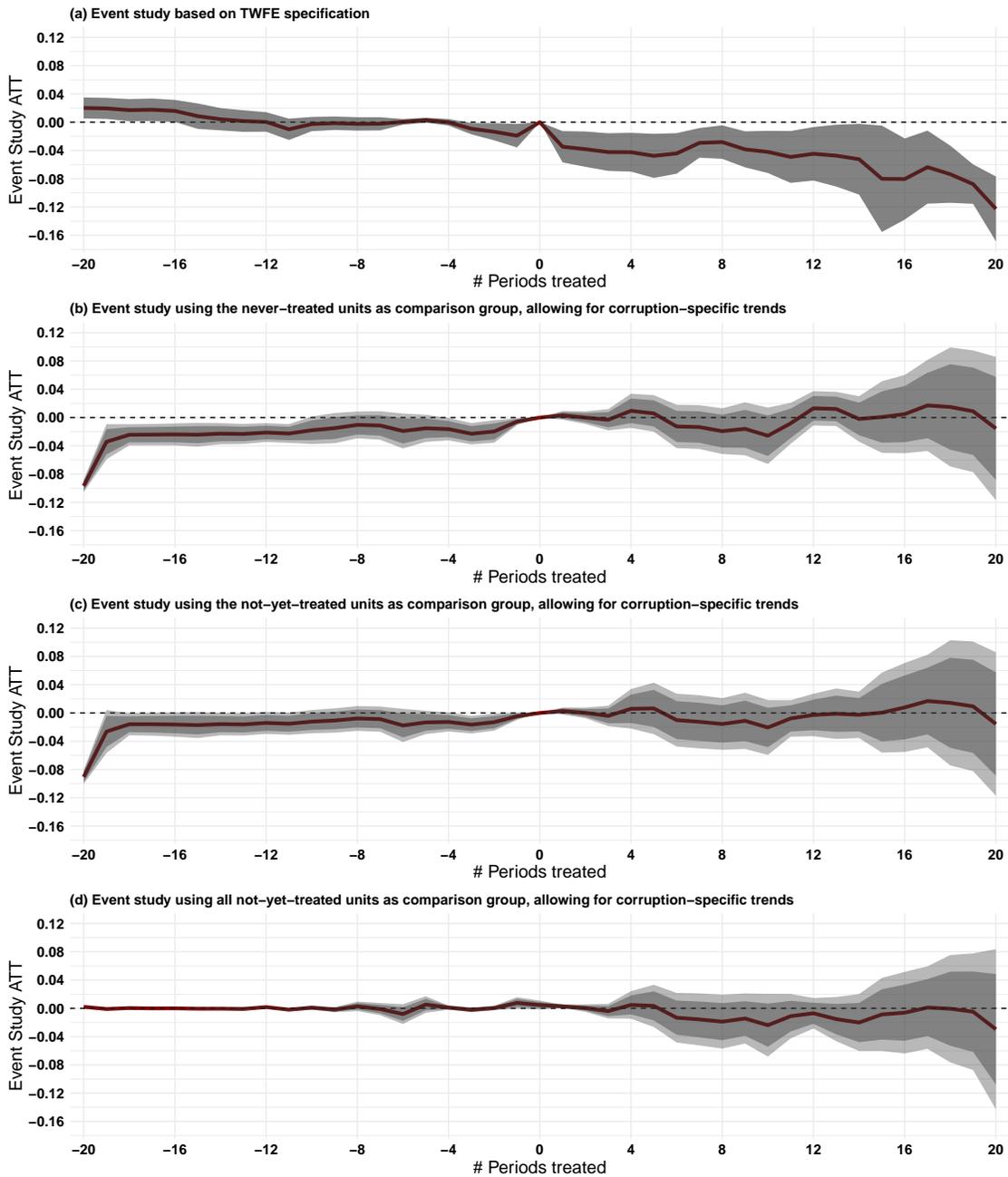
6.3.2 Corrupt vs. Non-corrupt Results

Next, we analyze whether the effect of state authorization on violation rates vary depending on whether a state has a long prevalence of corruption. Panel (a) of Figure 6.4 displays the OLS estimates of β_e^c's, together with the 90% pointwise confidence intervals. All standard errors are clustered at the state level. Consistent with the findings from Grooms (2015), the results suggest that states with high levels of corruption have a lower violation rate after authorization relative to non-corrupt states, and the relative drop in the violation rate appears to increase with elapsed treatment time.

Panels (b), (c), and (d) of Figure 6.4 present the event-study estimates (6.10) based on the PTAs 6.1, 6.2, and 6.3 that allow for corruption-specific trends, whereas Panels (b), (c), and (d) of Figure 6.5 present the event-study estimates (6.10) based on the PTAs 6.4, 6.5, and 6.6 that do not allow for corruption-specific trends. For comparison purposes, Panel (a) of Figures 6.4 and 6.5 displays the OLS estimates of the β_e^c's associated with the TWFE specification (6.2). Like before, all estimators are weighted by total facilities in a state, all standard errors are clustered at the state level, and we report both pointwise and simultaneous 90% confidence intervals.
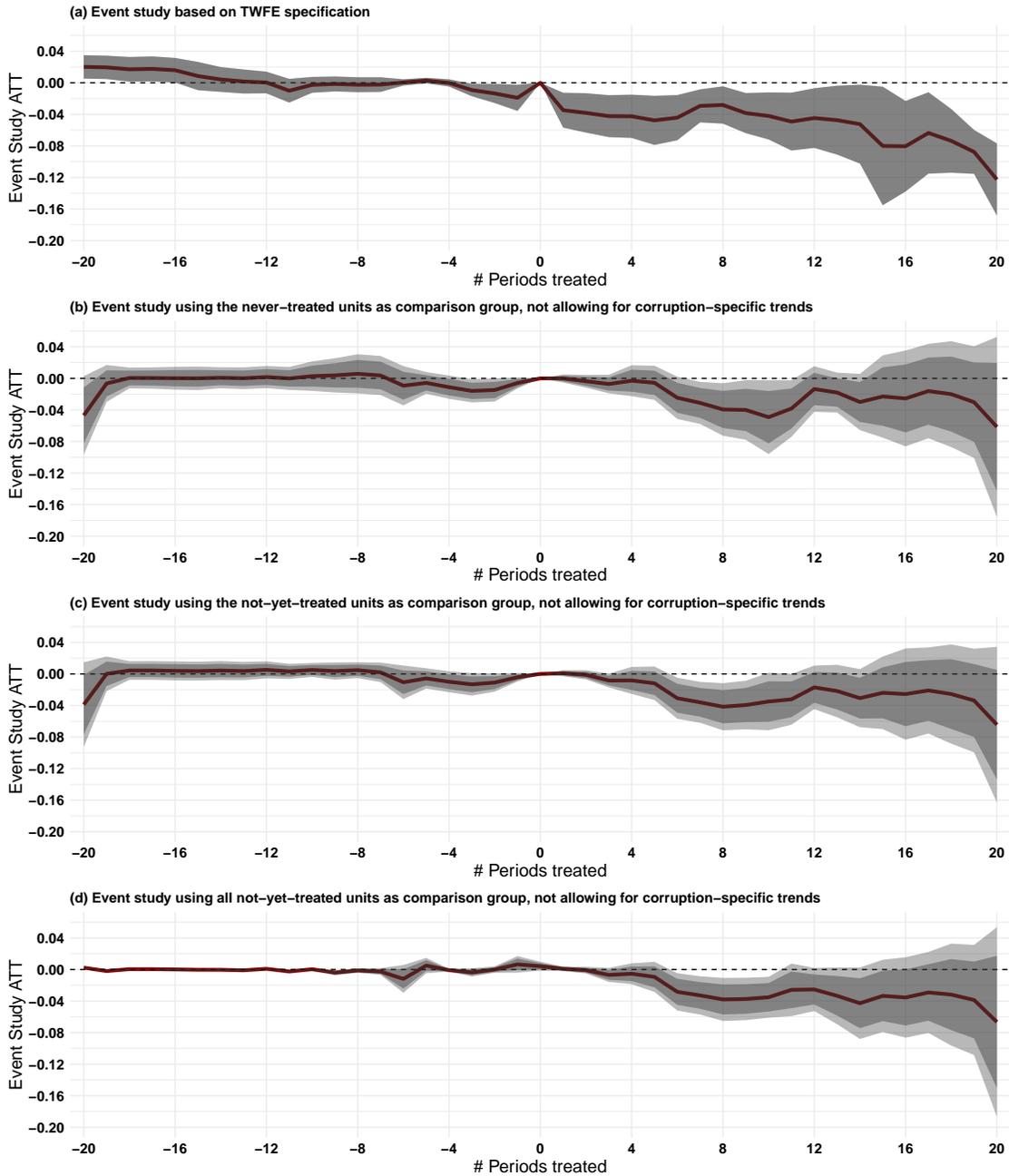
The results in Figures 6.4 and 6.5 reveal the practical relevance of being explicit about the underlying PTA in a given application. For instance, Figure 6.4 suggests that when one invokes Assumption 6.1 (Panel (b)), Assumption 6.2 (Panel (c)), or Assumption 6.3 (Panel

Figure 6.4: Event-study analysis of violation rate: difference between corrupt and non-corrupt states allowing different counterfactual trends between corrupt and non-corrupt states

(a) Event study based on TWFE specification

(b) Event study using the never-treated units as comparison group, allowing for corruption-specific trends

(c) Event study using the not-yet-treated units as comparison group, allowing for corruption-specific trends

(d) Event study using all not-yet-treated units as comparison group, allowing for corruption-specific trends

Notes: Red line displays the point estimate, dark-shaded area the 90% pointwise confidence interval, and the light-shaded are the 90% simultaneous confidence band. Panel (a) displays the results based on the OLS estimates of the β_e^c's in the TWFE specification (6.2); Panel (b) displays the results based on the event-study estimator (6.9) that relies on the PTA 6.1; Panel (c) displays the results based on the event-study estimator (6.9) that relies on the PTA 6.2; Panel (d) displays the results based on the event-study estimator (6.9) that relies on the PTA 6.3. All standard errors are clustered at the state level, though the standard errors in Panel (a) are based on analytical results, whereas those in Panel (b)-(d) are based on the multiplicative bootstrap procedure discussed in Algorithm B.1, which is similar to C&S proposal (we use 1,000 bootstrap draws). The critical value for the simultaneous confidence bands is computed using Algorithm B.1.

Figure 6.5: Event-study analysis of violation rate: difference between corrupt and non-corrupt states assuming same counterfactual trends for corrupt and non-corrupt states

(a) Event study based on TWFE specification

(b) Event study using the never–treated units as comparison group, not allowing for corruption–specific trends

(c) Event study using the not–yet–treated units as comparison group, not allowing for corruption–specific trends

(d) Event study using all not–yet–treated units as comparison group, not allowing for corruption–specific trends

Notes: Red line displays the point estimate, dark-shaded area the 90% pointwise confidence interval, and the light-shaded are the 90% simultaneous confidence band. Panel (a) displays the results based on the OLS estimates of the β_e^c's in the TWFE specification (6.2); Panel (b) displays the results based on the event-study estimator (6.9) that relies on the PTA 6.4; Panel (c) displays the results based on the event-study estimator (6.9) that relies on the PTA 6.5; Panel (d) displays the results based on the event-study estimator (6.9) that relies on the PTA 6.6. All standard errors are clustered at the state level, though the standard errors in Panel (a) are based on analytical results, whereas those in Panel (b)-(d) are based on the multiplicative bootstrap procedure discussed in Algorithm B.1, which is similar to C&S proposal (we use 1,000 bootstrap draws). The critical value for the simultaneous confidence bands is computed using Algorithm B.1.

(d)) and allows corruption-specific counterfactual trends, one finds that state authorization affected violation rates in a similar manner, regardless of whether a state is corrupt or not. Such results are in sharp contrast with those associated with the TWFE specification (6.2) in Panel (a).

On the other hand, the results in Figure 6.5 suggest that when one invokes Assumption 6.4 (Panel (b)), Assumption 6.5 (Panel (c)), or Assumption 6.6 (Panel (d)) and therefore rules out differential potential trends between corrupt and non-corrupt states, one finds that corrupt states experience a large decrease in violation rates after state authorization relative to non-corrupt states. Although such results are in line with those based on TWFE in Panel (a), we note that the results based on (6.10) suggests milder treatment effect differences between corrupt and non-corrupt states.

Next, we aggregate these treatment effect curves into a scalar, easy to interpret parameter. Table 6.2 reports the plug-in estimators for the two aggregated treatment effect parameters, shown in (6.11) and (6.12). For comparison, we report in Column (7) the OLS estimate of β_{fe}^c associated with the TWFE specification, shown in (6.13).

Table 6.2: Effect of authorization on violation rate: corrupt vs. not corrupt states.

Summary measures	Allow for corrupt-specific trends			Not allow corrupt-specific trend			TWFE
	Never-treated (1)	Not-yet-treated (2)	All Not-yet-treated (3)	Never-treated (4)	Not-yet-treated (5)	All Not-yet-treated (6)	(7)
$ATT^{simple,1-0}$	-0.007	-0.008	-0.014	-0.035	-0.035	-0.033	—
	(0.014)	(0.014)	(0.013)	(0.012)	(0.012)	(0.010)	—
	[-0.030, 0.017]	[-0.031, 0.016]	[-0.036, 0.008]	[-0.054, -0.016]	[-0.054, -0.015]	[-0.049, -0.017]	—
$\delta^{e,avg,1-0}$	-0.001	-0.002	-0.009	-0.024	-0.025	-0.028	—
	(0.014)	(0.016)	(0.014)	(0.013)	(0.012)	(0.011)	—
	[-0.024, 0.022]	[-0.028, 0.024]	[-0.031, 0.013]	[-0.045, -0.003]	[-0.045, -0.005]	[-0.047, -0.009]	—
TWFE	—	—	—	—	—	—	-0.037
	—	—	—	—	—	—	(0.010)
	—	—	—	—	—	—	[-0.054, -0.020]

Notes: The point estimates, cluster-robust standard errors (in parenthesis), and 90% confidence interval (in brackets) for the effect of state authorization on violation rates. $ATT^{simple,2-0}$ is as defined in (6.2) and denotes the difference of the weighted average of all post-treatment $ATT(g,t;c)'s$ between corrupt and non-corrupt states. $\delta^{e,avg,1-0}$ is as defined in (6.12) and denotes difference of the time-average of all event-study parameters $\delta^{es}(e)$, $e > 0$, between corrupt and non-corrupt states. TWFE refers to the ordinary least square estimates of β_{fe}^c in the TWFE linear regression specification (6.13), which is invariant to the comparison group being used. Columns (1)-(6) display the results that relies on the PTA 6.1-6.6, respectively. Standard errors are clustered at the state level, and, with the exception of the TWFE summary measure, are computed using the multiplicative bootstrap procedure presented in Algorithm B.1, which is akin to C&S proposal. We use 1,000 bootstrap draws.

The results in Table 6.2 reinforce the message from Figures 6.4 and 6.5: when one allows for corruption-specific trends and relies on PTAs 6.1, 6.2, or 6.3 (Columns (1), (2), and (3), respectively), one finds essentially no evidence that the effect of state authorization on violation rates varies by state corruption. On the other hand, when one relies on the "stronger" PTAs 6.4, 6.5, or 6.6 (Columns (4), (5), and (6), respectively), one finds evidence that corrupt states experienced a large decrease in violation rate after state authorization relative to non-corrupt states. This latter result is in agreement with the TWFE specification, whereas the former is not.

7 Conclusion

In this paper, we have highlighted the important role played by the parallel trends assumption in event-study settings in terms of identification, estimation and summary of different treatment effects parameters. We first showed that, when there is variation in treatment timing, researchers may adopt different types of parallel trends assumptions and identify/estimate different treatment effect parameters. Next, we discussed the practical implications of adopting different parallel trends assumptions, and discussed how one constructs estimators that make use of all the restrictions implied by the underlying PTA. Here, we documented an interesting "robustness" vs. "efficiency" trade-off in terms of the strength of the underlying PTA, and argue that one should take this into consideration whenever employing a DID-type of analysis. Importantly, we advocate that one should always attempt to be explicit about the parallel trends assumption invoked in the study, as this usually translates into a more transparent and objective analysis. We showed how one can form semiparametrically efficient DID estimators by fully exploiting all the empirical content of underlying PTA via the traditional GMM approach. We also proposed an alternative, simpler to use DID estimator that does not restrict pre-treatment trends when one wants to use "not-yet-treated" units as a comparison group, and, at the same time, makes use of more groups than other available DID estimators. Finally, we illustrated the practical importance of being explicit about the PTA via an empirical application about the effect of the transition from federal to state management of the Clean Water Act on compliance rates. Our results suggest that the conclusion that corrupt states see a decline in the violation rate after program authorization relative to non-corrupt treated states depends on the type of PTA adopted.

References

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

Anderson, Michael L., "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484).

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

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, "Synthetic Difference in Differences," *arXiv preprint arXiv:1812.09970*, 2018.

Athey, Susan and Guido W Imbens, "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," *arxiv preprint arXiv:1808.05293*, 2018.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," *Unpublished Manuscript, Department of Economics, Harvard University*, 2017.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *arXiv preprint arXiv:1803.09015*, 2020.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey, "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 2013, *81* (2).

Cunningham, Scott, *Causal Inference: The Mixtape, v.1.7* 2018.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9).

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *The Review of Economics and Statistics*, 2019, *101* (3).

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic, "Broken or Fixed Effects?," *Journal of Econometric Methods*, 2018, *8* (1).

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *NBER Working Paper No. 25018*, 2019.

Grooms, Katherine K, "Enforcing the Clean Water Act: The effect of state-level corruption on compliance," *Journal of Environmental Economics and Management*, 2015, *73*.

Han, Sukjin, "Identification in Nonparametric Models for Dynamic Treatment Effects," *Journal of Econometrics*, 2019, *Forthcoming*.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4).

Heckman, James J., Hidehiko Ichimura, Jefrey Smith, and Petra Todd, "Characterizing Selection Bias using Experimental Data," *Econometrica*, 1998, *66* (5).

Laporte, Audrey and Frank Windmeijer, "Estimation of panel data models with binary indicators when treatment effects are not constant over time," *Economics Letters*, 2005, *88* (3).

Lechner, Michael, "The Estimation of Causal Effects by Difference-in-Difference Methods," *Foundations and Trends in Econometrics*, 2010, *4* (3).

Rambachan, Ashesh and Jonathan Roth, "An Honest Approach to Parallel Trends," *Working Paper, Department of Economics, Harvard University*, 2019.

Romano, Joseph P. and Michael Wolf, "Stepwise multiple testing as formalized data snooping," *Econometrica*, 2005, *73* (4).

_ , **Azeem M. Shaikh, and Michael Wolf**, "Hypothesis Testing in Econometrics," *Annual Review of Economics*, 2010, *2* (1).

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *Working Paper, Department of Economics, Harvard University*, 2020.

Rubin, Donald B., "The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials," *Statistics in Medicine*, 2007, *26* (1).

_ , "For objective causal inference, design trumps analysis," *Annals of Applied Statistics*, 2008, *2* (3).

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

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects," *Working Paper, Department of Economics, MIT*, 2020.

Wooldridge, Jeffrey M., "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and Statistics*, 2005, *87* (2).