How to Improve the Difference-in-Differences Design with Multiple Pre-treatment Periods*

Naoki Egami† Soichiro Yamauchi‡

This version: December 10, 2019
First draft: December 6, 2019

Abstract

While difference-in-differences (DID) was originally developed with one pre- and one post-treatment periods, data from additional pre-treatment periods is often available. How can researchers improve the DID design with such multiple pre-treatment periods under what conditions? We first use potential outcomes to clarify three benefits of multiple pre-treatment periods: (1) assessing the parallel trends assumption, (2) improving estimation accuracy, and (3) allowing for a more flexible parallel trends assumption. We then propose a new estimator, *double* DID, which combines all the benefits through the generalized method of moments and contains the two-way fixed effects regression as a special case. In a wide range of applications where several pre-treatment periods are available, the double DID improves upon the standard DID both in terms of identification and estimation accuracy. We illustrate the proposed method in a study of how recentralization affects public services.

*The methods proposed in this article can be implemented via the open-source statistical software R package `DIDdesign`. We are grateful to Edmund Malesky, Cuong Viet Nguyen, and Anh Tran for providing us with data and answering our questions. We also thank Adam Glynn, Chad Hazlett, Shiro Kuriwaki, Ian Lundberg, John Marshall, Xiang Zhou, and participants of the 2019 Summer Meetings of the Political Methodology Society and the 2019 American Political Science Association Annual Conference for helpful comments and discussions.

†Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Pre-doctoral Fellow, Department of Government, Harvard University, Cambridge MA 02138. Email: negami@princeton.edu. URL: `http://scholar.princeton.edu/negami`

‡Graduate student, Department of Government, Harvard University, Cambridge MA 02138. Email: syamauchi@g.harvard.edu. URL: `https://soichiroy.github.io/`

1 Introduction

Over the last few decades, social scientists have developed and applied various approaches to make credible causal inferences from observational data. One of the most popular and successful is a difference-in-differences (DID) design (Bertrand et al., 2004; Angrist and Pischke, 2008). In its most basic form, we compare treatment and control groups over two time periods — one before and the other after the treatment assignment.

In practice, it is common to apply the DID method with additional pre-treatment periods. However, in contrast to the basic two-time-period case, there are a number of different ways to analyze the DID with multiple pre-treatment periods. One popular approach is to apply the two-way fixed effects regression to the entire time periods and supplement it with various robustness checks (e.g., Dube et al., 2013; Truex, 2014; Earle and Gehlbach, 2015; Hall, 2016; Larreguy and Marshall, 2017). Another is to stick with the two-time-period DID and limit the use of additional pre-treatment periods only to the assessment of pre-treatment trends (e.g., Ladd and Lenz, 2009; Bechtel and Hainmueller, 2011; Bullock and Clinton, 2011; Keele and Minozzi, 2013; Garfias, 2018). This variation of approaches raises an important practical question: how should analysts incorporate multiple pre-treatment periods into the DID design and under what assumptions? By addressing this point, we make several contributions to improve the DID design with multiple pre-treatment periods.

We first use potential outcomes (Neyman, 1923; Rubin, 1974) to clarify three well-known benefits of multiple pre-treatment periods. We focus on examining their required assumptions that are often unstated in practice. (1) *Assessing parallel trends*: Researchers are often advised to assess whether treatment and control groups have the same trends in pre-treatment periods (Angrist and Pischke, 2008). Although we echo this well known recommendation, we emphasize that this standard procedure does not evaluate the parallel trends assumption itself, which is untestable due to counterfactual outcomes. Rather, it tests the extended parallel trends assumption, which requires parallel trends over longer time periods. We discuss how to formally test this assumption to make the DID design more robust. (2) *Improving estimation accuracy*: When parallel trends also hold in pre-treatment periods, the DID design is not only

more credible but we can also use additional observations from pre-treatment periods for the DID estimation. This can increase estimation accuracy, and thus, reduce standard errors. We show how to achieve this efficient estimation, both nonparametrically and within the two-way fixed effects regression. (3) *Allowing for a more flexible parallel trends assumption*: When the parallel trends assumption is implausible, one strategy is to adjust for bias in the standard DID by taking into account non-parallel trends in pre-treatment periods (Mora and Reggio, 2012, 2019). We clarify that this estimator requires a generalization of parallel trends, called the parallel trends-in-trends assumption, that allows for linear time-varying unmeasured confounding that changes over time at some fixed rate.

Our main contribution is to propose a new, simple estimator that achieves all three benefits together. In its core, our proposed approach, *double difference-in-differences* (double DID), combines two DID estimators via the generalized method of moments (GMM) (Hansen, 1982). Within the GMM framework, the double DID has natural two steps: (1) assess underlying parallel trends assumptions, and (2) employ an unbiased, efficient DID estimator under the extended parallel trends or parallel trends-in-trends assumption. By internalizing these two steps, the double DID can achieve higher estimation accuracy than the standard DID when the trends of treatment and control groups are parallel. Even when trends are not parallel, it allows for the identification of causal effects as far as unmeasured confounders vary over time linearly. Importantly, the proposed double DID contains two-way fixed effects regression estimators as special cases and further improves in terms of identification and estimation accuracy via the GMM. We also generalize our methodologies to settings with any number of *pre-* and *post*-treatment periods. The proposed approach can be implemented in a companion R package.

This paper builds on the large literature of time-series cross-sectional data (e.g., De Boef and Keele, 2008; Beck and Katz, 2011; Blackwell and Glynn, 2018) and is connected to other popular methodologies for making causal inferences. Generalizing the well known case of two periods and two groups (e.g., Abadie, 2005), recent papers use potential outcomes to unpack the nonparametric connection between the DID and two-way fixed effects regression estimators, thereby proposing extensions to relax strong parametric and causal assumptions (e.g.,

Athey and Imbens, 2018; Goodman-Bacon, 2018; Strezhnev, 2018; Imai and Kim, 2019a,b). Our paper also uses potential outcomes to clarify nonparametric foundations on the use of multiple pre-treatment periods. The key difference is that, while this recent literature mainly considers the identification under the parallel trends assumption, we study both estimation accuracy and identification under more flexible assumptions of trends in a canonical DID setup where treatment assignment happens only once. Another class of popular methods is the synthetic control method (Abadie et al., 2010) and their recent extensions (e.g., Xu, 2017; Athey et al., 2017; Ben-Michael et al., 2018; Hazlett and Xu, 2018) that estimate a weighted average of control units to approximate a treated unit. As carefully noted in those papers, such methodologies require long pre-treatment periods to accurately estimate a pre-treatment trajectory of the treated unit (Abadie et al., 2010); for example, Xu (2017) recommends collecting more than ten pre-treatment periods. In contrast, our proposed double DID relies on the parallel trends or parallel trends-in-trends assumptions, and therefore, it is best applied in common DID settings where researchers have several pre-treatment periods.

This paper proceeds as follows. In Section 2, we start with the standard DID design with two time periods and clarify quantities of interest, assumptions, and estimators. Section 3 discusses the three benefits of using multiple pre-treatment periods. In Section 4, we propose the double DID that combines the three benefits within a single method. Section 5 presents simulation evidence of how the proposed method outperforms existing DID estimators by using pre-treatment periods more effectively. We provide an empirical illustration of the double DID in Section 6 based on Malesky et al. (2014), which studies how the abolition of elected councils affects local public services. Finally, we conclude by discussing the limitations and potential extensions of the proposed approach.

2 Difference-in-Differences with Two Time Periods

The difference-in-differences (DID) design is one of the most widely used methods to make causal inferences from observational studies (Imbens and Wooldridge, 2009). At their most basic, the DID design consists of treatment and control groups measured at two time periods, before and after the treatment assignment. While the DID design with two time periods

is well known in political science, we review it here to fix ideas for settings with multiple pre-treatment periods.

As our running example, we focus on a study of how the abolition of elected councils affects local public services (Malesky et al., 2014). This work uses the DID design to examine the effect of recentralization efforts in Vietnam. The abolition of elected councils, the main treatment of interest, was implemented in 2009 to about 12% of all the communes, the smallest administrative units that the paper considers. For each commune, a variety of outcomes, such as the quality of infrastructure, were measured both in 2008 and 2010, before and after the abolition of elected councils. With this DID design, Malesky et al. (2014) aim to estimate the causal effect of abolishing elected councils on various measures of local public services. To introduce the setup for the DID design, we focus only on the basic aspects of the study here and discuss further details when we reanalyze it in Section 6.

To begin with, let D_{it} be the treatment for unit i in time period t so that $D_{it} = 1$ if the unit is treated in time period t and $D_{it} = 0$ otherwise. For example, $D_{it} = 1$ would represent a commune that abolishes elected councils at time t. Under the basic DID design, we consider two time periods $t \in \{1, 2\}$ before and after the treatment implementation. Thus, a treatment group receives the treatment only at the second time period $D_{i1} = 0$ and $D_{i2} = 1$, whereas a control group never gets treated $D_{i1} = D_{i2} = 0$. We refer to the treatment group as $G_i = 1$ and the control group as $G_i = 0$. Outcome Y_{it} is measured both before and after the treatment $t \in \{1, 2\}$. In our running example, Y_{i1} and Y_{i2} are the qualities of infrastructure in a given commune in 2008 and 2010, before and after the abolition of elected councils in 2009. In addition to panel data where the same units are measured over time, the DID design accommodates repeated cross-sectional data as in our running example, in which different communes are sampled at two time periods.

To define causal effects of interest, we rely on the potential outcomes framework (Neyman, 1923; Rubin, 1974). For each time period, $Y_{it}(1)$ represents the quality of infrastructure that commune i would achieve in time period t if commune i had abolished elected councils. Similarly, $Y_{it}(0)$ represents the potential quality of infrastructure in time period t if commune i had kept elected councils. For an individual commune, the causal effect of abolishing elected

councils on the quality of infrastructure in time period t is the difference $Y_{it}(1) - Y_{it}(0)$. As the treatment is assigned in the second time period, causal effects are defined only for the second time period $Y_{i2}(1) - Y_{i2}(0)$.

In the DID design, because the average causal effect for all units is difficult to identify without strong assumptions, such as constant effects, we often focus on estimating the average treatment effect for treated units (ATT) (Angrist and Pischke, 2008):

$$\tau = \mathbb{E}[Y_{i2}(1) - Y_{i2}(0) \mid G_i = 1], \tag{1}$$

where the expectation is over units in the treatment group $G_i = 1$ so that this causal estimand is the average of individual causal effects for units that received the treatment. For example, this quantity represents the average effect of abolishing elected councils on the quality of infrastructure in 2010 for communes that abolished elected councils.

In the DID design with two time periods, we identify the ATT based on the widely-used assumption of *parallel trends* — if a treatment group had not received the treatment in the second period, its outcome trend would have been the same as the trend of control group's outcomes (Angrist and Pischke, 2008). Figure 1 visualizes the parallel trends. In terms of potential outcomes, the parallel trends assumption is:

Assumption 1 (Parallel Trends)

$$\mathbb{E}[Y_{i2}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1] \;=\; \mathbb{E}[Y_{i2}(0) \mid G_i = 0] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0], \tag{2}$$

where the left hand side is the trend in outcomes for the treatment group $G_i = 1$ and the right is the one for the control group $G_i = 0$.

Under the parallel trends assumption, we estimate the ATT via a difference-in-differences.

$$\widehat{\tau}_{\texttt{DID}} = \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right), \tag{3}$$

where n_{1t} and n_{0t} are the number of units in the treatment and control groups at time $t \in \{1, 2\}$, respectively. The first two terms represent the difference between the infrastructure qualities in 2010 and 2008 for the treatment group and the last two terms correspond to the same difference for the control group. Figure 1 graphically shows the DID estimator.

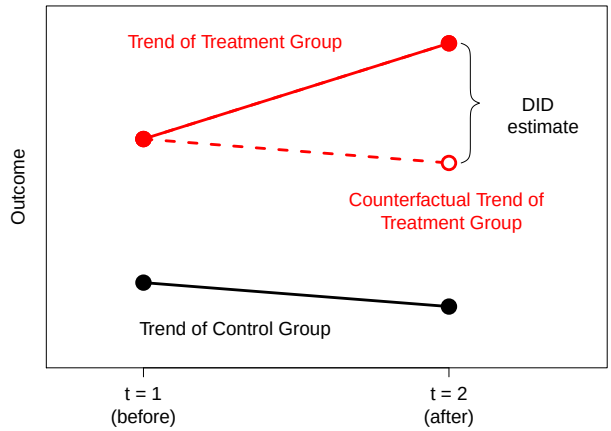

Figure 1: Parallel Trends Assumption and DID estimator. *Note:* The parallel trends assumption means that the trend of the control group (black solid line) is the same as the trend of the treatment group in the absence of the treatment (red dotted line). The difference in the trends of the treatment (red solid line) and control groups (black solid line) is the DID estimate.

In practice, we can compute the DID estimator via a linear regression in which we regress outcome Y_{it} on an intercept, treatment group indicator G_i, time indicator I_t (equal to 1 if post-treatment and 0 otherwise) and the interaction between the treatment group indicator and the time indicator $G_i \times I_t$.

$$Y_{it} \sim \alpha + \theta G_i + \gamma I_t + \beta(G_i \times I_t), \tag{4}$$

where $(\alpha, \theta, \gamma, \beta)$ are corresponding coefficients. In this case, a coefficient of the interaction term β is numerically equal to the DID estimator. Importantly, the linear regression is used here only to compute the nonparametric DID estimator (Equation (3)) and thus it does not require any parametric modeling assumption. Furthermore, when we analyze panel data in which the same units are observed repeatedly over time, we obtain exactly the same estimate via a linear regression with unit and time fixed effects. This numerical equivalence in the two-time-period case is often the justification of the two-way fixed effects regression as the DID design (Angrist and Pischke, 2008).

3 Three Benefits of Multiple Pre-treatment Periods

While the most basic DID design only requires data from one post- and one pre-treatment periods, additional pre-treatment periods are often available in applied contexts. For example,

Malesky et al. (2014) collected data on public services in 2006 as well as in 2008 and 2010. In fact, researchers have utilized multiple pre-treatment periods explicitly or implicitly to strengthen the DID design. Unfortunately, however, assumptions behind different uses of pre-treatment periods have often remained unstated. In this section, we use potential outcomes to discuss three well-known practical benefits of multiple pre-treatment periods: (1) assessing the parallel trends assumption, (2) improving estimation accuracy, and (3) allowing for a more flexible parallel trends assumption. Our focus is on clarifying necessary assumptions that are often implicit in practice. We discuss each of them in turn and summarize the discussion in Section 3.4.

To describe benefits of multiple pre-treatment periods, this section considers two pre-treatment time periods $t \in \{0, 1\}$ and one post-treatment period $t = 2$. In our running example, two pre-treatment periods are 2006 and 2008, and one post-treatment period is 2010.

3.1 Assessing Parallel Trends Assumption

The first and the most common use of pre-treatment periods is to assess the identification assumption of parallel trends. Because the validity of the DID design rests on the parallel trends assumption, it is critical to evaluate its plausibility in any application. However, the parallel trends assumption itself involves counterfactual outcomes, and thus analysts cannot empirically test it directly. Instead, we often investigate whether trends for treatment and control groups are parallel in pre-treatment periods (Angrist and Pischke, 2008). For example, researchers assess whether trends in the infrastructure quality from 2006 to 2008 — before the treatment implementation in 2009 — are the same for treatment and control communes.

Thus, researchers often estimate the DID for the pre-treatment periods:

$$\left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right), \tag{5}$$

where the first two terms represent the trend in pre-treatment periods for the treatment group and the last two terms quantify the trend for the control group. We then check whether the DID estimate on pre-treatment periods is statistically distinguishable from zero. For example,

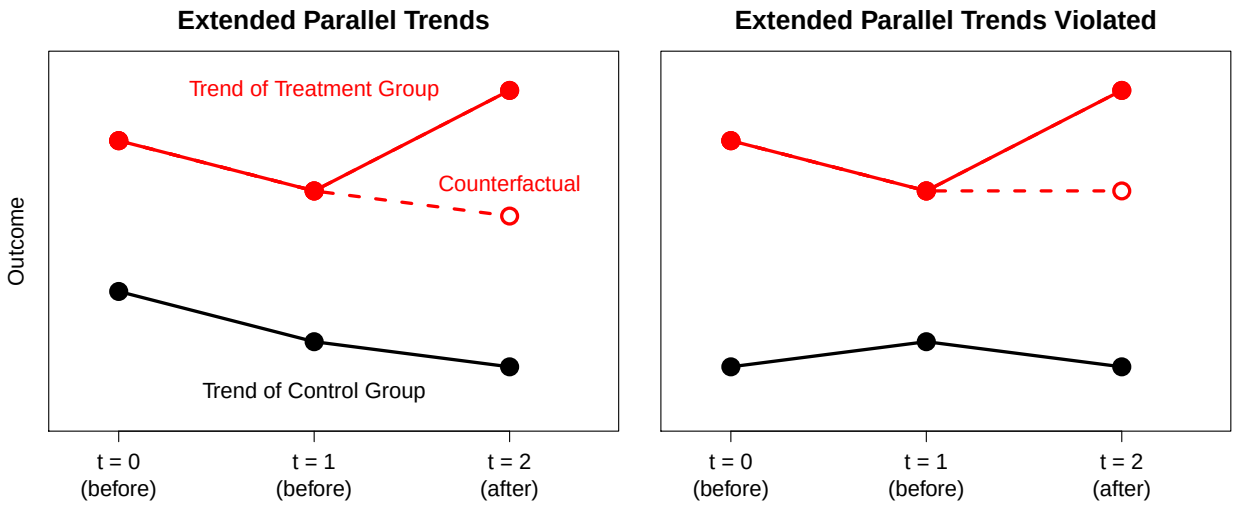

Figure 2: Parallel Pre-treatment Trends (left) and Non-Parallel Pre-treatment Trends (right). *Note:* The extended parallel trends assumption means that the trend of the control group (black solid line) is the same as the trend of the treatment group from $t = 0$ to $t = 1$ as well as from $t = 1$ to $t = 2$ (red solid line and red dotted line).

we can apply the DID estimator to 2006 and 2008 as if 2008 were the post-treatment period and see whether the estimate would be close to zero. In Figure 2, while a DID estimate on the pre-treatment periods would be close to zero for the left panel, it would be negative for the right panel in which two groups have different pre-treatment trends. In Appendix B.4, we show that a robustness check about leads effects (Angrist and Pischke, 2008) — include leads of the treatment variable into the two-way fixed effects regression and check whether their coefficients are zero — is equivalent to this DID on pre-treatment periods.

What are the underlying assumptions behind this test of pre-treatment trends? The basic idea is that if trends are parallel from 2006 to 2008, it is more likely that the parallel trends assumption holds for 2008 and 2010. Hence, instead of considering parallel trends only from 2008 to 2010, the test evaluates the two related parallel trends together. By doing so, this popular test tries to make the DID design falsifiable.

At its core, this approach does not test the parallel trends assumption itself (Assumption 1), which is untestable due to counterfactual outcomes. Instead, it tests the *extended parallel trends* assumption — the parallel trends hold for pre-treatment periods, from $t = 0$ to $t = 1$, as well as from a pre-treatment period $t = 1$ to a post-treatment period $t = 2$:

Assumption 2 (Extended Parallel Trends)

$$\begin{cases} \mathbb{E}[Y_{i2}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i2}(0) \mid G_i = 0] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0] \\ \mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i0}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(0) \mid G_i = 0] - \mathbb{E}[Y_{i0}(0) \mid G_i = 0], \end{cases} \quad (6)$$

where the first line is the same as the standard parallel trends assumption (Equation (2)) and the second line is the parallel trends for pre-treatment periods $t = 0$ and $t = 1$. This assumption means that treatment and control groups have parallel trends of the infrastructure quality from 2008 to 2010 as well as in pre-treatment periods, from 2006 to 2008. Because outcome trends are observable in pre-treatment periods, the test of pre-treatment trends (Equation (5)) directly tests this assumption.

Therefore, many DID studies that exploit the test on pre-treatment trends can be seen as the DID design under the extended parallel trends assumption. Because the extended parallel trends assumption naturally nests the conventional parallel trends assumption, it is also sufficient for identifying the ATT and estimating it via the same DID estimator (Equation (3)).

One potential downside of this DID design under the extended parallel trends is that it makes a stronger assumption than the conventional parallel trends assumption, which only considers trends from 2008 to 2010. It is theoretically possible that researchers might miss opportunities to utilize the DID design when the parallel trends hold only from 2008 to 2010 but not from 2006 to 2008. Therefore, the DID design under the extended parallel trends assumption, which is often recommended in practice, is a more conservative approach because researchers force themselves to focus on robust research design where parallel trends hold for longer time periods. If successful, the DID design under the extended parallel trends assumption guarantees additional credibility.

3.2 Improving Estimation Accuracy

As we discussed above, many existing DID studies that utilize the test of pre-treatment trends can be viewed as the DID design with the extended parallel trends assumption. However, this extended parallel trends assumption is often made implicitly and thus, it is used only for assessing the parallel trends assumption. Fortunately, however, if the extended parallel trends

assumption holds, we can also estimate the ATT with higher accuracy, resulting in smaller standard errors.

This additional benefit becomes clear by simply restating the extended parallel trends assumption as follows.

$$\begin{cases} \mathbb{E}[Y_{i2}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i2}(0) \mid G_i = 0] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0] \\ \mathbb{E}[Y_{i2}(0) \mid G_i = 1] - \mathbb{E}[Y_{i0}(0) \mid G_i = 1] = \mathbb{E}[Y_{i2}(0) \mid G_i = 0] - \mathbb{E}[Y_{i0}(0) \mid G_i = 0], \end{cases} \quad (7)$$

where the first line is the same as the conventional parallel trends assumption and the second line means that the parallel trends hold from a pre-treatment period $t = 0$ to a post-treatment period $t = 2$.

Under the extended parallel trends assumption, there are two natural DID estimators for the ATT. The first is the same as before: the DID on $t = 1$ and $t = 2$. The second is similar but with the additional pre-treatment period: the DID on $t = 0$ and $t = 2$. In our running example, this means that we have a DID estimator using data from 2008 and 2010 and the other using data from 2006 and 2010.

$$\begin{aligned} \widehat{\tau}_{\texttt{DID}} &= \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right), \\ \widehat{\tau}_{\texttt{DID(2,0)}} &= \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right). \end{aligned} \quad (8)$$

Under the extended parallel trends assumption, both estimators are unbiased and consistent for the ATT. Thus, we can increase estimation accuracy by combining the two estimators, for example, simply averaging them.

$$\widehat{\tau}_{\texttt{e-DID}} = \frac{1}{2}\widehat{\tau}_{\texttt{DID}} + \frac{1}{2}\widehat{\tau}_{\texttt{DID(2,0)}}. \quad (9)$$

Intuitively, this extended DID estimator is more efficient because we have more observations to estimate counterfactual outcomes for the treatment group $\mathbb{E}[Y_{i2}(0) \mid G_i = 1]$. See Appendix A.1 for the formal derivation.

It is also useful to connect this to a commonly used regression estimator. As described in Section 2, the standard DID estimator $\widehat{\tau}_{\texttt{DID}}$ can be computed as a coefficient of the linear regression when applied to two time periods $t = 1$ and $t = 2$ (Equation (4)). When we have an additional pre-treatment period $t = 0$ and run the same linear regression with time fixed

effects for $t \in \{0, 1, 2\}$, a coefficient of the treatment variable is now equal to the extended DID estimator $\widehat{\tau}_{\texttt{e-DID}}$ as far as the number of observations does not vary across pre-treatment periods, that is, $n_{g0} = n_{g1}$ for $g = 0, 1$. Similarly, in the panel data settings, when we fit a linear regression with unit and time fixed effects to the entire time periods, a coefficient of the treatment variable is numerically equivalent to the extended DID estimator $\widehat{\tau}_{\texttt{e-DID}}$. We present the general results in Appendix B.2.

This means that we can use all three time points, {2006, 2008, 2010}, to estimate the ATT more efficiently under the extended parallel trends assumption. This connection also implies that if researchers can only justify the parallel trends assumption from 2008 to 2010 but not from 2006 to 2008, the conventional linear regression estimator in Equation (4) should only use data from 2008 and 2010 to avoid bias.

3.3 Allowing For A More Flexible Parallel Trends Assumption

In this section, we consider scenarios in which the extended parallel trends assumption may not be plausible. Fortunately, multiple pre-treatment periods are also useful in accounting for some deviation from the parallel trends assumption. We discuss a popular generalization of the difference-in-difference estimator, a *sequential* DID estimator, which removes bias due to certain violations of the parallel trends assumption (Mora and Reggio, 2012, 2019). We clarify an assumption behind this simple method and relate it to the parallel trends assumption.

To introduce the sequential DID estimator, we begin with the extended parallel trends assumption. As we described in Section 3.1, when the extended parallel trends assumption holds, a DID estimator applied to pre-treatment periods $t = 0$ and $t = 1$ should be zero in expectation. For example, when we apply the DID estimator to 2006 and 2008, before the abolition of elected councils in 2009, the estimate should be indistinguishable from zero if the extended parallel trends assumption holds. In contrast, when trends of treatment and control groups are not parallel, a DID estimate on pre-treatment periods would be non-zero. In its essence, the sequential DID estimator uses this DID estimate from pre-treatment periods to adjust for bias in the standard DID estimator. In particular, it subtracts the DID estimator on pre-treatment periods from the usual DID estimator that uses pre- and post-treatment

periods $t = 1$ and $t = 2$.

$$\widehat{\tau}_{\texttt{s-DID}} = \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right) \right\} - \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right) \right\}, \quad (10)$$

where the first four terms are equal to the standard DID estimator (Equation (3)) and the last four terms are the DID estimator applied to pre-treatment periods $t = 0$ and $t = 1$. In our running example, we can use this sequential DID estimator by first estimating the DID using 2008 and 2010 and then subtracting the DID based on 2006 and 2008. An idea behind this approach is that the DID estimator on pre-treatment periods captures deviation from the parallel trends assumption, and thus we subtract this bias from the usual DID estimator. When trends are parallel in pre-treatment periods, this estimator converges to the standard DID estimator.

Although we would expect that this estimator only requires an assumption weaker than the extended parallel trends, what exact assumption do we need for this sequential DID estimator? At its core, the parallel trends assumption means that differences between treatment and control groups due to unobserved confounders are constant over time. Instead of assuming this constant unmeasured confounding, the sequential DID estimator rests on the *parallel trends-in-trends* assumption — unobserved confounding increases or decreases over time but with some constant rate. Thus, although it cannot deal with all forms of time-varying unobserved confounders, the sequential DID estimator accounts for *linear time-varying* unmeasured confounding. For example, researchers might be worried that some treated communes have higher motivation for reforms, which is not measured, and the infrastructure qualities differ between treated and control communes due to this unobserved motivation. The parallel trends assumption means that the difference in the infrastructure qualities due to this unobserved confounder does not grow or decline over time. In contrast, the parallel trends-in-trends assumption accommodates a simple yet important case in which the unobserved difference in the infrastructure qualities does grow or decline with some fixed rate, which analysts do not need to specify. This parallel trends-in-trends assumption is a generalization of the conventional parallel trends assumption and formally written as follows.

Assumption 3 (Parallel Trends-in-Trends)

$$\{\mathbb{E}[Y_{i2}(0) \mid G_i = 1] - \mathbb{E}[Y_{i2}(0) \mid G_i = 0]\} - \{\mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0]\}$$
$$= \{\mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0]\} - \{\mathbb{E}[Y_{i0}(0) \mid G_i = 1] - \mathbb{E}[Y_{i0}(0) \mid G_i = 0]\} \quad (11)$$

where the left-hand side represents how the unobserved difference between treatment and control groups changes over time from $t = 1$ to $t = 2$. The right-hand side quantifies the same time-varying unmeasured differences from $t = 0$ to $t = 1$. Because the trend of time-varying unmeasured confounding is estimated from pre-treatment periods $t = 0$ to $t = 1$, researchers do not need to know a rate by which time-varying unobserved confounding increases or decreases. Under the parallel trends-in-trends assumption, the sequential DID estimator is unbiased and consistent for the ATT.

Figure 3 illustrates the difference between the extended parallel trends assumption (left panel) and the parallel trends-in-trends assumption (middle panel). We can see in the second row of the figure that the parallel trends-in-trends assumption allows for a linear change in bias over time, whereas the bias is constant over time in the extended parallel trends.

The sequential DID estimator is again connected to a widely used regression estimator. In particular, the sequential DID estimator (Equation (10)) can be computed as a linear regression in which we just replace outcome Y_{it} with the following transformed outcomes.

$$\Delta Y_{it} \sim \alpha_s + \theta_s G_i + \gamma_s I_t + \beta_s (G_i \times I_t), \quad (12)$$

where $\Delta Y_{it} = Y_{it} - (\sum_{i:\, G_i=1} Y_{i,t-1})/n_{1,t-1}$ if $G_i = 1$ and $\Delta Y_{it} = Y_{it} - (\sum_{i:\, G_i=0} Y_{i,t-1})/n_{0,t-1}$ if $G_i = 0$. Coefficients are denoted by $(\alpha_s, \theta_s, \gamma_s, \beta_s)$. In this case, a coefficient in front of the interaction term β_s is numerically identical to the sequential DID estimator. We provide the proof of this equivalence in Appendix B.3. In time-series econometrics, it is common to take the difference in outcomes in order to remove linear time trends before running regressions (Wooldridge, 2010). We also demonstrate that a common robustness check of including group- or unit-specific time trends (Angrist and Pischke, 2008) is also nonparametrically equivalent to the sequential DID estimator. Within the potential outcomes framework, we clarified that these common techniques are justified under the parallel trends-in-trends assumption.

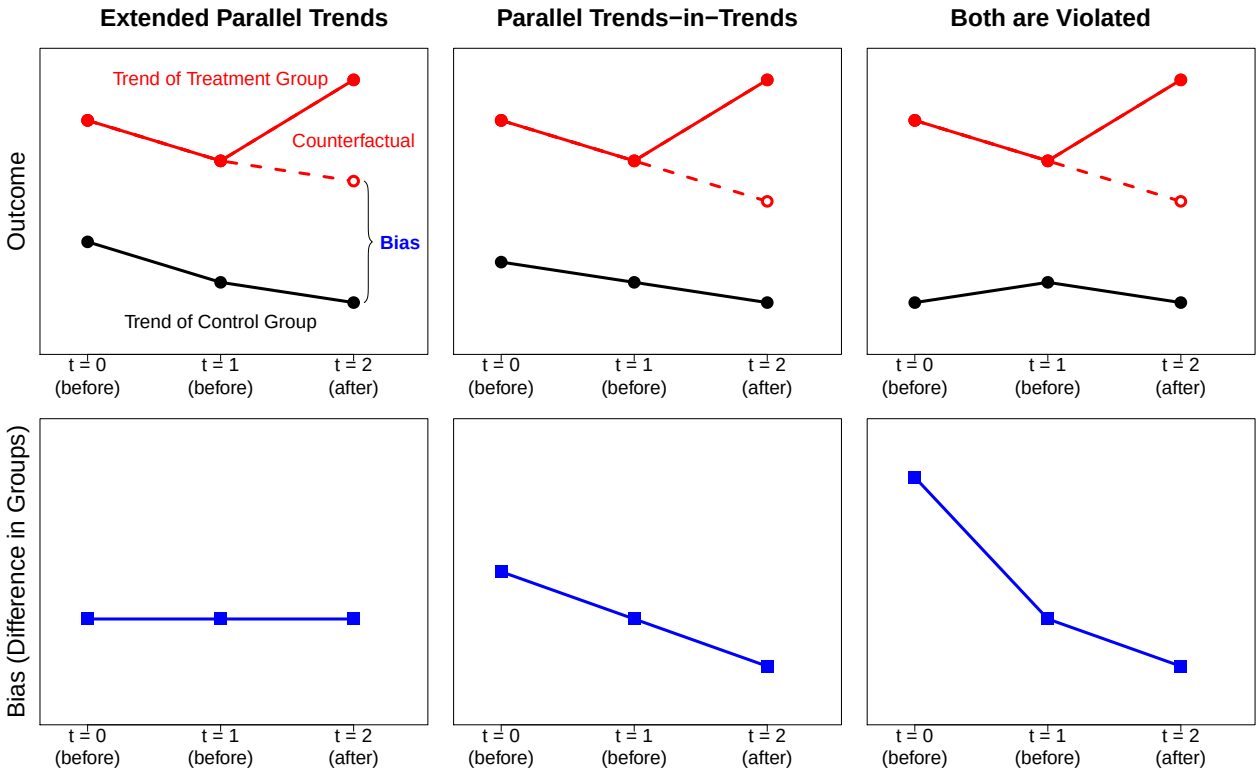

Figure 3: Comparing Extended Parallel Trends Assumption and Parallel Trends-in-Trends Assumption. *Note:* The extended parallel trends assumption (left column) means that the difference in the treatment and control groups (bias) is constant over time. The parallel trends-in-trends assumption (middle column) allows for linear time-varying unmeasured confounding. Both assumptions are violated in the right column.

Researchers might be worried not only about time-varying unmeasured confounders that change over time linearly but also about more general forms of time-varying unmeasured confounders. Fortunately, when we have more than two pre-treatment periods, we can further generalize the sequential DID estimator. In general, when we have K pre-treatment periods, it accounts for the $K-1$ degrees of polynomial functions of unmeasured confounders. In Appendix C, we show our proposed approach allows for this flexible time-varying confounding by incorporating the generalized sequential DID. However, we have to be aware of a key tradeoff; as we incorporate more flexible forms of confounding, standard errors of the sequential DID estimator get much larger. Another powerful approach to address unobserved time-varying confounding is based on a form of partial identification, such as sensitivity analysis (Keele et al., 2019b) and the bracketing bounds (Angrist and Pischke, 2008; Ding and Li, 2019; Keele et al., 2019a).

3.4 Summary of Three Benefits

Before we introduce our proposed estimator in the next section, we summarize the three benefits here to emphasize their close connections.

- **Assessing Parallel Trends Assumption:** Although the parallel trends assumption itself is not directly testable due to counterfactual outcomes, researchers can test the extended parallel trends assumption (Assumption 2). To do so, we rely on a DID estimator on pre-treatment periods from $t = 0$ to $t = 1$. See Equation (5).

- **Improving Estimation Accuracy:** When the extended parallel trends assumption holds, analysts can estimate the ATT with higher accuracy by using the extended DID estimator that naturally uses all the three-time periods $t \in \{0, 1, 2\}$. See Equation (9).

- **Allowing For A More Flexible Parallel Trends Assumption:** Even when the extended parallel trends assumption does not hold, i.e., unmeasured confounding is not constant over time, it might be reasonable to make the parallel trends-in-trends assumption that allows for linear time-varying unmeasured confounding (Assumption 3). Under this assumption, researchers can rely on a more flexible DID estimator, the sequential DID estimator, to estimate the ATT. See Equation (10).

4 Double Difference-in-Differences

While the most basic form of the DID design only requires one pre-treatment period, we see in the previous section that multiple pre-treatment periods provide the three related benefits. In this section, we propose a simple estimator, *double difference-in-differences* (double DID), that blends all the three benefits of multiple pre-treatment periods in a single framework. Here, we introduce the double DID with settings with two pre-treatment periods. We generalize the proposed method to any number of *pre-* and *post*-treatment periods in Appendix C.

4.1 How Double DID Combines Three Benefits

Our proposed double DID estimator is a generalization of existing DID estimators. In its essence, it takes a weighted average of the standard DID and the sequential DID.

$$\widehat{\tau}_{\texttt{d-DID}} = w_1 \times \widehat{\tau}_{\texttt{DID}} + w_2 \times \widehat{\tau}_{\texttt{s-DID}}, \tag{13}$$

where w_1 and w_2 are (potentially negative) weights for the standard DID and the sequential DID, respectively. Importantly, all of the popular estimators that we considered in the previous sections can be seen as special cases of this double DID estimator.

- When $w_1 = 1$ and $w_2 = 0$, the double DID is equal to the standard DID estimator $\widehat{\tau}_{\texttt{DID}}$ that only uses one pre- and one post-treatment periods $t = 1$ and $t = 2$ (Section 2).
- When $w_1 = 3/2$ and $w_2 = -1/2$, it is equal to the extended DID $\widehat{\tau}_{\texttt{e-DID}}$ that improves estimation accuracy when the extended parallel trends assumption holds (Section 3.2).
- When $w_1 = 0$ and $w_2 = 1$, it is equal to the sequential DID estimator $\widehat{\tau}_{\texttt{s-DID}}$ that is consistent for the ATT under the parallel trends-in-trends assumption (Section 3.3).

Thus, the choice of estimators is equivalent to the choice of different weights. A question in practice is, which estimator, i.e., weights w_1 and w_2, should we use under what condition? The double DID addresses this question within a framework of the generalized method of moments (GMM) (Hansen, 1982). In particular, the double DID estimator can be written as the GMM estimator that combines the standard DID estimator and the sequential DID estimator that minimizes the following objective function:

$$\widehat{\tau}_{\texttt{d-DID}} = \underset{\tau}{\operatorname{argmin}} \begin{pmatrix} \tau - \widehat{\tau}_{\texttt{DID}} \\ \tau - \widehat{\tau}_{\texttt{s-DID}} \end{pmatrix}^{\top} \mathbf{W} \begin{pmatrix} \tau - \widehat{\tau}_{\texttt{DID}} \\ \tau - \widehat{\tau}_{\texttt{s-DID}} \end{pmatrix} \tag{14}$$

where weight matrix $\mathbf{W}$ determines weights w_1 and w_2. Thus, we can estimate the optimal weights building on the theory of the efficient GMM. Specifically, the optimal weight matrix that minimizes the variance is the inverse of the variance-covariance matrix of DID estimators

(see Appendix C.3 for the generalized double DID).

$$\widehat{\mathbf{W}} = \begin{pmatrix} \widehat{\text{Var}}(\widehat{\tau}_{\texttt{DID}}) & \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}}) \\ \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}}) & \widehat{\text{Var}}(\widehat{\tau}_{\texttt{s-DID}}) \end{pmatrix}^{-1} \tag{15}$$

From this estimated optimal weight matrix $\widehat{\mathbf{W}}$, we can easily compute implied optimal weights w_1 and w_2 that balance the standard DID estimator and the sequential DID estimator. The optimal weights can be written as

$$w_1 = \frac{\widehat{\text{Var}}(\widehat{\tau}_{\texttt{s-DID}}) - \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}})}{\widehat{\text{Var}}(\widehat{\tau}_{\texttt{DID}}) + \widehat{\text{Var}}(\widehat{\tau}_{\texttt{s-DID}}) - 2\widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}})},$$
$$w_2 = \frac{\widehat{\text{Var}}(\widehat{\tau}_{\texttt{DID}}) - \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}})}{\widehat{\text{Var}}(\widehat{\tau}_{\texttt{DID}}) + \widehat{\text{Var}}(\widehat{\tau}_{\texttt{s-DID}}) - 2\widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}})}.$$

How do we implement this double DID estimator? The double DID estimator proceeds as a two-step process. The first step is to assess the underlying assumptions. We check the extended parallel trends assumption by applying the DID estimator on pre-treatment periods and testing whether the estimate is statistically distinguishable from zero at a conventional level (Equation (5)). Because this standard hypothesis testing approach has a risk of conflating evidence for parallel trends and statistical inefficiency, we also incorporate an equivalence approach (Wellek, 2010; Hartman and Hidalgo, 2018) in which we evaluate the null hypothesis that trends of two groups are *not* parallel in pre-treatment periods. Importantly, this first step of the double DID is equivalent to the overidentification test in the GMM framework, where we assume that the sequential DID estimator is correctly specified and test the null hypothesis that the standard DID estimator is correctly specified. When there are more than two pre-treatment periods, we can diagnose the parallel trends-in-trends assumption by applying the sequential DID estimator to pre-treatment periods.

The second step is the estimation of the ATT. When the extended parallel trends assumption is plausible, we first estimate the optimal weight matrix $\mathbf{W}$ and plug it into the double DID estimator (Equation (14)). When only the parallel trends-in-trends assumption is plausible, the double DID only contains one moment condition $\tau - \widehat{\tau}_{\texttt{s-DID}}$ and thus, is equal to the sequential DID estimator. When both assumptions are implausible, there is no cred-

ible estimator without making further stringent assumptions. In Appendix C, we discuss a generalized double DID that combines K moments using K pre-treatment periods.

This double DID estimator naturally enjoys the three benefits of multiple pre-treatment periods within a unified framework.

1. Assessing Underlying Assumptions The double DID incorporates the assessment of underlying assumptions in its first step as the overidentification test. When the trends in pre-treatment periods are not parallel, researchers have to pay the most careful attention to research design and use domain knowledge to assess the parallel trends-in-trends assumption.

2. Improving Estimation Accuracy When the extended parallel trends assumption holds, researchers can combine two DIDs with equal weights (i.e., the extended DID estimator, which is numerically equivalent to the two-way fixed effects regression) to increase estimation accuracy (Section 3.2). In this setting, the double DID further improves estimation accuracy because it selects the optimal weights as the GMM estimator. In Section 5, we use simulations to demonstrate that the double DID achieves even smaller standard errors than the extended DID estimator.

3. Allowing For A More Flexible Parallel Trends Assumption Under the parallel trends-in-trends assumption, the double DID estimator converges to the sequential DID estimator. However, when the extended parallel trends assumption holds, the double DID uses optimal weights and is not equal to the sequential DID. Thus, the double DID estimator avoids a dilemma of the sequential DID — it is consistent under a weaker assumption of the parallel trends-in-trends but is less efficient when the extended parallel trends assumption holds. By naturally changing weights, the double DID achieves high estimation accuracy under the extended parallel trends assumption and at the same time, allows for more flexible time-varying unmeasured confounding under the parallel trends-in-trends assumption.

4.2 Double DID Regression

Like other DID estimators, the double DID estimator is nicely connected to a widely-used regression approach. This connection is particularly useful when researchers would like to

control for pre-treatment covariates to make the DID design more robust and efficient.

To introduce the regression-based double DID estimator, we begin with the standard DID. As discussed in Section 2, the standard DID estimator is equivalent to a coefficient in the linear regression of Equation (4). Inspired by this connection, researchers often adjust for additional pre-treatment covariates as:

$$Y_{it} \sim \alpha + \theta G_i + \gamma I_t + \beta (G_i \times I_t) + \mathbf{X}_{it}^\top \boldsymbol{\rho}, \tag{16}$$

where we adjust for the additional pre-treatment covariates $\mathbf{X}_{it}$. Here, we make the parallel trends assumption *conditional* on pre-treatment covariates $\mathbf{X}_{it}$. The idea is that even when the parallel trends assumption might not hold without controlling for any covariates, trends of the two groups might be parallel after adjusting for observed covariates. For example, the conditional parallel trends assumption means that treatment and control groups have the same trends of the infrastructure quality after controlling for population size and GDP per capita.

The estimated coefficient $\widehat{\beta}$ is consistent for the ATT when this conditional parallel trends assumption holds and the parametric model is correctly specified. This parametric assumption might be strong, but it is common to all regression strategies, including non-causal settings, and can be assessed via usual model diagnostics.

The sequential DID estimator is extended similarly. Based on the connection to the linear regression of Equation (12), we can adjust for additional pre-treatment covariates as:

$$\Delta Y_{it} \sim \alpha_s + \theta_s G_i + \gamma_s I_t + \beta_s (G_i \times I_t) + \mathbf{X}_{it}^\top \boldsymbol{\rho}_s, \tag{17}$$

where $\Delta Y_{it} = Y_{it} - (\sum_{i:\, G_i=1} Y_{i,t-1})/n_{1,t-1}$ if $G_i = 1$ and $\Delta Y_{it} = Y_{it} - (\sum_{i:\, G_i=0} Y_{i,t-1})/n_{0,t-1}$ if $G_i = 0$. The estimated coefficient $\widehat{\beta}_s$ is consistent for the ATT under the *conditional* parallel trends-in-trends assumption and the conventional assumption of correct specification.

The double DID regression combines the two regression estimators via the GMM:

$$\widehat{\beta}_{\texttt{d-DID}} = \underset{\beta_d}{\operatorname{argmin}} \begin{pmatrix} \beta_d - \widehat{\beta} \\ \beta_d - \widehat{\beta}_s \end{pmatrix}^\top \widehat{\mathbf{W}} \begin{pmatrix} \beta_d - \widehat{\beta} \\ \beta_d - \widehat{\beta}_s \end{pmatrix} \tag{18}$$

where

$$\widehat{\mathbf{W}} = \begin{pmatrix} \widehat{\text{Var}}(\widehat{\beta}) & \widehat{\text{Cov}}(\widehat{\beta}, \widehat{\beta}_s) \\ \widehat{\text{Cov}}(\widehat{\beta}, \widehat{\beta}_s) & \widehat{\text{Var}}(\widehat{\beta}_s) \end{pmatrix}^{-1} \tag{19}$$

Thus, as the basic double DID estimator, the double DID regression also has two steps. The first step is to assess the underlying assumptions. Here, instead of using the standard DID estimator, we use the standard DID regression on pre-treatment periods to assess the conditional extended parallel trends assumption. The second step is to estimate the ATT while adjusting for pre-treatment covariates. Instead of using the double DID estimator without covariates, we implement the regression-based double DID estimator (Equation (18)).

5 Simulation Study

We conduct a simulation study to compare the performance of the various DID estimators discussed in this paper. We demonstrate two key results. First, the double DID is unbiased under the extended parallel trends assumption or under the parallel trends-in-trends assumption. Second, the double DID has the smallest standard errors among unbiased DID estimators. In particular, standard errors of the double DID are smaller than those of the extended DID (i.e., the two-way fixed effects estimator) even under the extended parallel trends assumption.

We compare three DID estimators — the double DID, the extended DID, and the sequential DID — using two scenarios. In the first scenario, the extended parallel trends assumption (Assumption 2) holds where the difference between potential outcomes under control $\mathbb{E}[Y_{it}(0) \mid G_i = 1] - \mathbb{E}[Y_{it}(0) \mid G_i = 0]$ is constant over time. This corresponds to time-invariant unmeasured confounding, and we expect that all the DID estimators are unbiased in this scenario. The second scenario represents the parallel-trends-in-trends assumption (Assumption 3) where unmeasured confounding varies over time linearly. Here, we expect that the double DID and the sequential DID are unbiased, whereas the extended DID is biased.

For each of the two scenarios, we consider the balanced panel data with n units and five-time periods where treatments are assigned at the last time period. We vary the number of units (n) from 100 to 1000 and evaluate the quality of estimators by absolute bias and

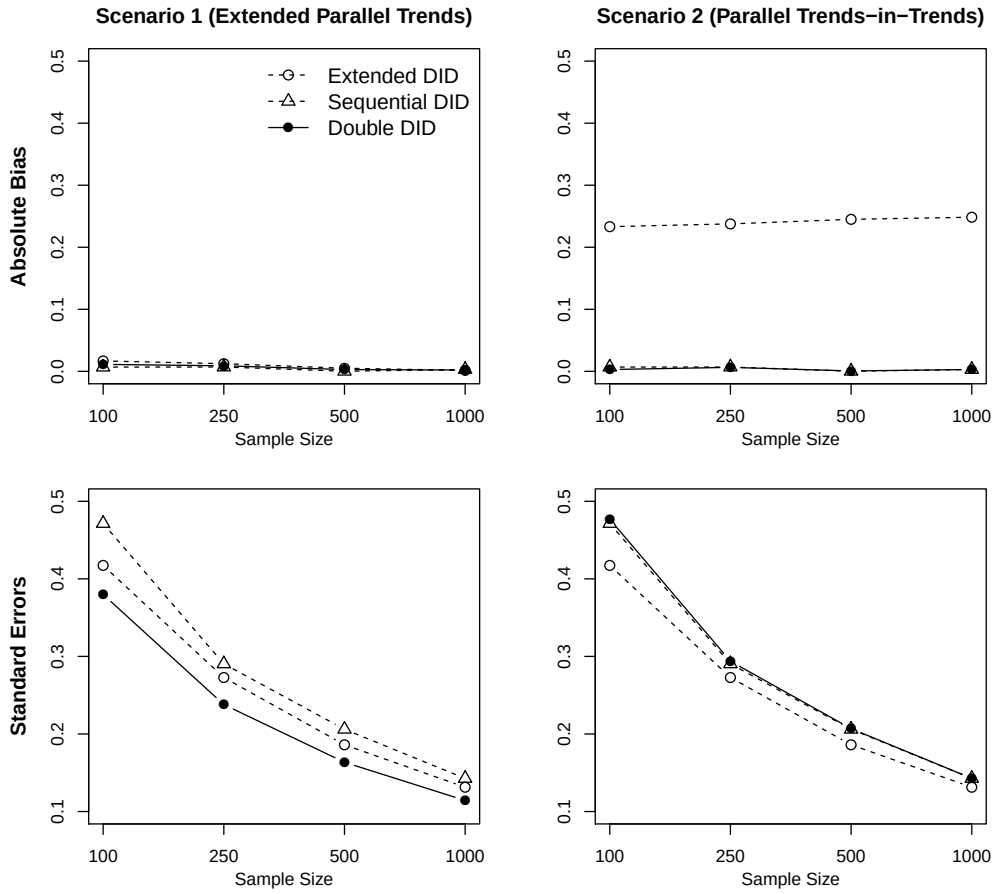

Figure 4: Comparing DID estimators in terms of the absolute bias and the standard errors. The first row shows that the double DID estimator (black circle with solid line) is unbiased under both scenarios. The second row demonstrates that the double DID has the smallest standard errors among unbiased DID estimators.

standard errors over 2000 Monte Carlo simulations. We describe the details of the simulation setup in Appendix D.

Figure 4 shows the results. To begin with the absolute bias, visualized in the first row, all estimators have little bias under the extended parallel trends assumption (Scenario 1), as expected from theoretical results. In contrast, under the parallel-trends-in-trends assumption (Scenario 2), the extended DID (white circle with dotted line) is biased, while the double DID (black circle with solid line) and the sequential DID (white triangle with dotted line) are unbiased.

The second row represents the standard errors of each estimator. Under the extended parallel trends assumption (the first column), the double DID estimator has the smallest standard

error, smaller than the extended DID estimator (i.e., the two-way fixed effects estimator). This efficiency gain comes from the fact that the double DID uses the GMM framework to optimally weight observations from different time periods, although the two-way fixed effects estimator uses equal weights to all pre-treatment periods. In Appendix D, we provide additional simulation results to illustrate that the efficiency gain of the double DID depends on the autocorrelation of errors over time (Bertrand et al., 2004).

Under the parallel trends-in-trends assumption (the second row; the second column), the double DID has almost the same standard error as the sequential DID. This shows that the double DID changes weights according to scenarios and solves a practical dilemma of the sequential DID — it is unbiased under the weaker assumption of the parallel trends-in-trends, but not efficient under the extended parallel trends. In Appendix D, we also show that when the autocorrelation of errors is small, standard errors of the double DID are smaller than those of the sequential DID even under the parallel trends-in-trends assumption.

6 Empirical Application

Malesky et al. (2014) utilize the standard DID design to study how the abolition of elected councils affects local public services in Vietnam. To estimate the causal effects, they rely on data from 2008 and 2010, which are before and after the abolition of elected councils in 2009. Then, the original authors supplement this main analysis by assessing trends in pre-treatment periods from 2006 to 2008. In this section, we apply the proposed method and illustrate how to improve this basic DID design.

Although Malesky et al. (2014) employ the exact same DID design to all of the thirty outcomes they consider, each outcome might require different assumptions as noted in the original paper. Here, we focus on reanalyzing three outcomes that have different patterns of pre-treatment periods. By doing so, we clarify how researchers can use the double DID method to transparently assess underlying assumptions and employ appropriate DID estimators under different settings. We provide an analysis of all the thirty outcomes in Appendix E.

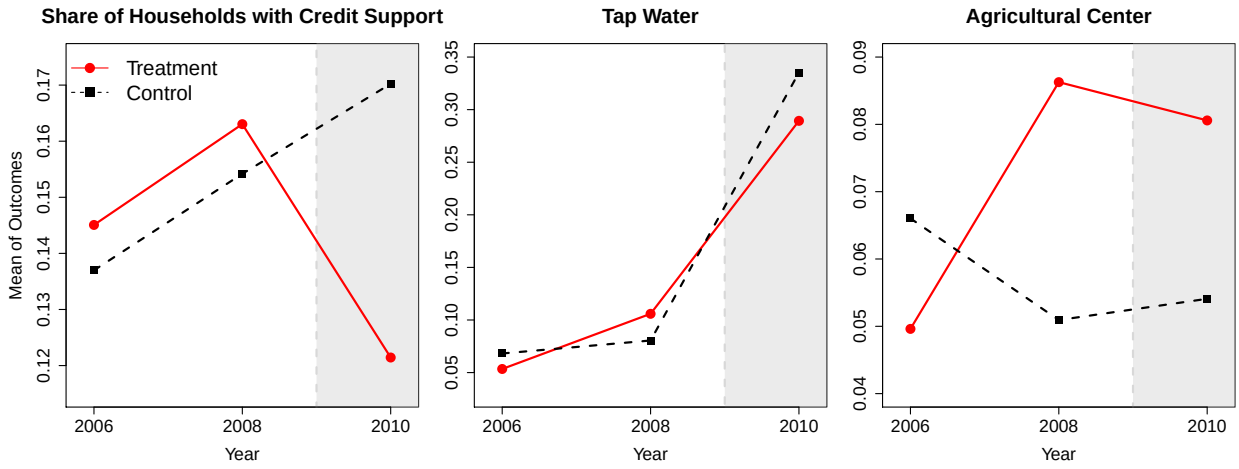

Figure 5: Visualizing Trends of Treatment and Control Groups. *Note*: We report trends for the treatment group (red solid line with circles) and the control group (black dotted line with squares). Two pre-treatment periods are 2006 and 2008. One post-treatment period, 2010, is indicated by the gray shaded area.

6.1 Visualizing and Assessing Underlying Assumptions

The first step of the DID design is to visualize trends of treatment and control groups. Figure 5 shows trends of three different outcomes; "Share of Households with Credit Support," "Tap Water," and "Agricultural Center."[1] Although the original analysis uses the same DID design for all of them, they have distinct trends in pre-treatment periods. The first outcome of "Share of Households with Credit Support " has similar trends in pre-treatment periods. For the second outcome, trends do not look parallel and yet have similar directions. On the other hand, for the third outcome of "Agricultural Center," trends of treatment and control groups have opposite signs. This visualization of trends is a transparent first step to assess the underlying assumptions necessary for the DID estimation.

The next step is to formally assess underlying assumptions. As in the original study, it is common to incorporate additional covariates to make the parallel trends assumption more

[1] "Share of Households with Credit Support" (proportion): Through the HEPR program and other social support programs, how many households/people received credit/loans? "Tap Water" (binary): What is the main source of drinking /cooking water for most people in this commune? "Agricultural Center" (binary): Is there any agriculture extension center in a given commune? Please see Malesky et al. (2014) for further details.

plausible. Based on detailed domain knowledge, Malesky et al. (2014) include four control variables; area size of each commune, population size, whether national-level city or not, and regional fixed effects. Thus, we assess the conditional extended parallel trends assumption by fitting the DID regression (Equation (16)) to pre-treatment periods from 2006 to 2008 where $\mathbf{X}_{it}$ includes the four control variables. If the conditional extended parallel trends assumption holds, estimates of the DID regression on pre-treatment trends should be close to zero. A traditional approach is to assess whether estimates are statistically distinguishable from zero with the conventional 5% level. Based on an equivalence approach that we recommend in Section 4, we also report a 95% equivalence confidence interval, which quantifies the smallest equivalence range supported by the observed data (Hartman and Hidalgo, 2018). For example, if the 95% equivalence confidence interval is $[-\nu, \nu]$, this means that the equivalence test rejects the hypothesis that the DID estimate on pre-treatment periods is larger than ν or smaller than $-\nu$ at the 5% level. Thus, the conditional extended parallel trends assumption is more plausible when the equivalence confidence interval is shorter.

The results are summarized in Table 1. For the first outcome, as a graphical presentation of Figure 5 suggests, statistical tests suggest the extended parallel trends assumption is plausible. The test of the conditional extended parallel trends yields the p-value of 0.587 (the third column), and similarly, the 95% equivalence confidence interval is $[-0.047, 0.047]$ (the fourth column), which is shorter than the other two outcomes discussed below. For the second outcome, the test of the parallel trends produces the p-value of 0.087, which is larger than the conventional level of 0.05. However, the 95% equivalence confidence interval, $[-0.107, 0.107]$, reveals that the parallel trends assumption is less plausible for this outcome than for the first outcome; the DID estimate on pre-treatment trends can be as large as ten percentage points. Finally, for the third outcome, "Agricultural Center," both traditional and equivalence approaches provide little evidence for parallel trends as graphically clear in Figure 5. The test of the parallel trends is rejected at the 5% level (p-value = 0.048) and the 95% equivalence confidence interval is relatively large, $[-0.091, 0.091]$. Although we only have two pre-treatment periods as in the original analysis, if more than two pre-treatment periods are available, researchers can assess the extended parallel trends-in-trends assumption

	Estimate	Standard Error	p-value	Equivalence CI
Share of Households with Credit Support	0.012	0.022	0.587	[-0.047, 0.047]
Tap Water	0.055	0.032	0.087	[-0.107, 0.107]
Agricultural Center	0.050	0.025	0.048	[-0.091, 0.091]

Table 1: Assessing underlying assumptions using the pre-treatment outcomes. *Note*: We evaluate the conditional extended parallel trends assumption for three different outcomes. The table reports DID estimates on pre-treatment trends, standard errors, p-values, and the 95% equivalence confidence interval.

in a similar way by applying the sequential DID estimator to pre-treatment periods. After assessing the underlying parallel trends assumptions, we now proceed to the estimation of the ATT via the double DID.

6.2 Estimating Causal Effects

Within the double DID framework, we select appropriate DID estimators after the empirical assessment of underlying assumptions. For the first outcome of "Share of Households with Credit Support," diagnostics in the previous section suggest that the extended parallel trends assumption is plausible. In such settings, the double DID is expected to produce similar point estimates with smaller standard errors compared to the conventional DID. The first plot of Figure 6 clearly shows this pattern. Using the standard DID estimator, the original estimate of the ATT on "Share of Households with Credit Support" was -0.054 (95% CI $= [-0.096, -0.012]$). Using the double DID estimator, an estimate is instead -0.048 (95% CI $= [-0.083, -0.013]$). By using the double DID estimator, we shrink standard errors by about 20%. Although we only have two pre-treatment periods here, when there are more pre-treatment periods, the efficiency gain of the double DID becomes even larger.

For the second outcome of "Tap Water," we did not have enough evidence to support the extended parallel trends assumption. Thus, instead of using the standard DID as in the original analysis, we rely on the parallel trends-in-trends assumption. In this case, the double DID estimates the ATT by allowing for linear time-varying unmeasured confounding

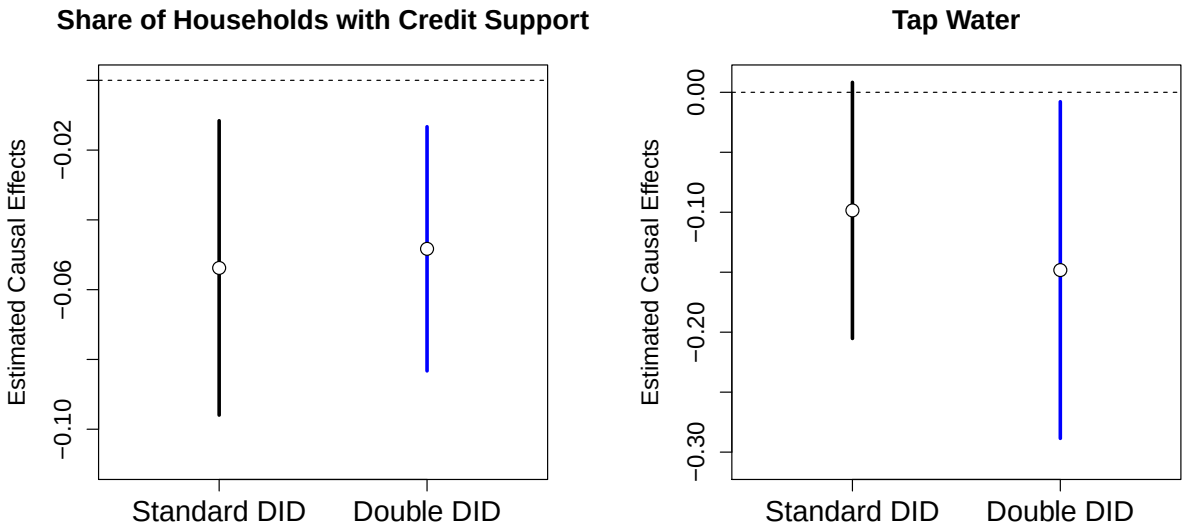

Figure 6: Estimating Causal Effects of Abolishing Elected Councils. *Note:* We compare estimates from the standard DID and the proposed double DID. For the first plot where the extended parallel trends assumption is plausible, the double DID produces a similar point estimate with smaller standard errors. For the second plot where only the parallel trends-in-trends assumption is plausible, the double DID estimator can still estimate the ATT, while the standard DID estimate is likely to be biased.

in contrast to the standard DID that still assumes constant unmeasured confounders. The second plot of Figure 6 shows the important difference between the two methods. Although the standard DID estimate is -0.098 (95% CI $= [-0.205, 0.008]$), the double DID estimate is -0.148 (95% CI $= [-0.289, -0.008]$). Given that the extended parallel trends assumption is implausible, this result suggests that the standard DID suffers from substantial bias, which is more than 30% of the more credible estimate based on the double DID. By incorporating non-parallel pre-treatment trends, the double DID shows that the original DID estimate was underestimated by a large amount. Finally, for the third outcome of "Agricultural Center," the previous diagnostics suggest that the extended parallel trends assumption is implausible. It is possible to use the double DID under the parallel trends-in-trends assumption. However, trends of treatment and control groups have opposite signs, implying the double DID estimates are highly sensitive to the parallel trends-in-trends assumption. Given that the parallel trends-in-trends assumption is also difficult to justify here, there is no credible estimator of the ATT without making additional stringent assumptions. While we mainly focused on the three

outcomes here, the double DID improves upon the standard DID in a similar way for the other outcomes as well (see Appendix E).

7 Concluding Remarks

While the most basic form of the DID only requires two time periods — one before and the other after treatment assignment, researchers can often collect data from several additional pre-treatment periods in a wide range of applications. In this article, we show that such multiple pre-treatment periods can help improve the standard DID design in three ways: (1) assessing underlying assumptions about parallel trends, (2) improving estimation accuracy and (3) enabling more flexible DID estimators. We use the potential outcomes framework to clarify assumptions required to enjoy each benefit. We then introduce a simple method, the double DID, to combine all three benefits within the GMM framework.

This paper also synthesizes a number of estimators popular in analyzing the DID with multiple pre-treatment periods, e.g., nonparametric DID estimators focusing on the last two time periods, the two-way fixed effects regression fitted to all the data, and so on. This wide variation of approaches is partly because researchers rely on different intuitive analogies to the two-time-period DID, some focusing on nonparametric design aspects and others on parametric regression. In this article, we formalize the benefits of multiple pre-treatment periods that have been separately targeted by such existing approaches and show how to jointly achieve them within the double DID framework. Importantly, the double DID contains the popular two-way fixed effects regression and nonparametric DID estimators as special cases and use the GMM to further improve with respect to identification and estimation accuracy.

Although the proposed method provides a simple way to improve upon the standard DID, it has its own important limitations. While we only consider the canonical DID design where treatment assignment happens only once, this may not be the case in an array of important political science applications. One fruitful extension of the method would be to generalize the double DID method to other designs, such as the staggered adoption, where units receive treatments at different timing (Athey and Imbens, 2018). It is also of interest to consider how to assess underlying assumptions and allow for more flexible time trends in general time-series

settings where units may switch in and out of the treatment group at different points in time (e.g., Robins et al., 2000; Blackwell and Glynn, 2018; Imai et al., 2019).

Online Appendix

How to Improve the Difference-in-Differences Design with Multiple Pre-treatment Periods

A Identification and Consistency

This section provides proofs of identification and consistency for the two estimators we discussed in the paper; the extended DID and the sequential DID. For completeness, we also reproduce results that are well known in the literature.

A.1 Extended DID

In Equation (9) of the main text, we define the extended DID as the simple average of the two DID estimator, $\widehat{\tau}_{\texttt{e-DID}} = (\widehat{\tau}_{\texttt{DID}} + \tau_{\texttt{DID(2,0)}})/2$. Here we introduce the generalized extended DID estimator as a weighted average of $\widehat{\tau}_{\texttt{DID}}$ and $\tau_{\texttt{DID(2,0)}}$ and prove that it is consistent for τ under the extended parallel trends assumption (Assumption 2).

Result 1 (Consistency of the Extended DID Under Extended Parallel Trends) Under the extended parallel trends assumption (Assumption 2),

$$\widehat{\tau}_{\texttt{e-DID}} \xrightarrow{p} \tau$$

Proof. Because the parallel trends assumption is implied by the extended parallel trends assumption, we know that $\widehat{\tau}_{\texttt{DID}} \xrightarrow{p} \tau$ under the extended parallel trends assumption, using the well-known results for the standard DID (e.g., Angrist and Pischke, 2008). In addition, under the extended parallel trends assumption,

$$\tau = \Big\{\mathbb{E}[Y_{i2} \mid G_i = 1] - \mathbb{E}[Y_{i0} \mid G_i = 1]\Big\} - \Big\{\mathbb{E}[Y_{i0} \mid G_i = 0] - \mathbb{E}[Y_{i0} \mid G_i = 0]\Big\}.$$

Therefore, by the law of large numbers, $\widehat{\tau}_{\texttt{DID(2,0)}} \xrightarrow{p} \tau$ where $\widehat{\tau}_{\texttt{DID(2,0)}}$ is defined in Equation (8). Since we define the extended DID as the linear combination of the two, $\widehat{\tau}_{\texttt{e-DID}} = \lambda\widehat{\tau}_{\texttt{DID}} + (1 - \lambda)\widehat{\tau}_{\texttt{DID(2,0)}}$ for $\lambda \in \mathbb{R}$, $\widehat{\tau}_{\texttt{e-DID}}$ is also a consistent estimator for τ by the continuous mapping theorem. □

Next we prove the efficiency gain of the extended DID relative to the standard DID.

Result 2 (Efficiency Gain of the Extended DID) The variance of the extended DID is smaller than the larger one of the variance of $\widehat{\tau}_{\texttt{DID}}$ and $\widehat{\tau}_{\texttt{DID(2,0)}}$.

$$\mathrm{Var}(\widehat{\tau}_{\texttt{e-DID}}) \leq \max\{\mathrm{Var}(\widehat{\tau}_{\texttt{DID}}), \mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}})\}.$$

Under the assumption of the equal variance of $\widehat{\tau}_{\texttt{DID}}$ and $\widehat{\tau}_{\texttt{DID(2,0)}}$,

$$\mathrm{Var}(\widehat{\tau}_{\texttt{e-DID}}) \leq \mathrm{Var}(\widehat{\tau}_{\texttt{DID}}).$$

Proof. Define $\mathrm{Var}^* \equiv \max\{\mathrm{Var}(\widehat{\tau}_{\texttt{DID}}), \mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}})\}$. Then,

$$\begin{aligned}
\mathrm{Var}(\widehat{\tau}_{\texttt{e-DID}}) &= \mathrm{Var}(\lambda\widehat{\tau}_{\texttt{DID}} + (1-\lambda)\widehat{\tau}_{\texttt{DID(2,0)}}) \\
&= \lambda^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID}}) + (1-\lambda)^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}}) + 2\lambda(1-\lambda)\mathrm{Cov}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{DID(2,0)}}) \\
&\leq \lambda^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID}}) + (1-\lambda)^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}}) + 2\lambda(1-\lambda)\sqrt{\mathrm{Var}(\widehat{\tau}_{\texttt{DID}})\mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}})} \\
&\leq \lambda^2\mathrm{Var}^* + (1-\lambda)^2\mathrm{Var}^* + 2\lambda(1-\lambda)\mathrm{Var}^* = \mathrm{Var}^*
\end{aligned}$$

where the first line follows from the definition of the extended DID, the second from the definition of variance, the third from the Cauchy-Schwarz inequality, and the fourth from the definition of Var^*. Under the assumption that $\mathrm{Var}(\widehat{\tau}_{\texttt{DID}}) = \mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}})$,

$$\begin{aligned}
\mathrm{Var}(\widehat{\tau}_{\texttt{e-DID}}) &= \mathrm{Var}(\lambda\widehat{\tau}_{\texttt{DID}} + (1-\lambda)\widehat{\tau}_{\texttt{DID(2,0)}}) \\
&= \lambda^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID}}) + (1-\lambda)^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}}) + 2\lambda(1-\lambda)\mathrm{Cov}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{DID(2,0)}}) \\
&\leq \lambda^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID}}) + (1-\lambda)^2\mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}}) + 2\lambda(1-\lambda)\sqrt{\mathrm{Var}(\widehat{\tau}_{\texttt{DID}})\mathrm{Var}(\widehat{\tau}_{\texttt{DID(2,0)}})} \\
&= \mathrm{Var}(\widehat{\tau}_{\texttt{DID}}),
\end{aligned}$$

which completes the proof. □

A.2 Sequential DID

For completeness, we first reproduce the identification result (Mora and Reggio, 2019) that the ATT is identified under the parallel trends-in-trends assumption (Assumption 3).

Result 3 (Identification under Parallel Trends-in-Trends Assumption (Mora and Reggio, 2019)) Under the parallel trends-in-trends assumption (Assumption 3),

$$\begin{aligned}\tau = &\Big\{\big(\mathbb{E}[Y_{i2} \mid G_i = 1] - \mathbb{E}[Y_{i1} \mid G_i = 1]\big) - \big(\mathbb{E}[Y_{i2} \mid G_i = 0] - \mathbb{E}[Y_{i1} \mid G_i = 0]\big)\Big\} \\ &- \Big\{\big(\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i0} \mid G_i = 1]\big) - \big(\mathbb{E}[Y_{i1} \mid G_i = 0] - \mathbb{E}[Y_{i0} \mid G_i = 0]\big)\Big\}\end{aligned}$$

Proof. From the parallel trends-in-trends assumption, we can rewrite $\mathbb{E}[Y_{i2}(0) \mid G_i = 1]$ as

$$\begin{aligned}\mathbb{E}[Y_{i2}(0) \mid G_i = 1] &= \mathbb{E}[Y_{i2}(0) \mid G_i = 0] + \Big(\mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0]\Big) \\ &\quad + \Big(\mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0]\Big) - \Big(\mathbb{E}[Y_{i0}(0) \mid G_i = 1] - \mathbb{E}[Y_{i0}(0) \mid G_i = 0]\Big) \\ &= \mathbb{E}[Y_{i2} \mid G_i = 0] + \Big(\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i1} \mid G_i = 0]\Big) \\ &\quad + \Big(\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i1} \mid G_i = 0]\Big) - \Big(\mathbb{E}[Y_{i0} \mid G_i = 1] - \mathbb{E}[Y_{i0} \mid G_i = 0]\Big).\end{aligned}$$

Therefore,

$$\begin{aligned}\tau &= \mathbb{E}[Y_{i2}(1) - Y_{i2}(0) \mid G_i = 1] \\ &= \Big\{\Big(\mathbb{E}[Y_{i2} \mid G_i = 1] - \mathbb{E}[Y_{i2} \mid G_i = 0]\Big) - \Big(\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i1} \mid G_i = 0]\Big)\Big\} \\ &\quad - \Big\{\Big(\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i1} \mid G_i = 0]\Big) - \Big(\mathbb{E}[Y_{i0} \mid G_i = 1] - \mathbb{E}[Y_{i0} \mid G_i = 0]\Big)\Big\},\end{aligned}$$

which completes the proof. □

It is straightforward to show that $\widehat{\tau}_{\texttt{s-DID}}$ (Equation (10)) is a consistent estimator for τ under the parallel trends-in-trends assumption (Mora and Reggio, 2019).

B Nonparametric Equivalence to Regression Estimators

In this section, we prove the nonparametric connection between regression estimators and the three DID estimators we discussed in the paper.

B.1 Standard DID

B.1.1 Repeated Cross-Sectional Data

For the later use in this Appendix, we report the well-known result that the standard DID estimator $\widehat{\tau}_{\texttt{DID}}$ (Equation (3)) is equivalent to coefficient $\widehat{\beta}$ in the regression estimator (Equation (4)) (Abadie, 2005).

We define O_{it} to be an indicator variable taking the value 1 when individual i is observed in time period t. Using this notation, we prove the following result.

Result 4 (Nonparametric Equivalence of the Standard DID and Regression Estimator) We write the linear regression estimator (Equation (4)) as a solution to the following least squares problem.

$$(\widehat{\alpha}, \widehat{\theta}, \widehat{\gamma}, \widehat{\beta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=1}^{2} O_{it} \Big\{ Y_{it} - \alpha - \theta G_i - \gamma I_t - \beta (G_i \times I_t) \Big\}^2.$$

Then, $\widehat{\tau}_{\texttt{DID}} = \widehat{\beta}$.

Proof. By solving the least squares problem, we obtain

$$\begin{aligned}
\widehat{\alpha} &= \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \\
\widehat{\theta} &= \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \\
\widehat{\gamma} &= \frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \\
\widehat{\beta} &= \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right),
\end{aligned}$$

which completes the proof. □

B.1.2 Panel Data

Again, for the later use in the Appendix, we report the well-known result that the standard DID estimator $\widehat{\tau}_{\texttt{DID}}$ (Equation (3)) is equivalent to coefficient $\widehat{\beta}$ in the two-way fixed effects regression estimator in the panel data setting (Abadie, 2005).

Result 5 (Nonparametric Equivalence of the Standard DID and Two-way Fixed Effects Regression Estimator) We can write the two-way fixed effects regression estimator as a solution to the following least squares problem.

$$(\widehat{\alpha}, \widehat{\delta}, \widehat{\beta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=1}^{2} (Y_{it} - \alpha_i - \delta_t - \beta D_{it})^2.$$

Then, $\widehat{\tau}_{\texttt{DID}} = \widehat{\beta}$.

Proof. First we define $\overline{Y}_i = \sum_{t=1}^{2} Y_{it}/2$, $\overline{Y}_t = \sum_{i=1}^{n} Y_{it}/n$, $\overline{Y} = \sum_{i=1}^{n} \sum_{t=1}^{2} Y_{it}/2n$, $\overline{D}_i = \sum_{t=1}^{2} D_{it}/2$, $\overline{D}_t = \sum_{i=1}^{n} D_{it}/n$, and $\overline{D} = \sum_{i=1}^{n} \sum_{t=1}^{2} D_{it}/2n$. Then, we can transform the least squares problem into a well-known demeaned form.

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \sum_{t=1}^{2} (\widetilde{Y}_{it} - \beta \widetilde{D}_{it})^2$$

where $\widetilde{Y}_{it} = Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{Y}$ and $\widetilde{D}_{it} = D_{it} - \overline{D}_i - \overline{D}_t + \overline{D}$. Using this notation, we can express $\widehat{\beta}$ as

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{2} \widetilde{D}_{it} \widetilde{Y}_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{2} \widetilde{D}_{it}^2}$$

where $\widetilde{D}_{it}$ takes the following form,

$$\widetilde{D}_{it} = \begin{cases} 1/2 \cdot n_0/n & \text{if } G_i = 1, t = 2 \\ -(1/2) \cdot n_0/n & \text{if } G_i = 1, t = 1 \\ -(1/2) \cdot n_1/n & \text{if } G_i = 0, t = 2 \\ 1/2 \cdot n_1/n & \text{if } G_i = 0, t = 1, \end{cases}$$

where $n_1 = \sum_{i=1}^{n} G_i$ and $n_0 = \sum_{i=1}^{n} (1 - G_i)$. Then, the numerator can be written as

$$\sum_{i=1}^{n} \sum_{t=1}^{2} \widetilde{D}_{it} \widetilde{Y}_{it} = \frac{n_0}{2n} \left\{ \sum_{i=1}^{n} G_i \widetilde{Y}_{i2} - \sum_{i=1}^{n} G_i \widetilde{Y}_{i1} \right\} - \frac{n_1}{2n} \left\{ \sum_{i=1}^{n} (1 - G_i) \widetilde{Y}_{i2} - \sum_{i=1}^{n} (1 - G_i) \widetilde{Y}_{i1} \right\}$$

and the denominator is given as

$$\sum_{i=1}^{n} \sum_{t=1}^{2} \widetilde{D}_{it}^2 = 2n_1 \left(\frac{n_0}{2n} \right)^2 + 2n_0 \left(\frac{n_1}{2n} \right)^2 = \frac{n_1 n_0}{2n}.$$

Combining both terms, we get

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{2}\widetilde{D}_{it}\widetilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{2}\widetilde{D}_{it}^{2}} \\
&= \frac{1}{n_1}\left\{\sum_{i=1}^{n} G_i\widetilde{Y}_{i2} - \sum_{i=1}^{n} G_i\widetilde{Y}_{i1}\right\} - \frac{1}{n_0}\left\{\sum_{i=1}^{n}(1-G_i)\widetilde{Y}_{i2} - \sum_{i=1}^{n}(1-G_i)\widetilde{Y}_{i1}\right\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n} G_i(Y_{i2}-Y_{i1}) - \frac{1}{n_0}\sum_{i=1}^{n}(1-G_i)(Y_{i2}-Y_{i1}) \\
&= \widehat{\tau}_{\texttt{DID}},
\end{aligned}$$

which concludes the proof. □

B.2 Extended DID

B.2.1 Repeated Cross-Sectional Data

We consider a case in which there are two pre-treatment periods $t = \{0, 1\}$ and one post-treatment period $t = 2$. Using this notation, we prove the following result. Similar derivations are provided in Goodman-Bacon (2018) and Strezhnev (2018) for a different setting of the staggered adoption design.

Theorem 1 (Nonparametric Equivalence of the Extended DID and Regression Estimator) We focus on a linear regression estimator that is a solution to the following least squares problem.

$$(\widehat{\theta}, \widehat{\gamma}, \widehat{\beta}) = \text{argmin} \sum_{i=1}^{n}\sum_{t=0}^{2} O_{it}\left(Y_{it} - \theta G_i - \gamma_t - \beta D_{it}\right)^2.$$

Then, $\widehat{\beta} = \lambda\widehat{\tau}_{\texttt{DID}} + (1-\lambda)\widehat{\tau}_{\texttt{DID(2,0)}}$ where

$$\begin{aligned}
\lambda &= \frac{n_{11}n_{01}(n_{10}+n_{00})}{n_{11}n_{01}(n_{10}+n_{00}) + n_{10}n_{00}(n_{11}+n_{01})}, \\
1-\lambda &= \frac{n_{10}n_{00}(n_{11}+n_{01})}{n_{11}n_{01}(n_{10}+n_{00}) + n_{10}n_{00}(n_{11}+n_{01})}.
\end{aligned}$$

When the sample size of each group is fixed over time, i.e., $n_{11} = n_{10}$ and $n_{01} = n_{00}$, $\lambda = 1/2$ and therefore, $\widehat{\beta}$ is equivalent to the extended DID estimator of equal weights (Equation (9)).

Proof. By solving the least squares problem, we obtain

$$
\begin{aligned}
\widehat{\theta} &= \lambda \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right) + (1-\lambda) \left(\frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right) \\
\widehat{\gamma}_2 &= \frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} \\
\widehat{\gamma}_1 &= \frac{\sum_{i:\, G_i=1} Y_{i1} + \sum_{i:\, G_i=0} Y_{i1}}{n_{11}+n_{01}} - \frac{n_{11}}{n_{11}+n_{01}} \widehat{\theta} \\
\widehat{\gamma}_0 &= \frac{\sum_{i:\, G_i=1} Y_{i0} + \sum_{i:\, G_i=0} Y_{i0}}{n_{10}+n_{00}} - \frac{n_{10}}{n_{10}+n_{00}} \widehat{\theta} \\
\widehat{\beta} &= \lambda \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right) \right\} \\
&\quad + (1-\lambda) \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right) \right\},
\end{aligned}
$$

which completes the proof. □

B.2.2 Panel Data

We prove that the extended DID estimator $\widehat{\tau}_{\texttt{e-DID}}$ (Equation (9)) (equal weights; $\lambda = 1/2$) is equivalent to coefficient $\widehat{\beta}$ in the two-way fixed effects regression estimator in the panel data setting with $t = \{0, 1, 2\}$. Similar derivations are provided in Goodman-Bacon (2018) and Strezhnev (2018) for a different setting of the staggered adoption design.

Theorem 2 (Nonparametric Equivalence of the Extended DID and Two-way Fixed Effects Regression Estimator) We can write the two-way fixed effects regression estimator as a solution to the following least squares problem.

$$
(\widehat{\alpha}, \widehat{\delta}, \widehat{\beta}) = \operatorname{argmin} \sum_{i=1}^{n} \sum_{t=0}^{2} (Y_{it} - \alpha_i - \delta_t - \beta D_{it})^2.
$$

Then, $\widehat{\tau}_{\texttt{e-DID}} = \widehat{\beta}$.

Proof. First we define $\overline{Y}_i = \sum_{t=0}^{2} Y_{it}/3$, $\overline{Y}_t = \sum_{i=1}^{n} Y_{it}/n$, $\overline{Y} = \sum_{i=1}^{n} \sum_{t=0}^{2} Y_{it}/3n$, $\overline{D}_i = \sum_{t=0}^{2} D_{it}/3$, $\overline{D}_t = \sum_{i=1}^{n} D_{it}/n$, and $\overline{D} = \sum_{i=1}^{n} \sum_{t=0}^{2} D_{it}/3n$. Then, we can write the two-way fixed effects estimator as a two-way demeaned estimator,

$$
\widehat{\beta} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} \sum_{t=0}^{2} (\widetilde{Y}_{it} - \beta \widetilde{D}_{it})^2 = \frac{\sum_{i=1}^{n} \sum_{t=0}^{2} \widetilde{D}_{it} \widetilde{Y}_{it}}{\sum_{i=1}^{n} \sum_{t=0}^{2} \widetilde{D}_{it}^2},
$$

as in Result 5, where $\widetilde{Y}_{it} = Y_{it} - \overline{Y}_{i} - \overline{Y}_{t} + \overline{Y}$ and $\widetilde{D}_{it} = D_{it} - \overline{D}_{i} - \overline{D}_{t} + \overline{D}$. Importantly, $\widetilde{D}_{it}$ takes the following form:

$$\widetilde{D}_{it} = \begin{cases} 2/3 \cdot n_0/n & \text{if } G_i = 1, t = 2 \\ -1/3 \cdot n_0/n & \text{if } G_i = 1, t = 0, 1 \\ -2/3 \cdot n_1/n & \text{if } G_i = 0, t = 2 \\ 1/3 \cdot n_1/n & \text{if } G_i = 0, t = 0, 1, \end{cases}$$

where $n_1 = \sum_{i=1}^{n} G_i$ and $n_0 = \sum_{i=1}^{n}(1 - G_i)$. Then, the numerator can be written as

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{t=0}^{2} \widetilde{D}_{it}\widetilde{Y}_{it} \\
&= \sum_{i=1}^{n} G_i\left(\frac{2n_0}{3n}\right)\widetilde{Y}_{i2} - \sum_{i=1}^{n}\sum_{t=0}^{1} G_i\left(\frac{n_0}{3n}\right)\widetilde{Y}_{it} + \sum_{i=1}^{n}(1-G_i)\left(\frac{-2n_1}{3n}\right)\widetilde{Y}_{i2} + \sum_{i=1}^{n}\sum_{t=0}^{1}(1-G_i)\left(\frac{n_1}{3n}\right)\widetilde{Y}_{it} \\
&= \sum_{i=1}^{n} G_i\left(\frac{n_0}{3n}\right)\{\widetilde{Y}_{i2} - \widetilde{Y}_{i1}\} + \sum_{i=1}^{n} G_i\left(\frac{n_0}{3n}\right)\{\widetilde{Y}_{i2} - \widetilde{Y}_{i0}\} \\
&\quad - \left\{\sum_{i=1}^{n}(1-G_i)\left(\frac{n_1}{3n}\right)\{\widetilde{Y}_{i2} - \widetilde{Y}_{i1}\} + \sum_{i=1}^{n}(1-G_i)\left(\frac{n_1}{3n}\right)\{\widetilde{Y}_{i2} - \widetilde{Y}_{i0}\}\right\} \\
&= \frac{n_0}{3n}\left\{\sum_{i=1}^{n} G_i\{Y_{i2} - Y_{i1}\} + \sum_{i=1}^{n} G_i\{Y_{i2} - Y_{i0}\}\right\} - \frac{n_1}{3n}\left\{\sum_{i=1}^{n}(1-G_i)\{Y_{i2} - Y_{i1}\} + \sum_{i=1}^{n}(1-G_i)\{Y_{i2} - Y_{i0}\}\right\}.
\end{aligned}$$

The denominator can be written as

$$\sum_{i=1}^{n}\sum_{t=0}^{2} \widetilde{D}_{it}^2 = \frac{n_0 n_1}{n} \cdot \frac{2}{3}.$$

Combining the two terms, we have

$$\begin{aligned}
\widehat{\beta} &= \frac{1}{2n_1}\left\{\sum_{i=1}^{n} G_i\{Y_{i2} - Y_{i1}\} + \sum_{i=1}^{n} G_i\{Y_{i2} - Y_{i0}\}\right\} \\
&\quad - \frac{1}{2n_0}\left\{\sum_{i=1}^{n}(1-G_i)\{Y_{i2} - Y_{i1}\} + \sum_{i=1}^{n}(1-G_i)\{Y_{i2} - Y_{i0}\}\right\} \\
&= \frac{1}{2}\left\{\frac{1}{n_1}\sum_{i=1}^{n} G_i\{Y_{i2} - Y_{i1}\} - \frac{1}{n_0}\sum_{i=1}^{n}(1-G_i)\{Y_{i2} - Y_{i1}\}\right\} \\
&\quad + \frac{1}{2}\left\{\frac{1}{n_1}\sum_{i=1}^{n} G_i\{Y_{i2} - Y_{i0}\} - \frac{1}{n_0}\sum_{i=1}^{n}(1-G_i)\{Y_{i2} - Y_{i0}\}\right\} \\
&= \frac{1}{2}\widehat{\tau}_{\texttt{DID}} + \frac{1}{2}\widehat{\tau}_{\texttt{DID(2,0)}},
\end{aligned}$$

which completes the proof. □

B.3 Sequential DID

B.3.1 Repeated Cross-Sectional Data

We prove that the sequential DID estimator $\widehat{\tau}_{\texttt{s-DID}}$ (Equation (10)) is equivalent to a coefficient in a regression estimator with transformed outcomes.

Theorem 3 (Nonparametric Equivalence of the Sequential DID and Regression Estimator) We focus on a linear regression estimator with a transformed outcome.

$$(\widehat{\alpha}, \widehat{\theta}, \widehat{\gamma}, \widehat{\beta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=1}^{2} O_{it} \Big\{ \Delta Y_{it} - \alpha - \theta G_i - \gamma I_t - \beta (G_i \times I_t) \Big\}^2,$$

where

$$\Delta Y_{it} = \begin{cases} Y_{i2} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} & \text{if } G_i = 1, t = 2 \\ Y_{i1} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} & \text{if } G_i = 1, t = 1 \\ Y_{i2} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} & \text{if } G_i = 0, t = 2 \\ Y_{i1} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} & \text{if } G_i = 0, t = 1. \end{cases}$$

Then, $\widehat{\tau}_{\texttt{s-DID}} = \widehat{\beta}$.

Proof. Using Result 4, we obtain

$$\begin{aligned} \widehat{\beta} &= \left(\frac{\sum_{i:\, G_i=1} \Delta Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} \Delta Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} \Delta Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} \Delta Y_{i1}}{n_{01}} \right) \\ &= \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right) \right\} \\ &\qquad - \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right) \right\}, \end{aligned}$$

which completes the proof. □

Next, we prove that the sequential DID estimator $\widehat{\tau}_{\texttt{s-DID}}$ (Equation (10)) is also equivalent to a coefficient in a regression estimator with group-specific time trends. Mora and Reggio (2019) derive similar results by making the parametric assumption of the conditional expectations. We prove nonparametric equivalence without making any assumptions about conditional expectations.

Theorem 4 (Nonparametric Equivalence of the Sequential DID and Regression Estimator with Group-Specific Time Trends) We focus on a linear regression estimator with group-specific time trends.

$$(\widehat{\theta}, \widehat{\gamma}, \widehat{\beta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=0}^{2} O_{it} \Big\{ Y_{it} - \theta_0 G_i - \theta_1 (G_i \times t) - \gamma_t - \beta D_{it} \Big\}^2.$$

Then, $\widehat{\tau}_{\texttt{s-DID}} = \widehat{\beta}$.

Proof. By solving the least squares problem, we obtain

$$\begin{aligned}
\widehat{\theta}_0 &= \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \\
\widehat{\theta}_1 &= \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right) - \left(\frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right) \\
\widehat{\gamma}_2 &= \frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}}, \quad \widehat{\gamma}_1 = \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}}, \quad \widehat{\gamma}_0 = \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \\
\widehat{\beta} &= \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right) \right\} \\
&\qquad - \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right) \right\},
\end{aligned}$$

which completes the proof. □

B.3.2 Panel Data

We prove that the sequential DID estimator $\widehat{\tau}_{\texttt{s-DID}}$ (Equation (10)) is equivalent to a coefficient in the two-way fixed effects regression estimator with transformed outcomes.

Theorem 5 (Nonparametric Equivalence of the Sequential DID and Two-way Fixed Effects Regression Estimator) We focus on the two-way fixed effects regression estimator with transformed outcomes.

$$(\widehat{\alpha}, \widehat{\delta}, \widehat{\beta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=1}^{2} (\Delta Y_{it} - \alpha_i - \delta_t - \beta D_{it})^2,$$

where $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$. Then, $\widehat{\tau}_{\texttt{s-DID}} = \widehat{\beta}$.

Proof. As in Result 5, we can focus on the demeaned form.

$$\widehat{\beta} = \text{argmin} \sum_{i=1}^{n} \sum_{t=1}^{2} (\widetilde{\Delta Y}_{it} - \beta \widetilde{D}_{it})^2,$$

where $\widetilde{\Delta Y}_{it} = \Delta Y_{it} - \overline{\Delta Y}_i - \overline{\Delta Y}_t + \overline{\Delta Y}$, $\overline{\Delta Y}_i = \sum_{t=1}^{2} \Delta Y_{it}/2$, $\overline{\Delta Y}_t = \sum_{i=1}^{n} \Delta Y_{it}/n$, and $\overline{\Delta Y} = \sum_{i=1}^{n}\sum_{t=1}^{2} \Delta Y_{it}/2n$. Similarly, $\widetilde{D}_{it} = D_{it} - \overline{D}_i - \overline{D}_t + \overline{D}$, $\overline{D}_i = \sum_{t=1}^{2} D_{it}/2$, $\overline{D}_t = \sum_{i=1}^{n} D_{it}/n$, and $\overline{D} = \sum_{i=1}^{n}\sum_{t=1}^{2} D_{it}/2n$. Using Result 5,

$$
\begin{aligned}
\widehat{\beta} &= \frac{1}{n_1}\sum_{i=1}^{n} G_i(\Delta Y_{i2} - \Delta Y_{i1}) - \frac{1}{n_0}\sum_{i=1}^{n}(1-G_i)(\Delta Y_{i2} - \Delta Y_{i1}) \\
&= \left\{\frac{1}{n_1}\sum_{i=1}^{n} G_i(Y_{i2} - Y_{i1}) - \frac{1}{n_0}\sum_{i=1}^{n}(1-G_i)(Y_{i2} - Y_{i1})\right\} \\
&\qquad - \left\{\frac{1}{n_1}\sum_{i=1}^{n} G_i(Y_{i1} - Y_{i0}) - \frac{1}{n_0}\sum_{i=1}^{n}(1-G_i)(Y_{i1} - Y_{i0})\right\} \\
&\equiv \widehat{\tau}_{\texttt{s-DID}},
\end{aligned}
$$

which concludes the proof. □

Next, we prove that the sequential DID estimator $\widehat{\tau}_{\texttt{s-DID}}$ (Equation (10)) is also equivalent to a coefficient in the two-way fixed effects regression estimator with individual-specific time trends.

Theorem 6 (Nonparametric Equivalence of the Sequential DID and Two-way Fixed Effects Regression Estimator with Individual-Specific Time Trends) We focus on the two-way fixed effects regression estimator with individual-specific time trends

$$
(\widehat{\alpha}, \widehat{\xi}, \widehat{\delta}, \widehat{\beta}) = \text{argmin} \sum_{i=1}^{n}\sum_{t=0}^{2}(Y_{it} - \alpha_i - (\xi_i \times t) - \delta_t - \beta D_{it})^2.
$$

Then, $\widehat{\tau}_{\texttt{s-DID}} = \widehat{\beta}$.

Proof. By solving the least squares problem, we obtain

$$
\begin{aligned}
&\sum_{i:\, G_i=1} Y_{i2} = (\widehat{\beta} + \widehat{\gamma}_2)n_1 + \sum_{i:\, G_i=1}\widehat{\alpha}_i + 2\sum_{i:\, G_i=1}\widehat{\xi}_i, \quad \sum_{i:\, G_i=0} Y_{i2} = \widehat{\gamma}_2 n_0 + \sum_{i:\, G_i=0}\widehat{\alpha}_i + 2\sum_{i:\, G_i=0}\widehat{\xi}_i \\
&\sum_{i:\, G_i=1} Y_{i1} = \widehat{\gamma}_1 n_1 + \sum_{i:\, G_i=1}\widehat{\alpha}_i + \sum_{i:\, G_i=1}\widehat{\xi}_i, \quad \sum_{i:\, G_i=0} Y_{i1} = \widehat{\gamma}_1 n_0 + \sum_{i:\, G_i=0}\widehat{\alpha}_i + \sum_{i:\, G_i=0}\widehat{\xi}_i \\
&\sum_{i:\, G_i=1} Y_{i0} = \widehat{\gamma}_0 n_1 + \sum_{i:\, G_i=1}\widehat{\alpha}_i, \quad \sum_{i:\, G_i=0} Y_{i0} = \widehat{\gamma}_0 n_0 + \sum_{i:\, G_i=0}\widehat{\alpha}_i.
\end{aligned}
$$

Therefore, we get

$$
\widehat{\beta} = \left\{\left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_1} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_1}\right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_0} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_0}\right)\right\}
$$

$$- \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_1} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_1} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_0} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_0} \right) \right\},$$

which completes the proof. □

B.4 Connection to the Leads Test

Here we formally prove the connection between the test of pre-treatment periods discussed in Section 3.1 and the well known leads test (Angrist and Pischke, 2008). The leads test includes $D_{i,t+1}$ into a linear regression and check whether a coefficient of $D_{i,t+1}$ is zero.

B.4.1 Repeated Cross-Sectional Data

In the repeated cross-sectional data setting, the leads test considers the following linear regression.

$$(\widehat{\theta}, \widehat{\gamma}, \widehat{\beta}, \widehat{\zeta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=0}^{1} O_{it} \left(Y_{it} - \theta G_i - \gamma_t - \beta D_{it} - \zeta D_{i,t+1} \right)^2.$$

Then, because $D_{it} = 0$ for all units in $t = \{0, 1\}$, this least squares problem is the same as

$$(\widehat{\theta}, \widehat{\gamma}, \widehat{\zeta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=0}^{1} O_{it} \left(Y_{it} - \theta G_i - \gamma_t - \zeta D_{i,t+1} \right)^2.$$

Then, using Result 4,

$$\widehat{\zeta} = \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right),$$

which is the standard DID estimator to the pre-treatment periods $t = 0, 1$. □

B.4.2 Panel Data

In the panel data setting, the leads test considers the following two-way fixed effects regression.

$$(\widehat{\alpha}, \widehat{\delta}, \widehat{\beta}, \widehat{\zeta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=0}^{1} (Y_{it} - \alpha_i - \delta_t - \beta D_{it} - \zeta D_{i,t+1})^2.$$

Again, this least squares problem is the same as

$$(\widehat{\alpha}, \widehat{\delta}, \widehat{\zeta}) = \text{argmin} \sum_{i=1}^{n} \sum_{t=0}^{1} (Y_{it} - \alpha_i - \delta_t - \zeta D_{i,t+1})^2.$$

Then, using Result 5,

$$\widehat{\zeta} = \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_1} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_1} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_0} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_0} \right),$$

which is the standard DID estimator to the pre-treatment periods $t = 0, 1$. □

C Generalized Double Difference-in-Differences

C.1 The Setup

Suppose we observe outcome Y_{it} for $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T^*, \ldots, T\}$. We define the binary treatment variable to be $D_{it} \in \{0, 1\}$. The treatment is assigned right after time period T^*, and thus, time periods $t \in \{T^* + 1, \ldots, T\}$ are the post-treatment periods and time periods $t \in \{1, \ldots, T^*\}$ are the pre-treatment periods. We denote the treatment group as $G_i = 1$ and $G_i = 0$ otherwise. Note that $D_{it} = 0$ for $t \in \{1, \ldots, T^*\}$ for all units. We focus on post-treatment period $T^* + s$ as the main target where $s > 0$. Therefore, our main quantity of interest is the average treatment effect on the treated (ATT) at time $T^* + s$.

$$\tau(s) \equiv \mathbb{E}[Y_{i,T^*+s}(1) - Y_{i,T^*+s}(0) \mid G_i = 1].$$

C.2 Assumptions

Here, we provide a generalization of the parallel trends assumption, which incorporates both the standard parallel trends assumption and the parallel trends-in-trends assumption.

Assumption 4 (*k*-th Order Parallel Trends) For some integer k such that $1 \leq k \leq T^*$,

$$\Delta_s^k \left(\mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = 1]\right) = \Delta_s^k \left(\mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = 0]\right),$$

where Δ_s^k is the k-th order difference operator defined recursively as follows.
When $k = 1$,

$$\Delta_s \left(\mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = g]\right) \equiv \mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = g] - \mathbb{E}[Y_{iT^*}(0) \mid G_i = g],$$

When $k = 2$,

$$\Delta_s^2 \left(\mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = g]\right) \equiv \Delta_s \left(\mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = g]\right) - s\Delta \left(\mathbb{E}[Y_{i,T^*}(0) \mid G_i = g]\right)$$

In general,

$$\Delta_s^k \left(\mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = g]\right)$$

$$\equiv \Delta_s^{k-1}\left(\mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = g]\right) - M_s^k \Delta^{k-1}\left(\mathbb{E}[Y_{i,T^*}(0) \mid G_i = g]\right),$$

$$= \mathbb{E}[Y_{i,T^*+s}(0) \mid G_i = g] - \mathbb{E}[Y_{iT^*}(0) \mid G_i = g] - \sum_{j=1}^{k-1} M_s^{j+1} \Delta^j \left(\mathbb{E}[Y_{i,T^*}(0) \mid G_i = g]\right),$$

where $M_s^\ell = \prod_{j=1}^{\ell-1}(s+j-1)/\prod_{j=1}^{\ell-1} j$ for $\ell \geq 2$. When $s = 1$ and $k = 1$, this assumption is equivalent to the standard parallel trends assumption (Assumption 1). The parallel-trends-in-trends assumption (Assumption 3) corresponds to a case when $s = 1$ and $k = 2$.

To understand Assumption 4, we can consider a simpler but stronger assumption. The k-th order parallel trends assumption (Assumption 4) is implied by

$$\mathbb{E}[Y_{it}(0) \mid G_i = 1] - \mathbb{E}[Y_{it}(0) \mid G_i = 0] = \alpha + \sum_{p=1}^{k-1} \Gamma_p t^p,$$

with arbitrary $(\alpha, \mathbf{\Gamma})$. This representation shows that the first order parallel trends assumption (the standard parallel trends assumption; Assumption 1) is implied by the time-invariant confounding; the second order parallel trends assumption (the parallel trends-in-trends assumption; Assumption 3) is implied by the linear time-varying confounding; and in general, the k-th order parallel trends assumption is implied by the k-th order polynomial confounding.

When $s = 1$, Assumption 4 is equivalent to the Parallel-(k) assumption in Mora and Reggio (2012, 2019). However, when $s > 1$, the two assumptions differ because our assumption only uses outcomes at the target time period $T^* + s$ and pre-treatment periods $t \leq T^*$. In contrast, the assumption in Mora and Reggio (2012, 2019) uses other post-treatment outcomes together and combines the identification of the ATT for multiple post-treatment outcomes. Our assumption allows us to clarify the identification assumption for each post-treatment period separately.

C.3 Identification and Estimation

Under the k-th order parallel trends assumption, the ATT is identified as follows.

$$\tau(s) = \Delta_s^k\left(\mathbb{E}[Y_{i,T^*+s} \mid G_i = 1]\right) - \Delta_s^k\left(\mathbb{E}[Y_{i,T^*+s} \mid G_i = 0]\right).$$

Because each conditional expectation can be consistently estimated via its sample analogue,

$$\widehat{\tau}_k(s) = \Delta_s^k\left(\frac{\sum_{i:\, G_i=1} Y_{i,T^*+s}}{n_{1,T^*+s}}\right) - \Delta_s^k\left(\frac{\sum_{i:\, G_i=0} Y_{i,T^*+s}}{n_{0,T^*+s}}\right)$$

is a consistent estimator for the ATT under the k-th order parallel trends assumption. When $s = 1$ and $k = 1$, this corresponds to the standard DID estimator (Equation (3)). When $s = 1$ and $k = 2$, this is equal to the sequential DID estimator (Equation (10)). While existing approaches (e.g., Angrist and Pischke, 2008; Mora and Reggio, 2012, 2019) consider each estimator separately, we propose combining multiple DID estimators within the GMM framework as follows.

The generalized double DID combines all T^* moment conditions (the total number of pre-treatment periods).

$$\widehat{\tau}(s) = \underset{\tau}{\operatorname{argmin}}\, \boldsymbol{g}(\tau)^\top \widehat{\mathbf{W}} \boldsymbol{g}(\tau) \tag{20}$$

where $\boldsymbol{g}(\tau) = (\tau - \widehat{\tau}_1(s), \ldots, \tau - \widehat{\tau}_{T^*}(s))^\top$. Based on the theory of the efficient GMM (Hansen, 1982), the optimal weight matrix is $\widehat{\mathbf{W}} = \operatorname{Var}(\widehat{\tau}_{(1:T^*)}(s))^{-1}$ where $\operatorname{Var}(\cdot)$ is the variance-covariance matrix and $\widehat{\tau}_{(1:T^*)}(s) = (\widehat{\tau}_1(s), \ldots, \widehat{\tau}_{T^*}(s))^\top$.

When $T^* = 1$, this converges to the standard DID estimator (Equation (3)). When $T^* = 2$, this corresponds to the basic form of the double DID estimator (Equation (14)). Within the GMM framework, we can select moment conditions using the J-statistics (Hansen, 1982). We can similarly generalize the double DID regression.

To assess the extended parallel trends assumption, we can apply the double DID with $(T^* - 1)$ moments to pre-treatment periods $t \in \{1, \ldots, T^*\}$ as if the last pre-treatment period T^* is the target time period. Moments are $\boldsymbol{g}(\tau) = (\tau - \widehat{\tau}_1(0), \ldots, \tau - \widehat{\tau}_{T^*-1}(0))^\top$ where $\widehat{\tau}_k(0) = \Delta^k \left(\frac{\sum_{i:\, G_i=1} Y_{iT^*}}{n_{1T^*}} \right) - \Delta^k \left(\frac{\sum_{i:\, G_i=0} Y_{iT^*}}{n_{0T^*}} \right)$. Similarly, to assess the extended parallel trends-in-trends assumption, we can apply the double DID with $(T^* - 2)$ moments to pre-treatment periods. Moments are $\boldsymbol{g}(\tau) = (\tau - \widehat{\tau}_2(0), \ldots, \tau - \widehat{\tau}_{T^*-1}(0))^\top$.

D Simulation Study

We provide details of the simulation setup and additional results for the simulation study.

D.1 Simulation Design

We consider the balanced panel data with $T = 5$ ($t = \{0, 1, 2, 3, 4\}$) where the last period ($t = 4$) is treated as the post-treatment period. We vary the number of units at

each time period as $n \in \{100, 250, 500, 1000\}$. Thus, the total number of observations are $nT \in \{500, 1250, 2500, 5000\}$. We compare three estimators: the double DID, the extended DID, and the sequential DID.

Note that we consider four pre-treatment periods here, and thus the generalized double DID is not equal to the sequential DID even under the parallel trends-in-trends assumption because it combines two other moments and optimally weight them (see Appendix C). The equivalence between the sequential DID and the double DID holds only when there are two pre-treatment periods. We see below that the generalized double DID improves upon the sequential DID even under the parallel trends-in-trends assumption as they optimally weight observations from different time periods.

We study two scenarios: one under the extended parallel trends assumption (Assumption 2) and the other under the parallel-trends-in-trends assumption (Assumption 3). In the first scenario, the difference between potential outcomes under control $\mathbb{E}[Y_{it}(0) \mid G_i = 1] - \mathbb{E}[Y_{it}(0) \mid G_i = 0]$ is constant over time. In particular, we set

$$\mathbb{E}[Y_{it}(0) \mid G_i = g] = \alpha_t + 0.05 \times g \tag{21}$$

where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (1, 2, 3, 4, 5)$. In the second scenario, we allow for linear time-varying confounding. In particular, we set

$$\mathbb{E}[Y_{it}(0) \mid G_i = g] = \alpha_t + 0.1 \times g \times (t + 1) \tag{22}$$

where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (1, 2, 3, 4, 5)$.

Then, potential outcomes under control are drawn as follows. $Y_{it}(0) = \mathbb{E}[Y_{it}(0) \mid G_i] + \epsilon_{it}$ where ϵ_{it} follows the AR(1) process with autocorrelation parameter ρ. That is,

$$\begin{aligned}
\epsilon_{it} &= \rho \epsilon_{i,t-1} + \xi_{it}, \\
\epsilon_{i0} &= \mathcal{N}(0, 3/(1 - \rho^2)), \\
\xi_{it} &= \mathcal{N}(0, 3).
\end{aligned}$$

The causal effect is denoted by τ and thus, $Y_{it}(1) = \tau + Y_{it}(0)$ where we set $\tau = 0.2$. Finally, $Y_{it} = Y_{it}(0)$ for $t \leq 3$ (pre-treatment periods) and $Y_{it} = G_i Y_{it}(1) + (1 - G_i) Y_{it}(0)$ for $t = 4$

(post-treatment period). The half of the samples are in the treatment group ($G_i = 1$) and the other half is in the control group ($G_i = 0$).

In Figure 4, we set the autocorrelation parameter $\rho = 0.6$. This value is similar to the autocorrelation parameter used in famous simulation studies in Bertrand et al. (2004) ($\rho = 0.8$). We pick a smaller value to make our simulations harder as we see below. In Figure 7, we also provide additional results where we consider a full range of the autocorrelation parameters $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$ (the same positive autocorrelation values considered in Bertrand et al. (2004)). Figure 4 shows the absolute bias and the standard errors which are defined as

$$\text{absolute bias} = \left| \frac{1}{M} \sum_{m=1}^{M} (\widehat{\tau}_m - \tau) \right| \quad \text{and} \quad \text{standard error} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\widehat{\tau}_m - \tau)^2},$$

where M is the total number of Monte Carlo iterations.

The first row of Figure 7 shows that our results on the (absolute) bias do not change regardless of the autocorrelation of errors. In particular, the double DID is unbiased under the extended parallel trends assumption (the first column) or under the parallel trends-in-trends assumption (the second column). In terms of the standard errors (the second row), two results are important. First, under the extended parallel trends assumption (the first column), the standard errors of the double DID is the smallest for all the values of ρ and the efficiency gain relative to the extended DID (i.e., two-way fixed effects estimator) is large when the there is high auto-correlations (i.e., ρ is large). Second, under the parallel trends-in-trends assumption (the second column), the standard errors of the double DID is the smallest among unbiased DID estimators (the extended DID is biased). The efficiency gain relative to the sequential DID is large when ρ is small.

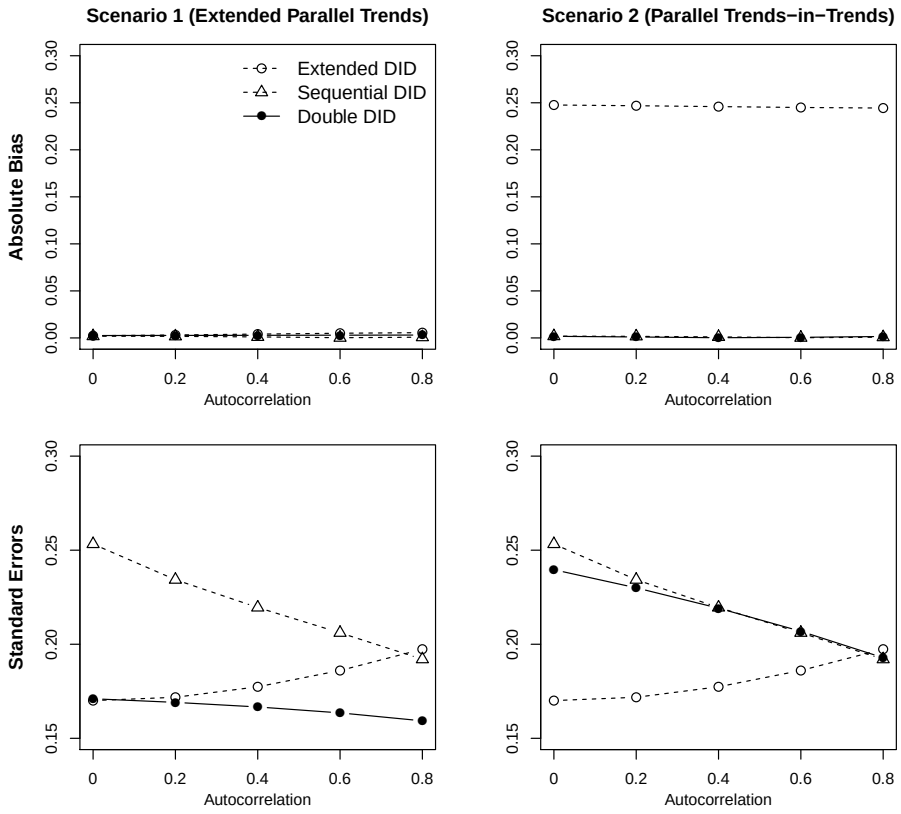

Figure 7: Comparing DID estimators in terms of the absolute bias and the standard errors according to the autocorrelation of errors. *Note*: The first row shows that the double DID estimator (black circle with solid line) is unbiased under both scenarios. The second row demonstrates that the double DID has the smallest standard errors among unbiased DID estimators. Under the extended parallel trends assumption (the first column), the efficiency gain relative to the extended DID (i.e., two-way fixed effects estimator) is large when the autocorrelation parameter ρ is large. Under the parallel trends-in-trends assumption (the second column), the efficiency gain relative to the sequential DID is large when ρ is small.

E Empirical Application

In Section 6, we focus on three outcomes to illustrate the advantage of the double DID estimator. In this section, we provide results for all the thirty outcomes. First, to assess the underlying parallel trends assumptions, we combine visualization and formal tests, as done in Section 6. In particular, we make the extended parallel trends assumption for fourteen outcomes of which treatment and control groups' pre-treatment trends have the same sign, and the 95% equivalence confidence interval is shorter than $[-0.10, 0.10]$. We rely on the parallel trends-in-trends assumption for six outcomes of which treatment and control groups' pre-treatment trends have the same sign, and the 95% equivalence confidence interval is wider than $[-0.10, 0.10]$. For the remaining ten outcomes of which treatment and control groups' pre-treatment trends have the opposite sign, it is difficult to justify either the extended parallel trends or parallel trends-in-trends assumption without additional information. Thus, there is no credible estimator for the ATT without making stronger assumptions. When there are more than two pre-treatment periods, researchers can apply the sequential DID estimator to pre-treatment periods in order to formally assess the extended parallel trends-in-trends assumption. We emphasize that, although we use the equivalence range of $[-0.10, 0.10]$ as a cutoff for an illustration, it is recommended to base this decision on substantive domain knowledge whenever possible in practice.

Table 2 shows results under the extended parallel trends assumption. As in Section 6, the double DID estimates are similar to those from the standard DID, and yet, standard errors are smaller because the double DID effectively uses pre-treatment periods within the GMM. Here, we only have two pre-treatment periods, but when there are more pre-treatment periods, the efficiency gain of the double DID becomes even larger. Table 3 shows results under the parallel trends-in-trends assumption. As in Section 6, the double DID estimates are often substantially different from those of the standard DID because the extended parallel trends assumption is implausible for these outcomes. Importantly, standard errors of the double DID are often larger than the standard DID. This is because the double DID needs to adjust for biases in the standard DID by using pre-treatment trends.

	Share of Households with Supported Crop		Share of Households with Agricultural Extension		Veterinarians		Share of Households Supported with Healthcare Fee		Public Health Project	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Standard DID	0.054	0.031	0.023	0.011	-0.043	0.026	0.036	0.023	0.083	0.039
Double DID	0.046	0.029	0.024	0.011	-0.040	0.023	0.035	0.020	0.072	0.036
	Education and Culture		**Share of Households Supported with Tuition fee**		**Upper Secondary School**		**Radio Broadcast**		**Post Office**	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Standard DID	0.084	0.056	0.001	0.003	0.002	0.037	0.062	0.049	0.046	0.031
Double DID	0.083	0.049	0.002	0.003	0.020	0.031	0.057	0.040	0.033	0.027
	Village with Post Office		**Share of Households with Credit Support**		**Share of Households with Business Tax Exemption**		**Village with Periodic Market**			
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error		
Standard DID	0.042	0.049	-0.054	0.022	0.001	0.008	0.001	0.045		
Double DID	0.038	0.042	-0.048	0.018	0.000	0.007	0.006	0.035		

Table 2: Comparing Standard DID and Double DID under Extended Parallel Trends Assumption. *Note:* The double DID estimates are similar to those from the standard DID, and yet, standard errors are smaller because the double DID effectively uses pre-treatment periods within the GMM.

	Public Transport		Development Project		Tap Water	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Standard DID	0.094	0.056	0.074	0.058	-0.098	0.054
Double DID	0.162	0.092	0.050	0.096	-0.148	0.072
	Staff to Support Crops		**Nonfarm Business**		**Village with Daily Market**	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Standard DID	-0.021	0.060	0.036	0.061	0.034	0.050
Double DID	-0.069	0.106	0.003	0.111	0.051	0.087

Table 3: Comparing Standard DID and Double DID under Parallel Trends-in-Trends Assumption. *Note:* The double DID estimates are often different from those of the standard DID because the extended parallel trends assumption is implausible for these outcomes.

References

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

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2017). Matrix Completion Methods for Causal Panel Data Models. Available at `https://arxiv.org/abs/1710.10251`.

Athey, S. and Imbens, G. W. (2018). Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption. National Bureau of Economic Research.

Bechtel, M. M. and Hainmueller, J. (2011). How Lasting is Voter Gratitude? An Analysis of the Short-and Long-Term Electoral Returns to Beneficial Policy. *American Journal of Political Science*, 55(4):852–868.

Beck, N. and Katz, J. N. (2011). Modeling Dynamics in Time-Series–Cross-Section Political Economy Data. *Annual Review of Political Science*, 14:331–352.

Ben-Michael, E., Feller, A., and Rothstein, J. (2018). The Augmented Synthetic Control Method. Available at `https://arxiv.org/abs/1811.04170`.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Blackwell, M. and Glynn, A. N. (2018). How To Make Causal Inferences With Time-Series Cross-Sectional Data Under Selection On Observables. *American Political Science Review*, 112(4):1067–1082.

Bullock, W. and Clinton, J. D. (2011). More a Molehill Than a Mountain: The Effects of the Blanket Primary on Elected Officials' Behavior from California. *The Journal of Politics*, 73(3):915–930.

De Boef, S. and Keele, L. (2008). Taking Time Seriously. *American Journal of Political Science*, 52(1):184–200.

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

Dube, A., Dube, O., and García-Ponce, O. (2013). Cross-Border Spillover: US Gun Laws and Violence in Mexico. *American Political Science Review*, 107(3):397–417.

Earle, J. S. and Gehlbach, S. (2015). The Productivity Consequences of Political Turnover: Firm-Level Evidence from Ukraine's Orange Revolution. *American Journal of Political Science*, 59(3):708–723.

Garfias, F. (2018). Elite Competition and State Capacity Development: Theory and Evidence from Post-Revolutionary Mexico. *American Political Science Review*, 112(2):339–357.

Goodman-Bacon, A. (2018). Difference-in-Differences with Variation in Treatment Timing. National Bureau of Economic Research.

Hall, A. B. (2016). Systemic Effects of Campaign Spending: Evidence From Corporate Contribution Bans in US State Legislatures. *Political Science Research and Methods*, 4(2):343–359.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029–1054.

Hartman, E. and Hidalgo, F. D. (2018). An Equivalence Approach to Balance and Placebo Tests. *American Journal of Political Science*, 62(4):1000–1013.

Hazlett, C. and Xu, Y. (2018). Trajectory Balancing: A General Reweighting Approach to Causal Inference with Time-Series Cross-Sectional Data. Available at SSRN: `https://ssrn.com/abstract=3214231`.

Imai, K. and Kim, I. S. (2019a). On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data. Available at `https://imai.fas.harvard.edu/research/files/FEmatch-twoway.pdf`.

Imai, K. and Kim, I. S. (2019b). When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science*, 63(2):467–490.

Imai, K., Kim, I. S., and Wang, E. (2019). Matching Methods for Causal Inference with Time-Series Cross-Sectional Data. Available at `https://imai.fas.harvard.edu/research/files/tscs.pdf`.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86.

Keele, L., Hasegawa, R., and Small, D. (2019a). Bracketing Bounds for Differences-in-Differences with an Application to Voter ID Laws. Presented at the 2019 Political Methodology Conference. Available at `https://polmeth.mit.edu/sites/default/files/documents/Keele_Paper.pdf`.

Keele, L. and Minozzi, W. (2013). How Much is Minnesota like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data. *Political Analysis*, 21(2):193–216.

Keele, L. J., Small, D. S., Hsu, J. Y., and Fogarty, C. B. (2019b). Patterns of Effects and Sensitivity Analysis for Differences-in-Differences. Available at `https://arxiv.org/abs/1901.01869`.

Ladd, J. M. and Lenz, G. S. (2009). Exploiting A Rare Communication Shift to Document the Persuasive Power of The News Media. *American Journal of Political Science*, 53(2):394–410.

Larreguy, H. and Marshall, J. (2017). The Effect of Education on Civic and Political Engagement in Nonconsolidated Democracies: Evidence from Nigeria. *Review of Economics and Statistics*, 99(3):387–401.

Malesky, E. J., Nguyen, C. V., and Tran, A. (2014). The Impact of Recentralization on Public Services: A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam. *American Political Science Review*, 108(1):144–168.

Mora, R. and Reggio, I. (2012). Treatment Effect Identification Using Alternative Parallel Assumptions. Working Paper 12-33 Economic Series (48), Universidad Carlos III. Available at `https://e-archivo.uc3m.es/bitstream/handle/10016/16065/we1233.pdf?sequence=1`.

Mora, R. and Reggio, I. (2019). Alternative Diff-in-Diffs Estimators with Several Pretreatment Periods. *Econometric Reviews*, 38(5):465–486.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with discussion). Section 9 (translated). *Statistical Science*, 5(4):465–472.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5):550–560.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688.

Strezhnev, A. (2018). Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs. Presented at the 2018 American Political Science Association Meeting.

Truex, R. (2014). The Returns to Office in a "Rubber Stamp" Parliament. *American Political Science Review*, 108(2):235–251.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Xu, Y. (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*, 25(1):57–76.