Using Multiple Pre-treatment Periods to Improve Difference-in-Differences and Staggered Adoption Designs*

Naoki Egami† Soichiro Yamauchi‡

This version: January 21, 2021
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 also generalize the double DID to the staggered adoption design where different units can receive the treatment in different time periods. We illustrate the proposed method with two empirical applications, covering both the basic DID and staggered adoption designs. We offer an open-source R package that implements the proposed methodologies.

*The methods proposed in this article can be implemented via the open-source statistical software R package `DIDdesign` available at `https://github.com/naoki-egami/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.

†Assistant Professor, Department of Political Science, Columbia University, New York NY 10027. Email: `naoki.egami@columbia.edu`; URL: `https://naokiegami.com`.

‡PhD Candidate, 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, Duflo and Mullainathan, 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.[1] However, in contrast to the basic two-time-period case, there are a number of different ways to analyze the DID design 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, Dube and García-Ponce, 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? In Section 2, we begin by examining three benefits of multiple pre-treatment periods using potential outcomes (Neyman, 1923; Rubin, 1974): (1) assessing the parallel trends assumption, (2) improving estimation accuracy, and (3) allowing for a more flexible parallel trends assumption. While these benefits have been discussed in the literature, we revisit them to clarify that each benefit requires different assumptions and different estimators. As a result, in practice, researchers tend to enjoy only a subset of the three benefits they can exploit from multiple pre-treatment periods. This methodological synthesis serves as a foundation to develop our new approach.

Our main contribution is to propose a new, simple estimator that achieves all the three benefits together. We use the generalized method of moments (GMM) framework (Hansen, 1982) to develop the *double difference-in-differences* (double DID). At its core, we combine two popular DID estimators: the standard DID estimator, which relies on the canonical parallel-

[1] In our literature review of *American Political Science Review* and *American Journal of Political Science* between 2015 and 2019, we found that about 63% of the papers that use the DID design have more than one pre-treatment period. See Appendix A for details about our literature review.

trends assumptions (Angrist and Pischke, 2008), and the sequential DID estimator (e.g., Lee, 2016; Mora and Reggio, 2019), which can allow for linear time-varying unmeasured confounding that changes over time at some fixed unknown rate. While each estimator itself is not new, the new combination of the two estimators via the GMM allows us to optimally exploit the three benefits of multiple pre-treatment periods in a way that previous methods could not.

First, we show that a test of pre-treatment-trends (the first benefit of the multiple pre-treatment trends) is equivalent to the over-identification test in the GMM framework. Therefore, we assess the underlying identification assumption as the first step of the GMM, and we clarify the exact identification assumption that the double DID estimator relies on. Second, we show that the proposed double DID estimator contains the two-way fixed effects estimator as a special case, which corresponds to a specific choice of the weight matrix in the GMM. By using the theory of the efficient GMM, we can choose the optimal weight matrix that increases estimation accuracy and reduces standard errors. Without assuming any constant treatment effects, our proposed estimator is consistent and more efficient than the two-way fixed effects estimator whenever the two-way fixed effects estimator is also consistent. Finally, when our first step of the pre-treatment-trend test finds that the parallel trends assumption is implausible, the double DID estimator reduces to the sequential DID estimator, which can unbiasedly estimates causal effects even under linear time-varying unmeasured confounding.

We then propose three extensions of our double DID estimator. First, we allow for any number of *pre*-treatment periods to further relax the parallel trends assumption (Section 3.4). This is important because researchers might be worried not only about time-varying unmeasured confounders that linearly change over time, but also about more general forms of time-varying unmeasured confounders. When there exist K pre-treatment periods, our proposed approach can accommodate time-varying unmeasured confounders that follow $(K-1)$th-order polynomial functions. Thus, we can account for more flexible forms of time-varying unmeasured confounding as we have more pre-treatment periods. Second, we also incorporate any number of *post*-treatment periods such that researchers can estimate not only short-term causal effects but also longer-term causal effects (Section 3.4). Finally, we generalize our double DID estimator to the staggered adoption design where different units can receive the treatment in different time periods (Section 4). This design is increasingly more popular in political science and in the social sciences in general (e.g., Athey and Imbens, 2018; Ben-Michael, Feller and Rothstein, 2019). We show that the double DID estimator achieves the same three benefits together even in this more general pattern of the treatment assignment.

We offer a companion R package DIDdesign that implements the proposed methods. We illustrate our proposed methods in Section 6 where we provide two empirical applications. The first application revisits Malesky, Nguyen and Tran (2014), which studies how the abolition of elected councils affects local public services. This serves as an example of the basic DID design where treatment assignment happens only once. Our second application is a reanalysis of Paglayan (2019), which examines the effect of granting collective bargaining rights to teacher's union on educational expenditures and teacher's salary. This is an example for the staggered adoption design where different units can receive the treatment in different time periods.

Related Literature. 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., Imai and Kim, 2011; Athey and Imbens, 2018; Goodman-Bacon, 2018; Strezhnev, 2018; Imai and Kim, 2019, 2020). 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. We do so both in the basic DID setup and in the staggered adoption design.

Our double DID estimator contains the sequential DID estimator (e.g., Lee, 2016; Mora and Reggio, 2019) as a special case. There are two key advantages over existing approaches. First, when the parallel trends assumption holds, the double DID optimally combines the standard DID and the sequential DID to improve efficiency, and thus is not equal to the sequential DID. Importantly, this avoids a dilemma of the sequential DID — it is consistent under assumptions weaker than the parallel trends, but is less efficient when the parallel trends assumption holds. Second, while the sequential DID has only been developed for the basic DID design where treatment assignment happens only once, we generalize it to the staggered adoption design, and further incorporate it into our proposed staggered-adoption double DID estimator.

Another class of popular methods is the synthetic control method (Abadie, Diamond and Hainmueller, 2010) and their recent extensions (e.g., Xu, 2017; Athey et al., 2017; Ben-Michael, Feller and Rothstein, 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, Diamond and Hainmueller, 2010); for example, Xu (2017) recommends collecting more than ten pre-treatment periods. In contrast, the proposed double DID can be applied as long as there are more than one pre-treatment period, and is better suited when there are a small to moderate number of pre-treatment periods.[2] However, we also show in Section 6.2 that the double DID can achieve performance comparable to variants of synthetic control methods even when there are a large number of pre-treatment periods. We offer additional discussions about relationships between our proposed approach and synthetic control methods in Section 5.3.

2 Three Benefits of Multiple Pre-treatment Periods

The difference-in-differences (DID) design is one of the most widely used methods to make causal inference from observational studies (Imbens and Wooldridge, 2009). At its most basic, the DID design consists of treatment and control groups measured at two time periods, before and after the treatment assignment. While this basic DID design only requires data from one post- and one pre-treatment period, additional pre-treatment periods are often available in applied contexts. Unfortunately, however, assumptions behind different uses of multiple 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. This section serves as a methodological foundation for developing a new approach in Sections 3 and 4.

2.1 The Setup for the Basic DID Design

As our running example, we focus on a study of how the abolition of elected councils affects local public services. Malesky, Nguyen and Tran (2014) use the DID design to examine the effect

[2] In our literature review of *APSR* and *AJPS*, we found that most DID applications have less than 10 pre-treatment periods. The median number of pre-treatment periods is 3.5 and, the mean number of pre-treatment periods is about 6 after removing one unique study that has more than 100 pre-treatment periods. See Appendix A for more details about our literature review.

of recentralization efforts in Vietnam. The abolition of elected councils, the main treatment of interest, was implemented in 2009 in about 12% of all the communes, which is the smallest administrative units that the paper considers. For each commune, a variety of outcomes related to public services, such as the quality of infrastructure, were measured in 2006, 2008 and 2010 — two pre-treatment periods and one post-treatment period. With this DID design, Malesky, Nguyen and Tran (2014) aim to estimate the causal effect of abolishing elected councils on various measures of local public services. To introduce the setup of the 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} denote the binary 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. In this section, we consider two pre-treatment time periods $t \in \{0, 1\}$ and one post-treatment period $t = 2$. We choose this setup here because it is sufficient for examining benefits of multiple pre-treatment periods, but we also generalize our discussions and methods to an arbitrary number of *pre-* and *post-*treatment periods (Section 3.4), and to the staggered adoption design (Section 4) for increasing applicability of our methods in practice. In our running example, two pre-treatment periods are 2006 and 2008, and one post-treatment period is 2010. Thus, the treatment group receives the treatment only at time $t = 2$; $D_{i0} = D_{i1} = 0$ and $D_{i2} = 1$, whereas units in the control group never gets treated $D_{i0} = D_{i1} = D_{i2} = 0$. We refer to the treatment group as $G_i = 1$ and the control group as $G_i = 0$. Outcome Y_{it} is measured at time $t \in \{0, 1, 2\}$. 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 three 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. $Y_{it}(0)$ is similarly defined. 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 time $t = 2$, $Y_{i2}(1) - Y_{i2}(0)$.

In the DID design, we are interested in 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 who receive the treatment.

DID with One Pre-Treatment Period

Before we discuss benefits of multiple pre-treatment periods from Section 2.2, we briefly review the DID with one pre-treatment period to fix ideas for settings with multiple pre-treatment periods.

In the most basic DID design where we have only one pre-treatment period (i.e., $t = 1$ is the pre-treatment period and $t = 2$ is the post-treatment period), researchers can identify the ATT based on the widely-used assumption of *parallel trends* — if the treatment group had not received the treatment in the second period, its outcome trend would have been the same as the trend of the outcome in the control group. (Angrist and Pischke, 2008).

Assumption 1 (Parallel Trends)**.**

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

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 the difference-in-differences estimator.

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

where n_{1t} and n_{0t} are the number of units in the treatment and control groups at time $t \in \{1, 2\}$, respectively.

In practice, we can compute the DID estimator via a linear regression. We regress the 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 $G_i \times I_t$.

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

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, $\widehat{\tau}_{\texttt{DID}}$. 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 such as constant treatment effects. 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). The above equivalence is formally shown in Appendix B.1 for completeness.

2.2 Benefit 1: Assessing Parallel Trends Assumption

We now consider how researchers can exploit multiple pre-treatment periods, while clarifying underlying necessary assumptions.

The first and the most common use of multiple 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 is implemented in 2009 — are the same for treatment and control communes.

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

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

We then check whether the DID estimate on pre-treatment periods is statistically distinguishable from zero. For example, we can apply the DID estimator to 2006 and 2008 as if 2008 were the post-treatment period, and assess whether the estimate would be close to zero. In Figure 1, a DID estimate on the pre-treatment periods would be close to zero for the left panel, while it would be negative for the right panel where two groups have different pre-treatment trends. In Appendix B.4, we show that a robustness check about leads effects (Angrist and Pischke, 2008), which incorporates 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 the 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

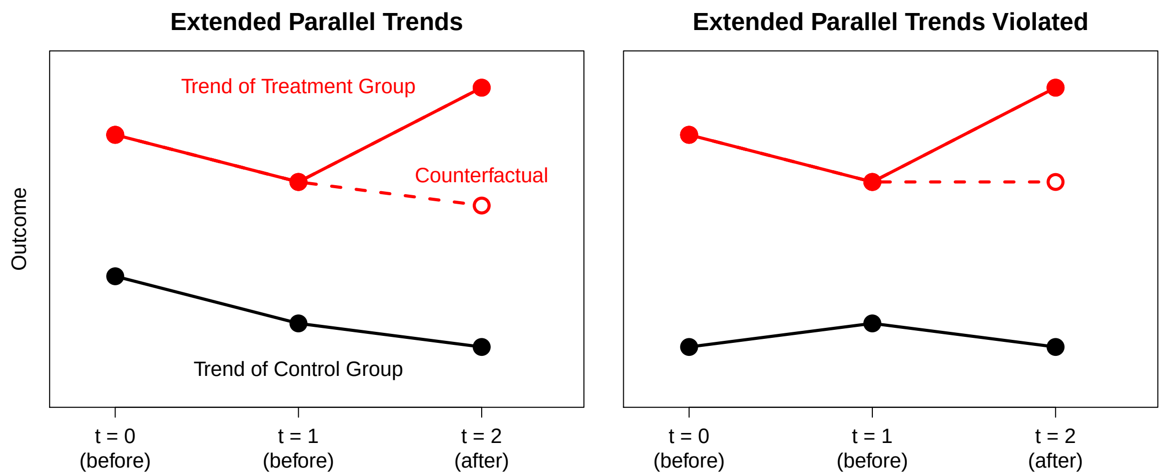

Figure 1: Parallel Pre-treatment Trends (left) and Non-Parallel Pre-treatment Trends (right).

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)**.**

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

Here, 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. 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 implies the conventional parallel trends assumption, Assumption 2 is also sufficient for identifying the ATT, and we can estimate it via the same DID estimator (equation (3)).

In summary, the first benefit is that researchers can assess the extended parallel trends assumption using the pre-treatment-trends test (equation (5)).

2.3 Benefit 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, 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.

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

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.

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

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.

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

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

In the panel data settings, we show that this extended DID estimator $\widehat{\tau}_{\texttt{e-DID}}$ is numerically equivalent to a coefficient of the treatment variable in the two-way fixed effects estimator fitted to the three time periods $t \in \{0, 1, 2\}$. We also present more general results about nonparametric relationships between the extended DID and the two-way fixed effects estimator in Appendix B.2.

To summarize, the second benefit of multiple pre-treatment periods is that researchers can use the extended DID estimator (equivalent to the two-way fixed effects estimator in the panel data) to increase statistical efficiency when the extended parallel trends assumption holds.

2.4 Benefit 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. 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 (e.g., Lee, 2016; Mora and Reggio, 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 2.2, 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. In contrast, when trends of treatment and control groups are not parallel, a DID estimate on pre-treatment periods would be non-zero. 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 $t = 2$.

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

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.

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, 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)**.**

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

Here, 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 2 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 replace the outcome Y_{it} with the transformed outcome that adjusts for the lagged

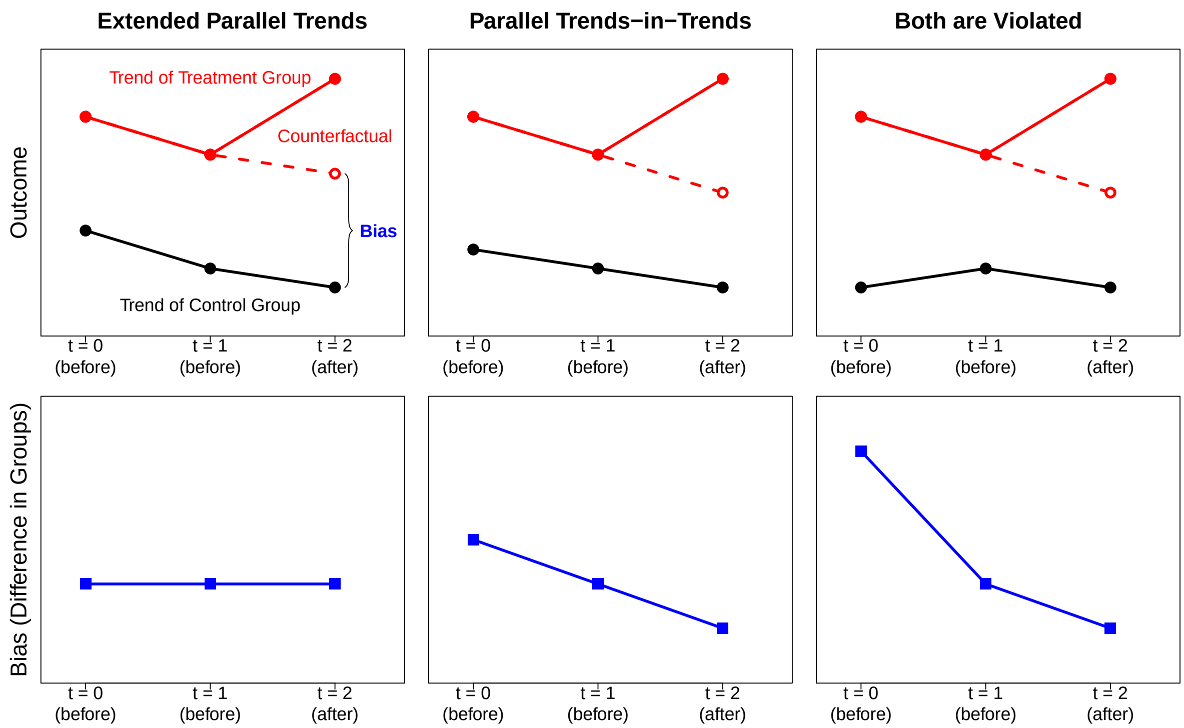

Figure 2: Comparing Extended Parallel Trends Assumption and Parallel Trends-in-Trends Assumption. *Note:* The extended parallel trends assumption (left column) means that the difference in the treatment and control groups (bias) is constant over time. The parallel trends-in-trends assumption (middle column) allows for linear time-varying unmeasured confounding. Both assumptions are violated in the right column.

outcome in a particular way.

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

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 (see Appendix B.3). Within the potential outcomes framework, we clarified that these common techniques are justified under the parallel trends-in-trends assumption.

In summary, the third benefit of multiple pre-treatment periods is that, researchers can use the sequential DID estimator (equation (11)) under the more flexible, parallel trends-in-trends assumption, even when the extended parallel trends assumption is violated (therefore, the two-way fixed effects estimator is biased).

Remark. 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 can account for the $K-1$ degrees of polynomial functions of unmeasured confounders. 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., 2019) and the bracketing bounds (Angrist and Pischke, 2008; Ding and Li, 2019; Ye et al., 2020). □

3 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. We have clarified that each benefit requires different assumptions and different estimators, and as a result, in practice, researchers tend to enjoy only a subset of the three benefits they can exploit from multiple pre-treatment periods. In this section, we propose a new, simple estimator, which we call the *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 also provide two extensions. First, we generalize the proposed method to any number of *pre*- and *post*-treatment periods in the DID design (Section 3.4 and Appendix C). Second, we then extend it to the staggered adoption design, where timing of the treatment assignment can vary across units, in Section 4.

3.1 Double DID via GMM

We propose the double DID estimator within a framework of the generalized method of moments (GMM) (Hansen, 1982). In particular, we combine the standard DID estimator and the sequential DID estimator via the GMM:

$$\widehat{\tau}_{\texttt{d-DID}} = \operatorname*{argmin}_{\tau} \begin{pmatrix} \tau - \widehat{\tau}_{\texttt{DID}} \\ \tau - \widehat{\tau}_{\texttt{s-DID}} \end{pmatrix}^{\top} \mathbf{W} \begin{pmatrix} \tau - \widehat{\tau}_{\texttt{DID}} \\ \tau - \widehat{\tau}_{\texttt{s-DID}} \end{pmatrix} \tag{13}$$

	Standard DID	Extended DID	Sequential DID
Weight Matrix **W**	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$

Table 1: Double DID as Generalization of Popular DID Estimators.

where $\mathbf{W}$ is a user-specified weight matrix of dimension 2×2.

The important property of the proposed double DID estimator is that it contains all of the popular estimators that we considered in the previous sections as special cases. By specifying an appropriate weight matrix $\mathbf{W}$, we can recover the standard DID, the extended DID, and the sequential DID estimators (see Table 1).

The advantage of the double DID estimator is that we can choose the optimal weight matrix by considering the identification assumption and estimation within the unifying framework of the GMM. The double DID estimator proceeds with the following two steps.

Step 1: Assessing the Underlying Assumptions

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)). Importantly, this first step of the double DID is equivalent to the over-identification 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.

Equivalence Approach. The standard hypothesis testing approach has a risk of conflating evidence for parallel trends and statistical inefficiency. For example, when sample size is small, even if pre-treatment trends of the treatment and control groups differ, a test of the difference might not be statistically significant due to large standard error, and analysts might "pass" the pre-treatment-trends test. To mitigate such concerns, 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.[3] By us-

[3] Liu, Wang and Xu (2020) propose a similar test for a different class of estimators, what they refer to as "counterfactual estimators," which share many properties with synthetic control methods. We discuss relation-

ing this approach, researchers can "pass" the pre-treatment-trends test only when estimated pre-treatment trends of the two groups are similar with high accuracy, thereby avoiding the aforementioned common mistake.

To facilitate the interpretation of the equivalence confidence interval, we report the standardized interval, which can be interpreted as the standard deviation from the baseline control mean. We provide technical details in Appendix E and provide an empirical example in Section 6.

Step 2: Estimation of the ATT

The second step is the estimation of the ATT. When the extended parallel trends assumption is plausible, we can estimate the optimal weight matrix $\mathbf{W}$ building on the theory of the efficient GMM (Hansen, 1982). Specifically, the optimal weight matrix that minimizes the variance is the inverse of the variance-covariance matrix of the DID estimators:

$$\widehat{\mathbf{W}} = \begin{pmatrix} \widehat{\text{Var}}(\widehat{\tau}_{\texttt{DID}}) & \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}}) \\ \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}}) & \widehat{\text{Var}}(\widehat{\tau}_{\texttt{s-DID}}) \end{pmatrix}^{-1} \tag{14}$$

This optimal weight matrix allows us to compute the weighted average of the standard DID and the sequential DID estimator such that the resulting variance is the smallest.

Remark. Importantly, under the extended parallel trends assumption, both the standard DID and the sequential DID estimator are consistent to the ATT, and thus, any weighted average is a consistent estimator. But this optimal weight matrix chooses the most efficient estimator among all consistent estimators. As we clarify more below, we do not use the weighted average of the standard DID and the sequential DID when the extended parallel trends assumption is violated. □

When only the parallel trends-in-trends assumption is plausible, the double DID only contains one moment condition $\tau - \widehat{\tau}_{\texttt{s-DID}} = 0$, and thus it is equal to the sequential DID estimator. This is equivalent to choosing the weight matrix $\mathbf{W}$ with $W_{11} = W_{12} = W_{21} = 0$ and $W_{22} = 1$ (the third column in Table 1).

When both assumptions are implausible, there is no credible estimator without making further stringent assumptions. However, when there are more than two pre-treatment periods, researchers can also use the proposed generalized K-DID (we will discuss in Section 3.4) to further relax the parallel trends-in-trends assumption.

ships between our method and synthetic control methods in Section 5.3.

3.2 Double DID Enjoys Three Benefits

The proposed 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 over-identification 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 2.3). In this setting, the double DID further improves estimation accuracy because it selects the optimal weights as the GMM estimator. In Section F, 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 the weight matrix in the GMM framework, 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.

3.3 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 2.1, 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:

$$Y_{it} \sim \alpha + \theta G_i + \gamma I_t + \beta(G_i \times I_t) + \mathbf{X}_{it}^\top \boldsymbol{\rho}, \tag{15}$$

where we adjust for the additional pre-treatment covariates $\mathbf{X}_{it}$. A coefficient of the interaction term $\widehat{\beta}$ is the standard DID regression estimator for the ATT. Here, we make the parallel trends assumption *conditional* on 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 conditionally 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:

$$\Delta Y_{it} \sim \alpha_s + \theta_s G_i + \gamma_s I_t + \beta_s(G_i \times I_t) + \mathbf{X}_{it}^\top \boldsymbol{\rho}_s, \tag{16}$$

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:

$$\widehat{\beta}_{\texttt{d-DID}} = \underset{\beta_d}{\operatorname{argmin}} \begin{pmatrix} \beta_d - \widehat{\beta} \\ \beta_d - \widehat{\beta}_s \end{pmatrix}^\top \mathbf{W} \begin{pmatrix} \beta_d - \widehat{\beta} \\ \beta_d - \widehat{\beta}_s \end{pmatrix} \tag{17}$$

where $\mathbf{W}$ is a user-specified weight matrix of dimension 2×2.

Thus, as the double DID estimator without covariates, 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 (17)).

3.4 Generalized K-Difference-in-Differences

We generalize the proposed method to any number of *pre*- and *post*-treatment periods in Appendix C. This generalization has two central benefits. First, it enables researchers to use longer *pre*-treatment periods to allow for even more flexible forms of unmeasured time-varying confounding. While the parallel trends-in-trends assumption (Assumption 3) accounts for linear time-varying unmeasured confounding, we can allow for time-varying unmeasured confounding that follows a $(K-1)$th order polynomial function when we have K pre-treatment periods. Second, we also allow for any number of *post*-treatment periods, and therefore, researchers can estimate not only short-term causal effects but also longer-term causal effects with this generalization.

4 Staggered Adoption Design

In this section, we extend the proposed double DID estimator to the staggered adoption design where timing of the treatment assignment can vary across units (Athey and Imbens, 2018; Strezhnev, 2018; Ben-Michael, Feller and Rothstein, 2019).

4.1 The Setup and Causal Quantities of Interest

In the staggered adoption (SA) design, different units can receive the treatment in different time periods. Once they receive the treatment, they remain exposed to the treatment afterwards. Therefore, $D_{it} = 1$ if $D_{im} = 1$ where $m < t$. We can thus summarize the information about the treatment assignment by the timing of the treatment A_i where $A_i \equiv \min \{t : D_{it} = 1\}$. When unit i never receives the treatment until the end of time T, we let $A_i = \infty$. For example, in many applications where researchers are interested in the causal effect of state- or local-level policies, units adopt policies in different time points and remain exposed to such policies once they introduce the policies. In Section 6.2, we provide its example based on Paglayan (2019). See Figure 3 for visualization of the SA design.

Following the recent literature on the SA design, we make two standard assumptions in the SA design: no anticipation assumption and invariance to history assumption (Athey and Imbens, 2018; Abadie, 2019; Ben-Michael, Feller and Rothstein, 2019; Imai and Kim, 2019). This implies that, for unit i in period t, the potential outcome $Y_{it}(1)$ represents the outcome of unit i that would realize in period t if unit i receives the treatment at or before period t. Similarly, $Y_{it}(0)$ represents the outcome of unit i that would realize in period t if unit i does

Year

	1997	1998	1999	2000	2001	2002	2003
State 1	0	0	1	1	1	1	1
State 2	0	0	0	0	1	1	1
State 3	0	0	0	0	0	0	0

Figure 3: Example of the Staggered Adoption Design. *Note:* We use gray cells of "1" to denote the treated observation and use white cells of "0" to denote the control observation. The main feature of the SA design is that once units receive the treatment, they remain exposed to the treatment.

not receive the treatment by period t. Finally, we generalize the group indicator G as follows.

$$G_{it} = \begin{cases} 1 & \text{if } A_i = t \\ 0 & \text{if } A_i > t \\ -1 & \text{if } A_i < t \end{cases} \tag{18}$$

where $G_{it} = 1$ represents units who receive the treatment at time t, and $G_{it} = 0$ ($G_{it} = -1$) indicates units who receive the treatment after (before) time t.

Under the SA design, the *staggered adoption ATT* (SA-ATT) at time t is defined as follows.

$$\tau^{\mathsf{SA}}(t) = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_{it} = 1],$$

which represents the causal effect of the treatment in period t on units with $G_{it} = 1$, who receive the treatment at time t. This is a straightforward extension of the standard ATT (equation (1)) in the basic DID setting. Researchers might also be interested in the *time-average staggered adoption ATT* (time-average SA-ATT).

$$\overline{\tau}^{\mathsf{SA}} = \sum_{t \in \mathcal{T}} \pi_t \tau^{\mathsf{SA}}(t),$$

where $\mathcal{T}$ represents a set of the time periods for which researchers want to estimate the ATT. For example, if a researcher is interested in estimating the ATT for the entire sample periods, one can take $\mathcal{T} = \{1, \dots, T\}$. The SA-ATT in period t, $\tau^{\mathsf{SA}}(t)$, is weighted by the proportion of units who receive the treatment at time t: $\pi_t = \sum_{i=1}^n \mathbf{1}\{A_i = t\} / \sum_{i=1}^n \mathbf{1}\{A_i \in \mathcal{T}\}$.

4.2 Double DID for Staggered Adoption Design

Under what assumptions can we identify the SA-ATT and the time-average SA-ATT? Here, we first extend the standard DID estimator under the parallel trends assumption and the sequential DID estimator under the parallel trends-in-trends assumption to the SA design. Formally, we define the standard DID estimator for the SA-ATT at time t as

$$\widehat{\tau}^{\mathsf{SA}}_{\texttt{DID}}(t) = \left(\frac{\sum_{i:\,G_{it}=1} Y_{it}}{n_{1t}} - \frac{\sum_{i:\,G_{it}=1} Y_{i,t-1}}{n_{1,t-1}}\right) - \left(\frac{\sum_{i:\,G_{it}=0} Y_{it}}{n_{0t}} - \frac{\sum_{i:\,G_{it}=0} Y_{i,t-1}}{n_{0,t-1}}\right),$$

which is consistent for the SA-ATT under the following parallel trends assumption in period t under the SA design:

$$\mathbb{E}[Y_{it}(0) \mid G_{it} = 1] - \mathbb{E}[Y_{i,t-1}(0) \mid G_{it} = 1] \ = \ \mathbb{E}[Y_{it}(0) \mid G_{it} = 0] - \mathbb{E}[Y_{i,t-1}(0) \mid G_{it} = 0].$$

Similarly, we can define the sequential DID estimator for the SA-ATT at time t as

$$\begin{aligned}\widehat{\tau}^{\mathsf{SA}}_{\texttt{s-DID}}(t) = &\left\{\left(\frac{\sum_{i:\,G_{it}=1} Y_{it}}{n_{1t}} - \frac{\sum_{i:\,G_{it}=1} Y_{i,t-1}}{n_{1,t-1}}\right) - \left(\frac{\sum_{i:\,G_{it}=0} Y_{it}}{n_{0t}} - \frac{\sum_{i:\,G_{it}=0} Y_{i,t-1}}{n_{0,t-1}}\right)\right\} \\ &- \left\{\left(\frac{\sum_{i:\,G_{it}=1} Y_{i,t-1}}{n_{1,t-1}} - \frac{\sum_{i:\,G_{it}=1} Y_{i,t-2}}{n_{1,t-2}}\right) - \left(\frac{\sum_{i:\,G_{it}=0} Y_{i,t-1}}{n_{0,t-1}} - \frac{\sum_{i:\,G_{it}=0} Y_{i,t-2}}{n_{0,t-2}}\right)\right\},\end{aligned}$$

which is consistent for the SA-ATT under the following parallel trends-in-trends assumption in period t under the SA design:

$$\begin{aligned}&\{\mathbb{E}[Y_{it}(0) \mid G_{it} = 1] - \mathbb{E}[Y_{it}(0) \mid G_{it} = 0]\} - \{\mathbb{E}[Y_{i,t-1}(0) \mid G_{it} = 1] - \mathbb{E}[Y_{i,t-1}(0) \mid G_{it} = 0]\} \\ &= \{\mathbb{E}[Y_{i,t-1}(0) \mid G_{it} = 1] - \mathbb{E}[Y_{i,t-1}(0) \mid G_{it} = 0]\} - \{\mathbb{E}[Y_{i,t-2}(0) \mid G_{it} = 1] - \mathbb{E}[Y_{i,t-2}(0) \mid G_{it} = 0]\}.\end{aligned}$$

Finally, combining the standard and sequential DID estimators, we can extend the double DID to the SA design as follows.

$$\widehat{\tau}^{\mathsf{SA}}_{\texttt{d-DID}}(t) = \underset{\tau^{\mathsf{SA}}(t)}{\operatorname{argmin}} \begin{pmatrix} \tau^{\mathsf{SA}}(t) - \widehat{\tau}^{\mathsf{SA}}_{\texttt{DID}}(t) \\ \tau^{\mathsf{SA}}(t) - \widehat{\tau}^{\mathsf{SA}}_{\texttt{s-DID}}(t) \end{pmatrix}^{\top} \mathbf{W}(t) \begin{pmatrix} \tau^{\mathsf{SA}}(t) - \widehat{\tau}^{\mathsf{SA}}_{\texttt{DID}}(t) \\ \tau^{\mathsf{SA}}(t) - \widehat{\tau}^{\mathsf{SA}}_{\texttt{s-DID}}(t) \end{pmatrix}$$

where $\mathbf{W}(t)$ is a user-specified weight matrix. Under the SA design as well, the standard DID and sequential DID estimators are special cases of our proposed double DID estimator with specific weight matrices.

Like the basic double DID estimator that we have proposed in Section 3.1, the double DID for the SA design also has two steps. The first step is to assess the underlying assumptions using the standard DID for the SA design with two time points $\{t-1, t-2\}$ for units who are not yet treated at time $t-1$, that is, $\{i : G_{it} \geq 0\}$. This is a generalization of the pre-treatment-trends test in the basic DID setup (Section 2.2). The second step is to estimate the SA-ATT at time t. When only the parallel trends-in-trends assumption is plausible, we choose weight matrix $\mathbf{W}(t)$ where $\mathbf{W}(t)_{11} = \mathbf{W}(t)_{12} = \mathbf{W}(t)_{21} = 0$ and $\mathbf{W}(t)_{22} = 1$, which converges to the sequential DID under the SA design. When the extended parallel trends assumption is plausible, we use the optimal weight matrix defined as $\mathbf{W}(t) = \text{Var}(\widehat{\tau}^{\mathsf{SA}}_{(1:2)}(t))^{-1}$ where $\text{Var}(\cdot)$ is the variance-covariance matrix and $\widehat{\tau}^{\mathsf{SA}}_{(1:2)}(s) = (\widehat{\tau}^{\mathsf{SA}}_{\mathtt{DID}}(t), \widehat{\tau}^{\mathsf{SA}}_{\mathtt{s\text{-}DID}}(t))^\top$. This optimal weight matrix provides us with the most efficient estimator (i.e., the smallest standard error). We provide further details on the implementation in Appendix D.

To estimate the time-average SA-DID, we extend the double DID as follows.

$$\widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{d\text{-}DID}} = \underset{\overline{\tau}^{\mathsf{SA}}}{\text{argmin}} \begin{pmatrix} \overline{\tau}^{\mathsf{SA}} - \widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{DID}} \\ \overline{\tau}^{\mathsf{SA}} - \widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{s\text{-}DID}} \end{pmatrix}^\top \overline{\mathbf{W}} \begin{pmatrix} \overline{\tau}^{\mathsf{SA}} - \widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{DID}} \\ \overline{\tau}^{\mathsf{SA}} - \widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{s\text{-}DID}} \end{pmatrix}$$

where

$$\widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{DID}} = \sum_{t \in \mathcal{T}} \pi_t \widehat{\tau}^{\mathsf{SA}}_{\mathtt{DID}}(t), \quad \text{and} \quad \widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{s\text{-}DID}} = \sum_{t \in \mathcal{T}} \pi_t \widehat{\tau}^{\mathsf{SA}}_{\mathtt{s\text{-}DID}}(t).$$

The optimal weight matrix $\overline{\mathbf{W}}$ is equal to $\overline{\mathbf{W}} = \text{Var}(\widehat{\overline{\tau}}^{\mathsf{SA}}_{(1:2)})^{-1}$ where $\widehat{\overline{\tau}}^{\mathsf{SA}}_{(1:2)} = (\widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{DID}}, \widehat{\overline{\tau}}^{\mathsf{SA}}_{\mathtt{s\text{-}DID}})^\top$.

4.3 Double DID Regression for Staggered Adoption Design

We now extend the double DID regression to the SA design setting. We first extend the standard DID regression (Section 3.3) to the SA design. In particular, to estimate the SA-ATT at time t, we can fit the following regression for units who are not yet treated at time $t-1$, that is, $\{i : G_{it} \geq 0\}$.

$$Y_{iv} \sim \alpha + \theta G_{it} + \gamma I_v + \beta^{\mathsf{SA}}(t)(G_{it} \times I_v) + \mathbf{X}_{iv}^\top \boldsymbol{\rho},$$

where $v \in \{t-1, t\}$ and the time indicator I_v (equal to 1 if $v = t$ and 0 if $v = t-1$). Note that G_{it} defines the treatment and control group at time t, and thus, it does not depend on time index v. The estimated coefficient $\widehat{\beta}^{\mathsf{SA}}(t)$ is consistent for the SA-ATT under the conditional parallel trends assumption.

Similarly, we can extend the sequential DID regression to the SA design. Using the connection to the linear regression of equation (12), we can adjust for additional pre-treatment covariates as:

$$\Delta Y_{iv} \sim \alpha_s + \theta_s G_{it} + \gamma_s I_v + \beta^{\mathsf{SA}}_s(t)(G_{it} \times I_v) + \mathbf{X}_{iv}^\top \boldsymbol{\rho}_s,$$

where $v \in \{t-1, t\}$ and $\Delta Y_{iv} = Y_{iv} - (\sum_{i:\, G_{it}=1} Y_{i,v-1})/n_{1,v-1}$ if $G_{it} = 1$ and $\Delta Y_{iv} = Y_{iv} - (\sum_{i:\, G_{it}=0} Y_{i,v-1})/n_{0,v-1}$ if $G_{it} = 0$. The estimated coefficient $\widehat{\beta}_s^{\mathsf{SA}}(t)$ is consistent for the SA-ATT under the conditional parallel trends-in-trends assumption.

Therefore, the double DID regression for the SA design combines the two regression estimators via the GMM:

$$\widehat{\beta}^{\mathsf{SA}}_{\texttt{d-DID}}(t) = \underset{\beta_d^{\mathsf{SA}}(t)}{\operatorname{argmin}} \begin{pmatrix} \beta_d^{\mathsf{SA}}(t) - \widehat{\beta}^{\mathsf{SA}}(t) \\ \beta_d^{\mathsf{SA}}(t) - \widehat{\beta}_s^{\mathsf{SA}}(t) \end{pmatrix}^{\top} \mathbf{W}(t) \begin{pmatrix} \beta_d^{\mathsf{SA}}(t) - \widehat{\beta}^{\mathsf{SA}}(t) \\ \beta_d^{\mathsf{SA}}(t) - \widehat{\beta}_s^{\mathsf{SA}}(t). \end{pmatrix}$$

where the choice of the weight matrix follows the same two-step procedure as Section 4.2. We also provide further details in Appendix D. The optimal weight matrix $\mathbf{W}(t)$ is equal to $\mathrm{Var}(\widehat{\beta}^{\mathsf{SA}}_{(1:2)})^{-1}$ where $\widehat{\beta}^{\mathsf{SA}}_{(1:2)} = (\widehat{\beta}^{\mathsf{SA}}(t), \widehat{\beta}_s^{\mathsf{SA}}(t))^{\top}$.

To estimate the time-average SA-ATT, we extend the double DID regression as follows.

$$\widehat{\overline{\beta}}^{\mathsf{SA}}_{\texttt{d-DID}} = \underset{\overline{\beta}_d^{\mathsf{SA}}}{\operatorname{argmin}} \begin{pmatrix} \overline{\beta}_d^{\mathsf{SA}} - \widehat{\overline{\beta}}^{\mathsf{SA}} \\ \overline{\beta}_d^{\mathsf{SA}} - \widehat{\overline{\beta}}_s^{\mathsf{SA}} \end{pmatrix}^{\top} \overline{\mathbf{W}} \begin{pmatrix} \overline{\beta}_d^{\mathsf{SA}} - \widehat{\overline{\beta}}^{\mathsf{SA}} \\ \overline{\beta}_d^{\mathsf{SA}} - \widehat{\overline{\beta}}_s^{\mathsf{SA}} \end{pmatrix}$$

where

$$\widehat{\overline{\beta}}^{\mathsf{SA}} = \sum_{t \in \mathcal{T}} \pi_t \widehat{\beta}^{\mathsf{SA}}(t), \quad \text{and} \quad \widehat{\overline{\beta}}_s^{\mathsf{SA}} = \sum_{t \in \mathcal{T}} \pi_t \widehat{\beta}_s^{\mathsf{SA}}(t).$$

The optimal weight matrix $\overline{\mathbf{W}}$ is equal to $\mathrm{Var}(\widehat{\overline{\beta}}^{\mathsf{SA}}_{(1:2)})^{-1}$ where $\widehat{\overline{\beta}}^{\mathsf{SA}}_{(1:2)} = (\widehat{\overline{\beta}}^{\mathsf{SA}}, \widehat{\overline{\beta}}_s^{\mathsf{SA}})^{\top}$.

5 Discussion

This section clarifies relationships between our proposed double DID and three existing methods: the two-way fixed effects estimator, the sequential DID estimator, and synthetic control methods.

5.1 Relationship with Two-Way Fixed Effects Estimator

While we contrast the double DID with the two-way fixed effects estimator throughout the paper, we summarize our discussion here. First, in the basic DID design, the two-way fixed effects estimator is a special case of the double DID with a specific choice of the weight matrix $\mathbf{W}$ (see Table 1). Therefore, whenever the two-way fixed effects estimator is consistent for the ATT, the double DID is a more efficient, consistent estimator of the ATT. This is because the double DID can choose the optimal weight matrix via the GMM, while the two-way fixed effects uses the pre-determined equal weights over time. Second, in the SA design, a large number of recent papers show that the widely-used two-way fixed effects estimator are in general inconsistent for the ATT due to treatment effect heterogeneity and implicit parametric assumptions (Abraham and Sun, 2018; Athey and Imbens, 2018; Strezhnev, 2018; Imai and Kim, 2020). In contrast, the proposed double DID in the SA design generalizes nonparametric DID estimators to allow for treatment effect heterogeneity, and thus, it does not suffer from the same problem.

5.2 Relationship with Sequential DID Estimator

Our double DID estimator contains the sequential DID estimator (e.g., Lee, 2016; Mora and Reggio, 2019) as a special case. Our proposed double DID improves over the sequential DID estimator in two ways. First, when the parallel trends assumption holds, the double DID optimally combine the standard DID and the sequential DID to improve efficiency, and it is not equal to the sequential DID. Therefore, it avoids a dilemma of the sequential DID — it is consistent under the parallel trends-in-trends assumption (weaker than the parallel trends assumption), but is less efficient when the parallel trends assumption holds. Second, while the sequential DID estimator has only been available for the basic DID design where treatment assignment happens only once, we generalize it to the staggered adoption design and further incorporate it into our staggered-adoption double DID estimator (Section 4).

5.3 Relationship with Synthetic Control Methods

Another relevant popular class of methods is the synthetic control methods. While the method was originally designed to estimate the causal effect on a *single* treated unit, recent extensions allow for multiple treated units and the staggered adoption design (e.g., Xu, 2017; Athey et al., 2017; Ben-Michael, Feller and Rothstein, 2018; Hazlett and Xu, 2018). Despite a wide variety of innovative extensions, they all share the same core feature: they require long pre-treatment periods to accurately estimate a pre-treatment trajectory of the treated units. For example, Xu (2017) recommends collecting more than ten pre-treatment periods. In contrast, the proposed double DID can be applied as long as there are more than one pre-treatment periods, and is better suited when there are a small to moderate number of pre-treatment periods.

When there are a large number of pre-treatment periods (i.e., long enough to apply the synthetic control methods), we recommend to apply both the synthetic control methods and proposed double DID, and evaluate robustness across those approaches. This is important because they rely on different identification assumptions. In fact, we show in Section 6.2, the double DID can recover credible estimates similar to more flexible variants of synthetic control methods even when there are a large number of pre-treatment periods. This robustness provides researchers with additional credibility for their causal estimates and underlying assumptions.

6 Empirical Application

6.1 Basic DID Design

Malesky, Nguyen and Tran (2014) utilize the basic DID design to study how the abolition of elected councils affects local public services in Vietnam. To estimate the causal effects of the institutional change, the original authors rely on data from 2008 and 2010, which are before and after the abolition of elected councils in 2009. Then, they supplement the 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, Nguyen and Tran (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 G.

6.1.1 Visualizing and Assessing Underlying Assumptions

The first step of the DID design is to visualize trends of treatment and control groups. Figure 4 shows trends of three different outcomes: "Education and Cultural Program," "Tap Water," and "Agricultural Center."[4] Although the original analysis uses the same DID design for all of them, they have distinct trends in the pre-treatment periods. The first outcome of "Education and Cultural Program" has parallel trends in pre-treatment periods. For the other two outcomes, trends do not look parallel in either of the cases. While the trends for the second outcome ("Tap Water") have similar directions, trends for the third outcome ("Agricultural Center") has opposite signs. This visualization of trends serves as 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

[4]"Education and Cultural Program" (binary): This variable takes one if there is a program that invests in culture and education in the commune. "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, Nguyen and Tran (2014) for further details.

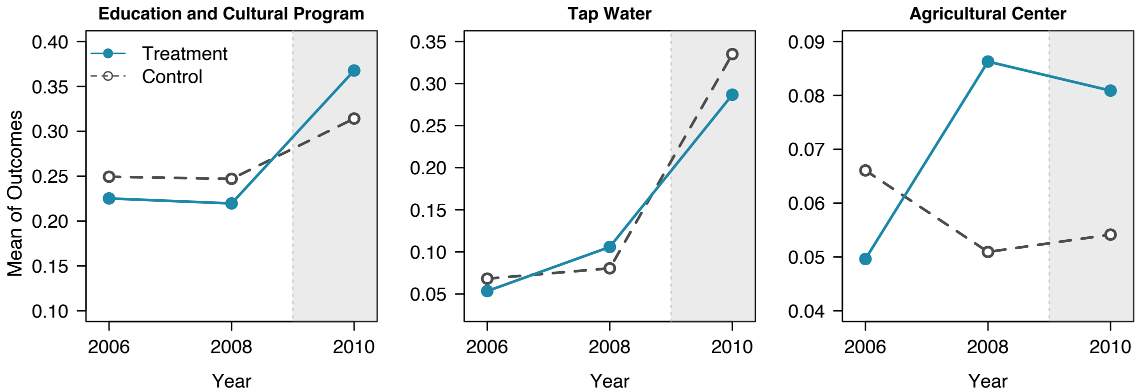

Figure 4: Visualizing Trends of Treatment and Control Groups. *Note:* We report trends for the treatment group (blue solid line with solid circles) and the control group (gray dashed line with hollow circles). Two pre-treatment periods are 2006 and 2008. One post-treatment period, 2010, is indicated by the gray shaded area.

plausible. Based on detailed domain knowledge, Malesky, Nguyen and Tran (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 (15)) 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.

While a traditional approach is to assess whether estimates are statistically distinguishable from zero with the conventional 5% or 10% level, we also report results based on an equivalence approach that we recommend in Section 3. Specifically, we compute the 95% standardized equivalence confidence interval, which quantifies the smallest equivalence range supported by the observed data (Hartman and Hidalgo, 2018). In the context of this application, the equivalance confidence interval is standardized based on the mean and standard deviation of the control group in 2006. For example, if the 95% standardized equivalence confidence interval is $[-\nu, \nu]$, this means that the equivalence test rejects the hypothesis that the DID estimate (standardized with respect to the baseline control outcome) 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 2. Standard errors are computed via block-bootstrap at the district level, where we take 2000 bootstrap iterations. For the first outcome, as the graphical presentation in Figure 4 suggests, a statistical test suggests that the extended parallel

	Estimate	Std. Error	p-value	95% Std. Equivalence CI
Education and Cultural Program	-0.07	0.096	0.940	$[-0.166, 0.166]$
Tap Water	0.166	0.084	0.048	$[-0.302, 0.302]$
Agricultural Center	0.198	0.095	0.037	$[-0.332, 0.332]$

Table 2: 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% standardized equivalence confidence intervals.

trends assumption is plausible. The test of the conditional extended parallel trends yields the p-value of 0.940 (the third column), and similarly, the 95% standardized equivalence confidence interval is $[-0.166, 0.166]$ (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 0.048, and the 95% standardized equivalence confidence interval, $[-0.302, 0.302]$, reveals that the parallel trends assumption is less plausible for this outcome, than for the first outcome. Finally, for the third outcome, "Agricultural Center," both traditional and equivalence approaches provide little evidence for parallel trends as graphically clear in Figure 4. The test of the parallel trends is rejected at the 5% level (p-value $= 0.037$) and the 95% standardized equivalence confidence interval is relatively large, $[-0.332, 0.332]$. 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 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.

6.1.2 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 "Education and Cultural Program," 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 estimator. The first plot of Figure 5 clearly shows this pattern. In the figure, we report point estimates as well as 90%

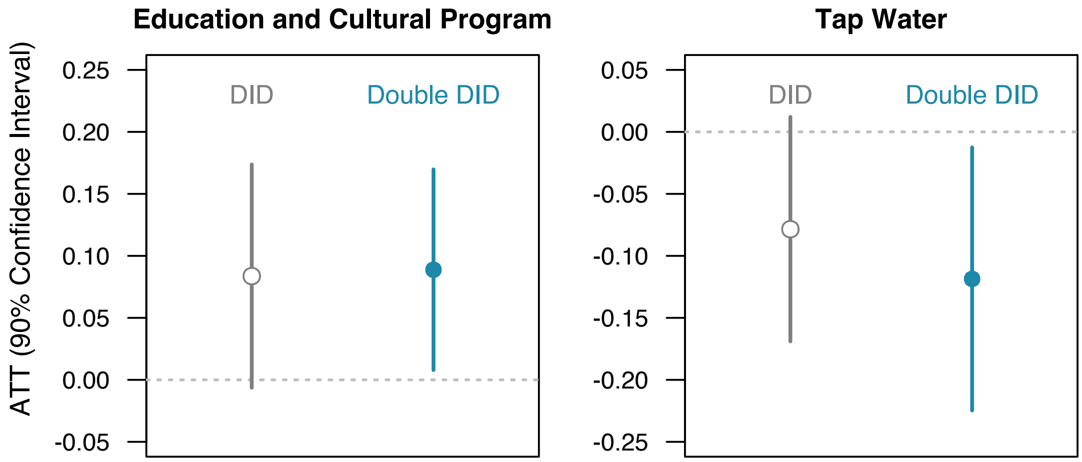

Figure 5: Estimating Causal Effects of Abolishing Elected Councils. *Note:* We compare estimates from the standard DID and the proposed double DID. For the first plot where the extended parallel trends assumption is plausible, the double DID produces a similar point estimate with smaller standard errors. 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.

confidence intervals following the original paper (see Figure 3 in Malesky, Nguyen and Tran (2014)). Using the standard DID estimator, the original estimate of the ATT on "Education and Cultural Program" was 0.084 (90% CI $= [-0.007, 0.174]$). Using the double DID estimator, an estimate is instead 0.088 (90% CI $= [0.008, 0.170]$). By using the double DID estimator, we shrink standard errors by about 10%. Although we only have two pre-treatment periods here, when there are more pre-treatment periods, the efficiency gain of the double DID can be 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 in contrast to the standard DID that still assumes constant unmeasured confounders. The second plot of Figure 5 shows the important difference between the two methods. While the standard DID estimates is -0.079 (90% CI $= [-0.169, 0.012]$), the double DID estimate is -0.119 (90% CI $= [-0.225, -0.013]$). Given that the extended parallel trends assumption is not plausible, this result suggests that the standard DID suffers from substantial bias (the bias of 0.04 corresponds to more than 50% of the original point estimate). 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 ("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 G).

6.2 Staggered Adoption Design

In this section, we apply the proposed double DID estimator to revisit Paglayan (2019), which uses the staggered adoption (SA) design to study the effect of granting collective bargaining rights to teacher's union on educational expenditures and teacher's salary. Paglayan (2019) applies the standard two-way fixed effect models to estimate the effect of the introduction of the mandatory bargaining law in the US states on the two outcome. The original author exploits the variation induced by the different introduction timing of the law: A few states introduced the law as early as in the mid 1960's, while some states, such as Arizona or Kentucky, never introduced the mandate. Among the states that granted the bargaining rights, the introduction timing varies from the mid 1960's to the mid 1980's (Nebraska was the last state that adopted the law).

6.2.1 Assessing Underlying Assumptions

We apply the proposed double DID for the SA design to the panel data consists of state-year observations. A state is treated at a particular year, if the state passes the law or has already passed the law of mandatory bargaining. Following the original study, we study two outcome: Per-pupil expenditure and annual teacher salary, both are on a log scale. There are 2,058 observations, containing 49 states (excluding Washington DC and Wisconsin, due to the short availability of the pre-treatment outcomes) and spanning from 1959 through 2000.

Figure 6 shows the variation of the treatment across states and over time. Cells in gray indicate state-year observations that are not treated and blue cells indicate the treated observations. We can observe that there are 14 unique treatment timings (the earliest is 1965 and

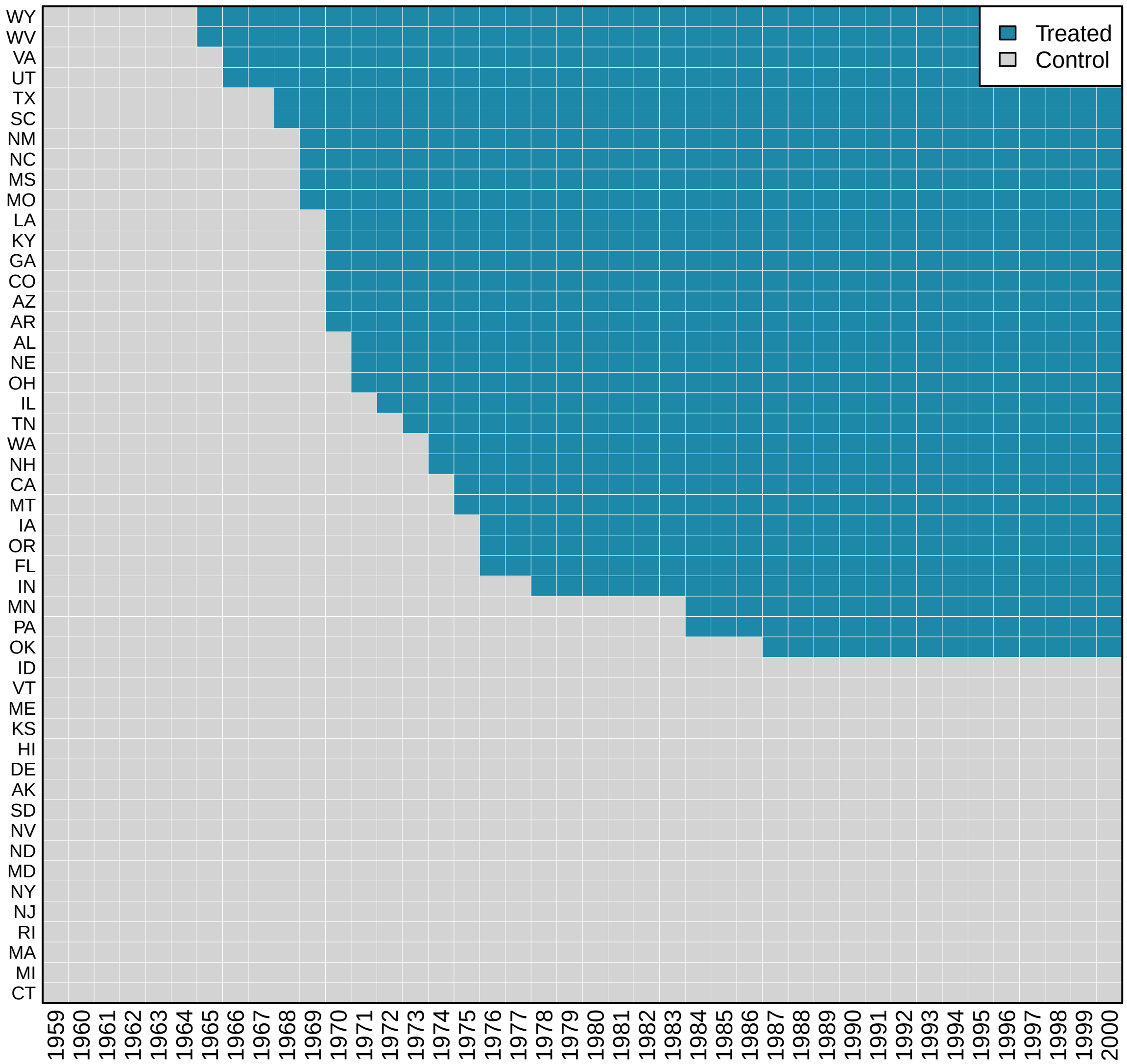

Figure 6: Treatment Variation Plot. *Note:* Cells in gray are state-year observations that are not treated (i.e., the mandatory bargaining law is not implemented), while cells in blue are observations that are under the treatment condition. Rows are sorted such that states that adopt the policy at earlier years are shown near the top, while states that never adopt the policy are shown near the bottom. The figure indicates that there are variations across states in adoption timings, and that some states never adopt the policy.

the latest is 1987) where the number of states at each treatment timing varies from one to six (the average number of states at a treatment timing is 2.3). We can also see that there is no reversal of a treatment status in that once a state adopts the policy, the state has never abolished it during the sample period.

We assess the underlying parallel trends assumption for the SA design by utilizing the pre-treatment outcome. As in the pre-treatment-trends test in the basic DID design, we apply the

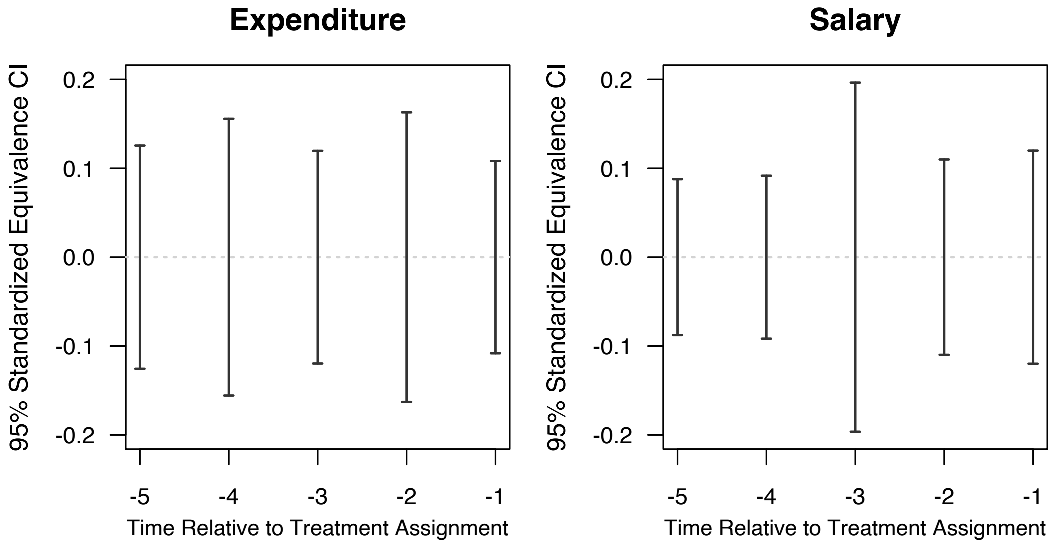

Figure 7: Assessing Underlying Assumptions Using the Pre-treatment Outcomes (Left: logged expenditure; Right: logged teacher salary). *Note:* We report the 95% standardized equivalence confidence intervals.

standard DID estimator for the SA design to pre-treatment periods. For example, to test the pre-treatment trends from $t-1$ to t for units who receive the treatment at time t, we estimate the SA-ATT using the outcome from $t-2$ and $t-1$ (See Section 4.2 for more details). To further facilitate interpretation, we standardize the outcome by the mean and standard deviation of the baseline control group, so that the effect can be interpreted relative to the control group.

Figure 7 shows 95% standardized equivalence confidence intervals for the two outcomes of interest (See Section 3.1 for details on the standardization procedure). It shows that for both outcomes, the equivalence confidence intervals are within 0.2 standard deviation from the means of the baseline control groups through $t-5$ to $t-1$. This suggests that the extended parallel trends assumption is plausible for both outcomes.

6.2.2 Estimating Causal Effects

We apply the double DID for the SA design as described in Section 4. The standard errors are computed by conducting the block bootstrap where the block is taken at the state level and we take 2000 bootstrap iterations. Analyses for the two outcomes are conducted separately. In addition to the proposed method, we apply two existing variants of synthetic control methods that can handle the staggered adoption design: the generalized synthetic control method, `gsynth`, (Xu, 2017) and the augmented synthetic control method, `augsynth`, (Ben-Michael, Feller and Rothstein, 2019). While the proposed double DID is better suited for settings where

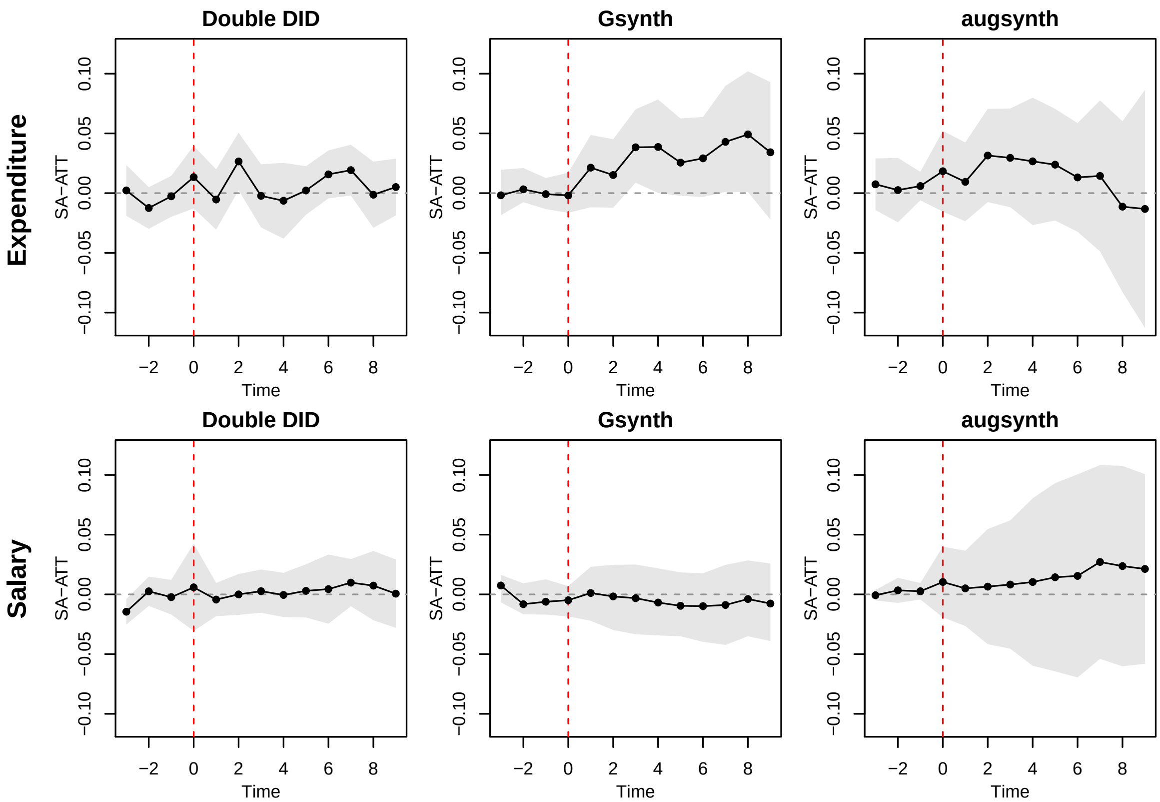

Figure 8: Plot of the Average Treatment Effect on the Treated on Two Outcomes. *Note:* We compare estimates from the double DID, the generalized synthetic control method, and the augmented synthetic control method. The causal estimates are similar across methods for both outcomes and treatment effects are not statistically significant at the conventional 5% level for most of the time periods.

there are a small to moderate number of pre-treatment periods, we evaluate, in the setting of long pre-treatment periods, whether it can achieve comparable performance to these variants of synthetic control methods that are primarily designed to deal with long pre-treatment periods (see more discussions in Section 5.3).

Figure 8 shows the estimates of the treatment on the per-pupil expenditure (the first row) and the teacher's salary (the second row), where both effects are on a log scale. We estimated the average treatment effect on the two outcomes ℓ periods after the treatment assignment where $\ell = \{0, 1, \ldots, 9\}$. Note that $\ell = 0$ corresponds to the contemporaneous effect. Each column corresponds to different estimators. The first column shows the proposed double DID estimator for the staggered adoption design, whereas the second (third) column shows estimates based on the generalized synthetic control method (the augmented synthetic control method). We can see that estimates are similar across methods for both outcomes and treatment effects are not statistically significant at the 5% level for most of the time periods. This result is consistent with the original finding of Paglayan (2019) that the granting collective bargaining rights did not increase the level of resources devoted to education.

As in this example, when there are a large number of pre-treatment periods, it is important to apply both synthetic control methods and the proposed double DID and evaluate robustness across those approaches. This is critical because they rely on different identification assumptions. We found such robustness in this application, which provides us with additional credibility.

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 basic DID design and the staggered adoption 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 propose a simple method, the double DID, to combine all three benefits within the GMM framework. Importantly, the double DID contains the popular two-way fixed effects regression and nonparametric DID estimators as special cases, and it use the GMM to further improve with respect to identification and estimation accuracy. Finally, we provide two key extensions. First, we accommodate any number of pre- and post-treatment periods, which allows for even more flexible forms of unmeasured time-varying confounding. Second, we generalize the double DID estimator to the staggered adoption design where timing of the treatment assignment can vary across units.

References

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

Abadie, Alberto. 2019. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* .

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 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.

Abraham, Sarah and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects." Available at `https://arxiv.org/abs/1804.05785`.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Athey, Susan and Guido W Imbens. 2018. "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." National Bureau of Economic Research.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi. 2017. "Matrix Completion Methods for Causal Panel Data Models." Available at `https://arxiv.org/abs/1710.10251`.

Bechtel, Michael M and Jens Hainmueller. 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, Nathaniel and Jonathan N Katz. 2011. "Modeling Dynamics in Time-Series–Cross-Section Political Economy Data." *Annual Review of Political Science* 14:331–352.

Ben-Michael, Eli, Avi Feller and Jesse Rothstein. 2018. "The Augmented Synthetic Control Method." Available at `https://arxiv.org/abs/1811.04170`.

Ben-Michael, Eli, Avi Feller and Jesse Rothstein. 2019. "Synthetic Controls and Weighted Event Studies with Staggered Adoption." *arXiv preprint arXiv:1912.03290* .

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

Blackwell, Matthew and Adam N Glynn. 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, Will and Joshua D Clinton. 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, Suzanna and Luke Keele. 2008. "Taking Time Seriously." *American Journal of Political Science* 52(1):184–200.

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

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

Earle, John S and Scott Gehlbach. 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, Francisco. 2018. "Elite Competition and State Capacity Development: Theory and Evidence from Post-Revolutionary Mexico." *American Political Science Review* 112(2):339–357.

Goodman-Bacon, Andrew. 2018. "Difference-in-Differences with Variation in Treatment Timing." National Bureau of Economic Research.

Hall, Andrew 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, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50(4):1029–1054.

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

Hazlett, Chad and Yiqing Xu. 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, Kosuke and In Song Kim. 2011. "Understanding and Improving Linear Fixed Effects Regression Models for Causal Inference." Available at `https://imai.fas.harvard.edu/research/files/FEmatchOld.pdf`.

Imai, Kosuke and In Song Kim. 2019. "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, Kosuke and In Song Kim. 2020. "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* .

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

Keele, Luke J, Dylan S Small, Jesse Y Hsu and Colin B Fogarty. 2019. "Patterns of Effects and Sensitivity Analysis for Differences-in-Differences." Available at `https://arxiv.org/abs/1901.01869`.

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

Ladd, Jonathan McDonald and Gabriel S Lenz. 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, Horacio and John Marshall. 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.

Lee, Myoung-jae. 2016. "Generalized Difference in Differences With Panel Data and Least Squares Estimator." *Sociological Methods & Research* 45(1):134–157.

Liu, Licheng, Ye Wang and Yiqing Xu. 2020. "A Practical Guide to Counterfactual Estimators for Causal Inference With Time-Series Cross-Sectional Data." *Available at SSRN 3555463* .

Malesky, Edmund J, Cuong Viet Nguyen and Anh Tran. 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, Ricardo and Iliana Reggio. 2019. "Alternative Diff-in-Diffs Estimators with Several Pretreatment Periods." *Econometric Reviews* 38(5):465–486.

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

Paglayan, Agustina S. 2019. "Public-Sector Unions and the Size of Government." *American Journal of Political Science* 63(1):21–36.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies." *Journal of Educational Psychology* 66(5):688.

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

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

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

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

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

Ye, Ting, Luke Keele, Raiden Hasegawa and Dylan S Small. 2020. "A Negative Correlation Strategy for Bracketing in Difference-in-Differences with Application to the Effect of Voter Identification Laws on Voter Turnout." *arXiv preprint arXiv:2006.02423* .

Online Appendix

Using Multiple Pre-treatment Periods to Improve Difference-in-Differences and Staggered Adoption Designs

Table of Contents

A Review of Papers in *APSR* and *AJPS*

We conduct a review of the literature to assess current practices of the difference-in-differences (DID) design. Specifically, we search articles published in *American Political Science Review* and *American Journal of Political Science* from 2015 to 2019. Some of the papers we reviewed were accepted in 2019 and were officially published in 2020. Using Google Scholar, we find articles that contains any of the following keywords: "two-way fixed effect", "two-way fixed effects", "difference in difference" or "difference in differences." We then manually select articles from the list that uses the basic DID design and the staggered adoption design (see the main text for details about the first two design). This procedure left us with a total of 25 articles, 11 from APSR and 14 from AJPS. Table A1 and A2 show the articles in the list published in APSR and AJPS, respectively.

To determine the number of pre-treatment periods, we manually assess the listed articles. Among the 25 articles, 20 articles use the basic DID design, and 5 articles use the staggered adoption design. When a paper uses the basic DID design, we can determine the length of the pre-treatment periods from the data description and the time of the treatment assignment. On the other hand, the pre-treatment periods for the staggered adoption and the general design are set to the total number of time-periods available in the data, as the length of pre-treatment periods varies across units.

Authors	**Year**	**Title**
O'brien, D. Z., & Rickne J.	2016	Gender Quotas And Women's Political Leadership
Garfias, F.	2018	Elite Competition and State Capacity Development: Theory and Evidence From Post-Revolutionary Mexico.
Martin, G. J., & Mccrain, J.	2019	Local News And National Politics
Blom-Hansen, J., Houlberg, K., Serritzlew, S., & Treisman, D.	2016	Jurisdiction Size and Local Government Policy Expenditure: Assessing The Effect of Municipal Amalgamation
Clinton, J. D., & Sances, M. W.	2018	The Politics of Policy: The Initial Mass Political Effects of Medicaid Expansion in The States
Malesky, E. J. , Nguyen, C. V., & Tran, A.	2014	The Impact of Recentralization on Public Services A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam.
Larsen, M. V., Hjorth, F., Dinesen, P. T., & Sønderskov, K. M.	2019	When Do Citizens Respond Politically to The Local Economy? Evidence From Registry Data on Local Housing Markets
Becher, M., & González, I. M.	2019	Electoral Reform and Trade-Offs in Representation
Selb, P., & Munzert, S.	2018	Examining A Most Likely Case for Strong Campaign Effects
Enos, R. D., Kaufman, A. R., & Sands, M. L.	2019	Can Violent Protest Change Local Policy Support?
Vasiliki Fouka	2019	How Do Immigrants Respond to Discrimination?

Table A1: DID papers on APSR.

Authors	**Year**	**Title**
Bechtel, M. M., Hangartner, D., & Schmid, L.	2016	Does compulsory voting increase support for leftist policy?
Bisgaard, M., & Slothuus, R.	2018	Partisan elites as culprits? How party cues shape partisan perceptual gaps.
Bischof, D., & Wagner, M.	2019	Do voters polarize when radical parties enter parliament?
Dewan, T., Meriläinen, J., & Tukiainen, J.	2020	Victorian voting: The origins of party orientation and class alignment.
Earle, J. S., & Gehlbach, S.	2015	The Productivity Consequences of Political Turnover: Firm-Level Evidence from Ukraine's Orange Revolution.
Enos, R. D.	2016	What the demolition of public housing teaches us about the impact of racial threat on political behavior.
Gingerich, D. W.	2019	Ballot Reform as Suffrage Restriction: Evidence from Brazil's Second Republic.
Hainmueller, J, & Hangartner, D.	2019	Does direct democracy hurt immigrant minorities? Evidence from naturalization decisions in Switzerland.
Holbein, J. B., & Hillygus, D. S.	2016	Making young voters: the impact of preregistration on youth turnout.
Jäger, K.	2020	When Do Campaign Effects Persist for Years? Evidence from a Natural Experiment.
Lindgren, K. O., Oskarsson, S., & Dawes, C. T.	2017	Can Political Inequalities Be Educated Away? Evidence from a Large-Scale Reform.
Lopes da Fonseca, M.	2017	Identifying the source of incumbency advantage through a constitutional reform.
Paglayan, AS.	2019	Public-Sector Unions and the Size of Government
Pardos-Prado, S., & Xena, C.	2019	Skill specificity and attitudes toward immigration.

Table A2: DID papers on AJPS.

B Nonparametric Equivalence to Regression Estimators

In this section, we provide results on the nonparametric connection between regression estimators and the three DID estimators we discussed in the paper. This section provides methodological foundations for our main methodological contributions, which we prove in Sections C and D.

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 1 (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.*

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

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

Proof. By solving the least squares problem, we obtain the following solutions:

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

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 2 (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 the demeaned treatment and outcome variables, $\overline{Y}_i = \sum_{t=1}^{2} Y_{it}/2$, $\overline{Y}_t = \sum_{i=1}^{n} Y_{it}/n$, $\overline{Y} = \sum_{i=1}^{n} \sum_{t=1}^{2} Y_{it}/2n$, $\overline{D}_i = \sum_{t=1}^{2} D_{it}/2$, $\overline{D}_t = \sum_{i=1}^{n} D_{it}/n$, and $\overline{D} = \sum_{i=1}^{n} \sum_{t=1}^{2} D_{it}/2n$.

Given these transformed variables, 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 report the following result.

Result 3 (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.*

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

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

$$\lambda = \frac{n_{11}n_{01}(n_{10}+n_{00})}{n_{11}n_{01}(n_{10}+n_{00}) + n_{10}n_{00}(n_{11}+n_{01})},$$

$$1-\lambda = \frac{n_{10}n_{00}(n_{11}+n_{01})}{n_{11}n_{01}(n_{10}+n_{00}) + n_{10}n_{00}(n_{11}+n_{01})}.$$

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 in 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 report that the extended DID estimator $\widehat{\tau}_{\texttt{e-DID}}$ (equation (9)) (equal weights: $\lambda = 1/2$) is equivalent to the estimated coefficient $\widehat{\beta}$ in the two-way fixed effects regression estimator in the panel data setting with $t = \{0, 1, 2\}$.

Result 4 (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.*

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

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,

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

as in Result 2, where $\widetilde{Y}_{it} = Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{Y}$ and $\widetilde{D}_{it} = D_{it} - \overline{D}_i - \overline{D}_t + \overline{D}$. 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

$$\widehat{\beta} = \frac{1}{2n_1} \left\{\sum_{i=1}^{n} G_i \{Y_{i2} - Y_{i1}\} + \sum_{i=1}^{n} G_i \{Y_{i2} - Y_{i0}\}\right\}$$

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

which completes the proof. □

B.3 Sequential DID

B.3.1 Repeated Cross-Sectional Data

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

Result 5 (Nonparametric Equivalence of the Sequential DID and Regression Estimator) *We focus on a linear regression estimator with a transformed outcome.*

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

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 1, we obtain

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

which completes the proof. □

Next, we clarify 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.

Result 6 (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.*

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

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 clarify 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.

Result 7 (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.*

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

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

Proof. As in Result 2, we can focus on the demeaned form.

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

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 2,

$$\widehat{\beta} = \frac{1}{n_1} \sum_{i=1}^{n} G_i (\Delta Y_{i2} - \Delta Y_{i1}) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - G_i)(\Delta Y_{i2} - \Delta Y_{i1})$$

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

which concludes the proof. □

Next, we clarify 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.

Result 8 (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*

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

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

Proof. By solving the least squares problem, we obtain that

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

Therefore, we get

$$
\begin{aligned}
\widehat{\beta} = &\left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_1} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_1} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_0} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_0} \right) \right\} \\
&- \left\{ \left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_1} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_1} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_0} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_0} \right) \right\},
\end{aligned}
$$

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 2.2 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.

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

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

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

Finally, using Result 1, we have

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

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.

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

Again, this least squares problem is the same as

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

Then, using Result 2, we have

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

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

C Generalized K-Difference-in-Differences

In this section, we propose the generalized K-DID, which extends the double DID in Section 3 to arbitrary number of *pre-* and *post-*treatment periods in the basic DID setting. We consider the staggered adoption design in Section 4.

C.1 The Setup and Causal Quantities of Interest

We first extend the setup to account for arbitrary number of pre- and post-treatment periods. Suppose we observe outcome Y_{it} for $i \in \{1, \ldots, n\}$ and $t \in \{0, 1, \ldots, T\}$. We define the binary treatment variable to be $D_{it} \in \{0, 1\}$. The treatment is assigned right before time period T^*, and thus, time periods $t \in \{T^*, \ldots, T\}$ are the post-treatment periods and time periods $t \in \{0, \ldots, T^* - 1\}$ are the pre-treatment periods. As in Section 2.1, 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 are interested in the causal effect at post-treatment time $T^* + s$ where $s \geq 0$. When $s = 0$, this corresponds to the contemporaneous treatment effect. By specifying different values of $s > 0$, researchers can study a variety of long-term causal effects of the treatment. Formally, our quantity of interest is the average treatment effect on the treated (ATT) at post-treatment time $T^* + s$.

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

For example, when $s = 3$, this could mean the causal effect of the policy after three years from its initial introduction. This definition is a generalization of the standard ATT: when $s = 0$, this quantity is equal to the ATT defined in equation (1).

C.2 Generalize Parallel Trends Assumptions

What assumptions do we need to identify the ATT at post-treatment time $T^* + s$? 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 A1 (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. For $g \in \{0, 1\}$,

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

when $k = 1$ and, in general,

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

where $M_s^\ell = \prod_{j=1}^{\ell-1}(s+j)/\prod_{j=1}^{\ell-1} j$ for $\ell \geq 2$. $\Delta^k\left(\mathbb{E}[Y_{i,T^*-1}(0) \mid G_i = g]\right)$ is also recursively defined as $\Delta^k\left(\mathbb{E}[Y_{i,T^*-1}(0) \mid G_i = g]\right) \equiv \Delta^{k-1}\left(\mathbb{E}[Y_{i,T^*-1}(0) \mid G_i = g]\right) - \Delta^{k-1}\left(\mathbb{E}[Y_{i,T^*-2}(0) \mid G_i = g]\right)$, and $\Delta^1\left(\mathbb{E}[Y_{i,T^*-m}(0) \mid G_i = g]\right) = \mathbb{E}[Y_{i,T^*-m}(0) \mid G_i = g] - \mathbb{E}[Y_{i,T^*-m-1}(0) \mid G_i = g]$ for $m = \{1, 2\}$. The standard parallel trends assumption and the parallel-trends-in-trends assumption are both special cases of this assumption. The k-th order parallel trends assumption reduces to the standard parallel trends assumption (Assumption 1) when $s = 1$ and $k = 1$, and to the parallel-trends-in-trends assumption (Assumption 3) when $s = 1$ and $k = 2$.

To further clarify the meaning of Assumption A1, we can consider a simpler but stronger condition. In particular, the k-th order parallel trends assumption (Assumption A1) is implied by the following p-th degree polynomial model of confounding.

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

with unknown parameters α and $\mathbf{\Gamma}$. Here, the left hand side of the equality captures the difference between the two groups (treatment and control) in terms of the mean of potential outcomes under the control condition. This representation shows that the standard parallel trends assumption (Assumption 1) is implied by the time-invariant confounding; 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.

C.3 Estimate ATT with Multiple Pre- and Post-Treatment Periods

We consider the identification and estimation of the ATT at post-treatment time $T^* + s$. Under the k-th order parallel trends assumption (Assumption A1), 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 at time $T^* + s$ under the k-th order parallel trends assumption. When $s = 0$ and $k = 1$, this estimator corresponds to the standard DID estimator (equation (3)). When $s = 0$ 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; Lee, 2016; Mora and Reggio, 2019) consider each estimator separately, we propose combining multiple DID estimators within the GMM framework.

In general, the generalized double DID combines K moment conditions where K is the number of pre-treatment periods researchers use. When there are more than two pre-treatment periods, we can naturally combine more than two DID estimators, improving upon the double DID in Section 3. Formally, the generalized double DID is defined as,

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

where $\boldsymbol{g}(\tau) = (\tau - \widehat{\tau}_1(s), \ldots, \tau - \widehat{\tau}_K(s))^\top$. Based on the theory of the efficient GMM (Hansen, 1982), the optimal weight matrix is $\widehat{\mathbf{W}} = \text{Var}(\widehat{\tau}_{(1:K)}(s))^{-1}$ where $\text{Var}(\cdot)$ is the variance-covariance matrix and $\widehat{\tau}_{(1:K)}(s) = (\widehat{\tau}_1(s), \ldots, \widehat{\tau}_K(s))^\top$. When $T^* = 2$, this converges to the standard DID estimator (equation (3)). When $T^* = 3$, this corresponds to the basic form of the double DID estimator (equation (13)). 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 generalized double DID to pre-treatment periods $t \in \{1, \ldots, T^* - 1\}$ as if the last pre-treatment period $T^* - 1$ is the target time period. Moments are $\boldsymbol{g}(\tau) = (\tau - \widehat{\tau}_1(0), \ldots, \tau - \widehat{\tau}_K(0))^\top$ where $\widehat{\tau}_k(0) = \Delta_s^k \left(\frac{\sum_{i:\, G_i=1} Y_{i,T^*-1}}{n_{1,T^*-1}} \right) - \Delta_s^k \left(\frac{\sum_{i:\, G_i=0} Y_{i,T^*-1}}{n_{0,T^*-1}} \right)$. Similarly, to assess the extended parallel trends-in-trends assumption, we can apply the generalized double DID to pre-treatment periods with moments $\boldsymbol{g}(\tau) = (\tau - \widehat{\tau}_2(0), \ldots, \tau - \widehat{\tau}_K(0))^\top$.

D Generalized K-DID for Staggered Adoption Design

Combining the setup introduced in Section C.1 and the one in Section 4.1, we propose the generalized K-DID for the SA design, which allows researchers to estimate long-term causal effects in the SA design. We focus on the SA-ATT at post-treatment time $t+s$ where t is the timing of the treatment assignment and $s \geq 0$ represents how far in the future we want estimate the ATT for. We first redefine the group indicator G to estimate the long-term SA-ATT at post-treatment time $t+s$. In particular, we define

$$G_{its} = \begin{cases} 1 & \text{if } A_i = t \\ 0 & \text{if } A_i > t+s \\ -1 & \text{otherwise} \end{cases}$$

where $G_{its} = 1$ represents units who receive the treatment at time t, and $G_{its} = 0$ indicates units who do not receive the treatment by time $t+s$. $G_{its} = -1$ includes other units who receive the treatment before time t or receive the treatment between $t+1$ and $t+s$. When $s = 0$, this definition corresponds to the group indicator in equation (18).

Formally, our first quantity of interest is the *staggered-adoption average treatment effect on the treated* (SA-ATT) at post-treatment time $t+s$.

$$\tau^{\mathsf{SA}}(s,t) \equiv \mathbb{E}[Y_{i,t+s}(1) - Y_{i,t+s}(0) \mid G_{its} = 1].$$

By averaging over time, we can also define the *time-average staggered-adoption average treatment effect on the treated* (time-average SA-ATT) at s periods after treatment onset.

$$\overline{\tau}^{\mathsf{SA}}(s) \equiv \sum_{t \in \mathcal{T}} \pi_t \tau^{\mathsf{SA}}(s,t),$$

where $\mathcal{T}$ represents a set of the time periods for which researchers want to estimate the ATT. The SA-ATT in period t, $\tau^{\mathsf{SA}}(t)$, is weighted by the proportion of units who receive the treatment at time t: $\pi_t = \sum_{i=1}^n \mathbf{1}\{A_i = t\} / \sum_{i=1}^n \mathbf{1}\{A_i \in \mathcal{T}\}$.

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 A2 (k-th Order Parallel Trends for Staggered Adoption Design) *For some integer k such that $1 \leq k \leq T$, and for $k \leq t \leq T-s$,*

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

where Δ_s^k is the k-th order difference operator defined in Assumption A1.

Under Assumption A2, the SA-ATT at post-treatment time $t+s$ is identified as follows.

$$\tau^{\mathsf{SA}}(s,t) = \Delta_s^k \left(\mathbb{E}[Y_{i,t+s} \mid G_{its} = 1]\right) - \Delta_s^k \left(\mathbb{E}[Y_{i,t+s} \mid G_{its} = 0]\right).$$

Since conditional expectations can be consistently estimated via the sample analogue,

$$\widehat{\tau}_k^{\mathsf{SA}}(s,t) = \Delta_s^k \left(\frac{\sum_{i:\, G_{its}=1} Y_{i,t+s}}{n_{1,t+s}}\right) - \Delta_s^k \left(\frac{\sum_{i:\, G_{its}=0} Y_{i,t+s}}{n_{0,t+s}}\right)$$

is a consistent estimator for the SA-ATT at post-treatment time $t+s$ under Assumption A2.

In general, we combine K DID estimators to obtain the generalized K-DID for the SA-ATT at post-treatment time $t+s$ as follows.

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

where $\boldsymbol{g}(\tau^{\mathsf{SA}}) = (\tau^{\mathsf{SA}} - \widehat{\tau}_1^{\mathsf{SA}}(s), \ldots, \tau^{\mathsf{SA}} - \widehat{\tau}_K^{\mathsf{SA}}(s))^\top$. The optimal weight matrix is $\widehat{\mathbf{W}} = \text{Var}(\widehat{\tau}_{(1:K)}^{\mathsf{SA}}(s))^{-1}$ where $\widehat{\tau}_{(1:K)}^{\mathsf{SA}}(s) = (\widehat{\tau}_1^{\mathsf{SA}}(s), \ldots, \widehat{\tau}_K^{\mathsf{SA}}(s))^\top$.

To estimate the time-average SA-ATT, we first define the time-average k-th order time-average DID estimator as,

$$\widehat{\overline{\tau}}_k^{\mathsf{SA}}(s) = \sum_{t \in \mathcal{T}} \pi_t \widehat{\tau}_k^{\mathsf{SA}}(s,t).$$

Finally, the generalized K-DID combines K moment conditions as follows.

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

where $\boldsymbol{g}(\overline{\tau}^{\mathsf{SA}}) = (\overline{\tau}^{\mathsf{SA}} - \widehat{\overline{\tau}}_1^{\mathsf{SA}}(s), \ldots, \overline{\tau}^{\mathsf{SA}} - \widehat{\overline{\tau}}_K^{\mathsf{SA}}(s))^\top$. The optimal weight matrix is $\widehat{\overline{\mathbf{W}}} = \text{Var}(\widehat{\overline{\tau}}_{(1:K)}^{\mathsf{SA}}(s))^{-1}$ where $\widehat{\overline{\tau}}_{(1:K)}^{\mathsf{SA}}(s) = (\widehat{\overline{\tau}}_1^{\mathsf{SA}}(s), \ldots, \widehat{\overline{\tau}}_K^{\mathsf{SA}}(s))^\top$.

E Equivalence Approach

Here, we provide technical details on the equivalence approach we introduced in Section 3.1. In the standard hypothesis testing, researchers usually evaluate the two-sided null hypothesis $H_0 : \delta = 0$ where $\delta = \{\mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i0}(0) \mid G_i = 1]\} - \{\mathbb{E}[Y_{i1}(0) \mid G_i = 0] - \mathbb{E}[Y_{i0}(0) \mid G_i = 0]\}$ when we are conducting the pre-treament-trends test. However, this approach has a risk of conflating evidence for parallel trends and statistical inefficiency. For example, when sample size is small, even if pre-treatment trends of the treatment and control groups differ (i.e., the null hypothesis is false), a test of the difference might not be statistically significant due to large standard error. And, analysts might "pass" the pre-treatment-trends test by not finding enough evidence for the difference.

The equivalence approach can mitigate this concern by flipping the null hypothesis, so that the rejection of the null can be the evidence for parallel trends. In particular, we consider two one-sided tests:

$$H_0 : \theta \geq \gamma_U, \quad \text{or} \quad \theta \leq \gamma_L$$

where (γ_U, γ_L) is a user-specified equivalence range. By rejecting this null hypothesis, researchers can provide statistical evidence for the alternative hypothesis:

$$H_0 : \gamma_L < \theta < \gamma_U,$$

which means that θ (i.e., the difference in pre-treament-trends across treatment and control groups) are within an interval $[\gamma_L, \gamma_U]$.

One difficulty of the equivalence approach is that researchers have to choose this equivalence range (γ_U, γ_L), which might not be straightforward in practice. To overcome this challenge, we follow Hartman and Hidalgo (2018) to estimate the 95% equivalence confidence interval, which is the smallest equivalence range supported by the observed data. Suppose we obtain $[-c, c]$ as the symmetric 95% equivalence confidence interval where $c > 0$ is some positive constant. Then, this means that if researchers think the absolute value of θ smaller than c is substantively negligible, the 5% equivalence test would reject the null hypothesis and provide the evidence for the parallel pre-treatment-trends. In contrast, if researchers think the absolute value of θ being c is substantively too large as bias in practice, the 5% equivalence test would fail to reject the null hypothesis and cannot provide the evidence for the parallel pre-treatment-trends. In sum, by estimating the equivalence confidence interval, readers of the analysis can decide how much evidence for the parallel pre-treatment-trends exists in the observed data. Researchers can estimate the 95% equivalence confidence interval by the following general two steps. First, estimate 90% confidence interval, which we denote by $[b_L, b_U]$. Second, we can obtain the symmetric 95% equivalence confidence interval as $[-b, b]$ where we define $b = \max\{|b_L|, |b_U|\}$. See Wellek (2010); Hartman and Hidalgo (2018) for more details.

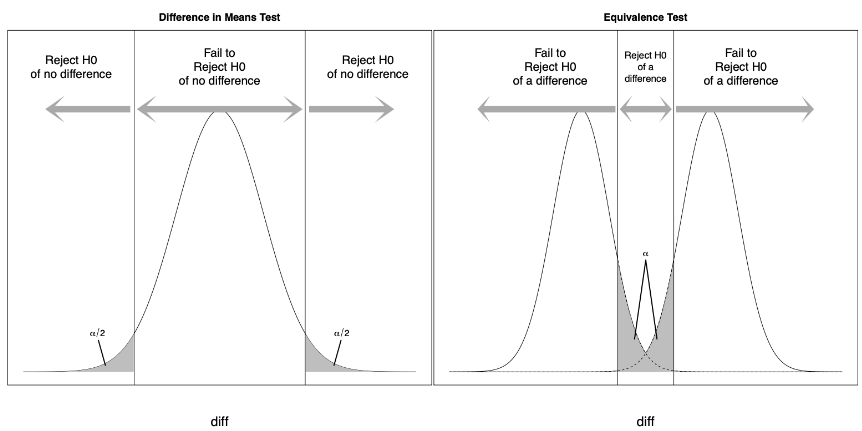

Figure A1: Figure 1 from Hartman and Hidalgo (2018) on the difference between the standard hypothesis testing and the equivalence testing.

F 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 standard errors over 2000 Monte Carlo simulations. We describe the details of the simulation setup next.

F.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

$$\mathbb{E}[Y_{it}(0) \mid G_i = g] = \alpha_t + 0.05 \times g \tag{1}$$

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

$$\mathbb{E}[Y_{it}(0) \mid G_i = g] = \alpha_t + 0.1 \times g \times (t + 1) \tag{2}$$

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 A2, 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 A3, 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)). Both figures show 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. Note that this standard error is a true standard error over the sampling distribution.

F.2 Results

Figure A2 shows the results when the autocorrelation parameter $\rho = 0.6$. 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.

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

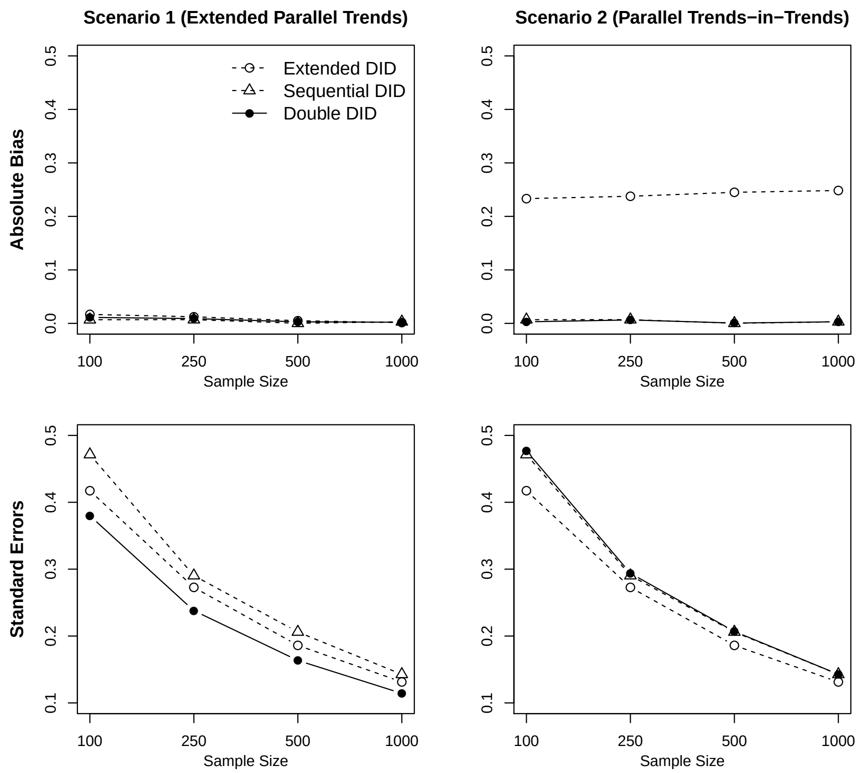

Figure A2: Comparing DID estimators in terms of the absolute bias and the standard errors. The first row shows that the double DID estimator (black circle with solid line) is unbiased under both scenarios. The second row demonstrates that the double DID has the smallest standard errors among unbiased DID estimators.

unbiased under the weaker assumption of the parallel trends-in-trends, but not efficient under the extended parallel trends.

In Figure A3, we 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)). We find 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.

The first row of Figure A3 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.

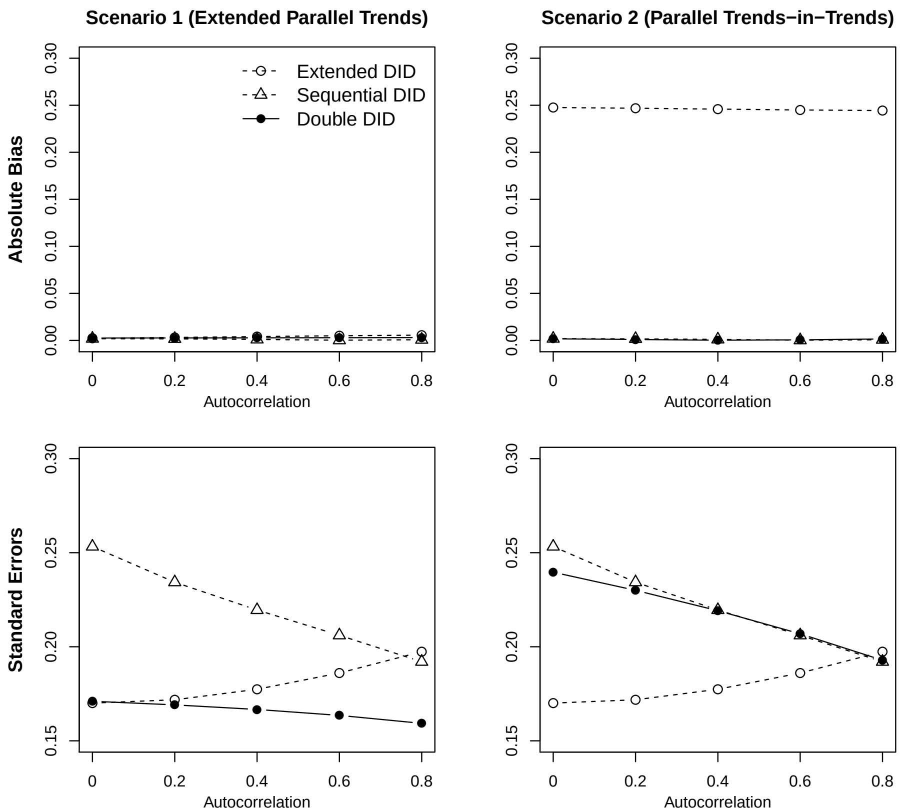

Figure A3: Comparing DID estimators in terms of the absolute bias and the standard errors according to the autocorrelation of errors. *Note*: The first row shows that the double DID estimator (black circle with solid line) is unbiased under both scenarios. The second row demonstrates that the double 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.

G Empirical Application

In Section 6, we have focused on three outcomes to illustrate the advantage of the double DID estimator. In this section, we provide results for all thirty outcomes analyzed in the original paper.

To assess the underlying parallel trends assumptions, we combine visualization and formal tests, as recommended in the main text. The assessment suggests that we can make the extended parallel trends assumption for fifteen outcomes. Specifically, for those fifteen outcomes, p-values for the null of pre-treatment parallel trends are above 0.10 (i.e., fail to reject the null at the conventional level), and the 95% standardize equivalence confidence interval is contained in the interval $[-0.2, 0.2]$. This means that the deviation from the parallel trends in the pre-treatment periods are less than 0.2 standard deviation of the control mean in 2006.

Figure A4 shows estimated treatment effects 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 can be even larger.

We rely on the parallel trends-in-trends assumption for eight outcomes out of the fifteen remaining outcomes. These outcomes have the 95% standardized equivalence confidence interval wider than $[-0.20, 0.20]$, but show that treatment and control groups' pre-treatment trends have the same sign. The same sign of the pre-treatment trends suggests that parallel trends-in-trends assumption, which can account for the linear time-varying unmeasured confounder, can be plausible for these outcomes, even though the stronger parallel trends assumption is possibly violated.

Figure A5 shows results under the parallel trends-in-trends assumption. As in Section 6, the double DID estimates are often 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.

For the remaining seven 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 $[-0.20, 0.20]$ as a cutoff for an illustration, it is recommended to base this decision on substantive domain knowledge whenever possible in practice.

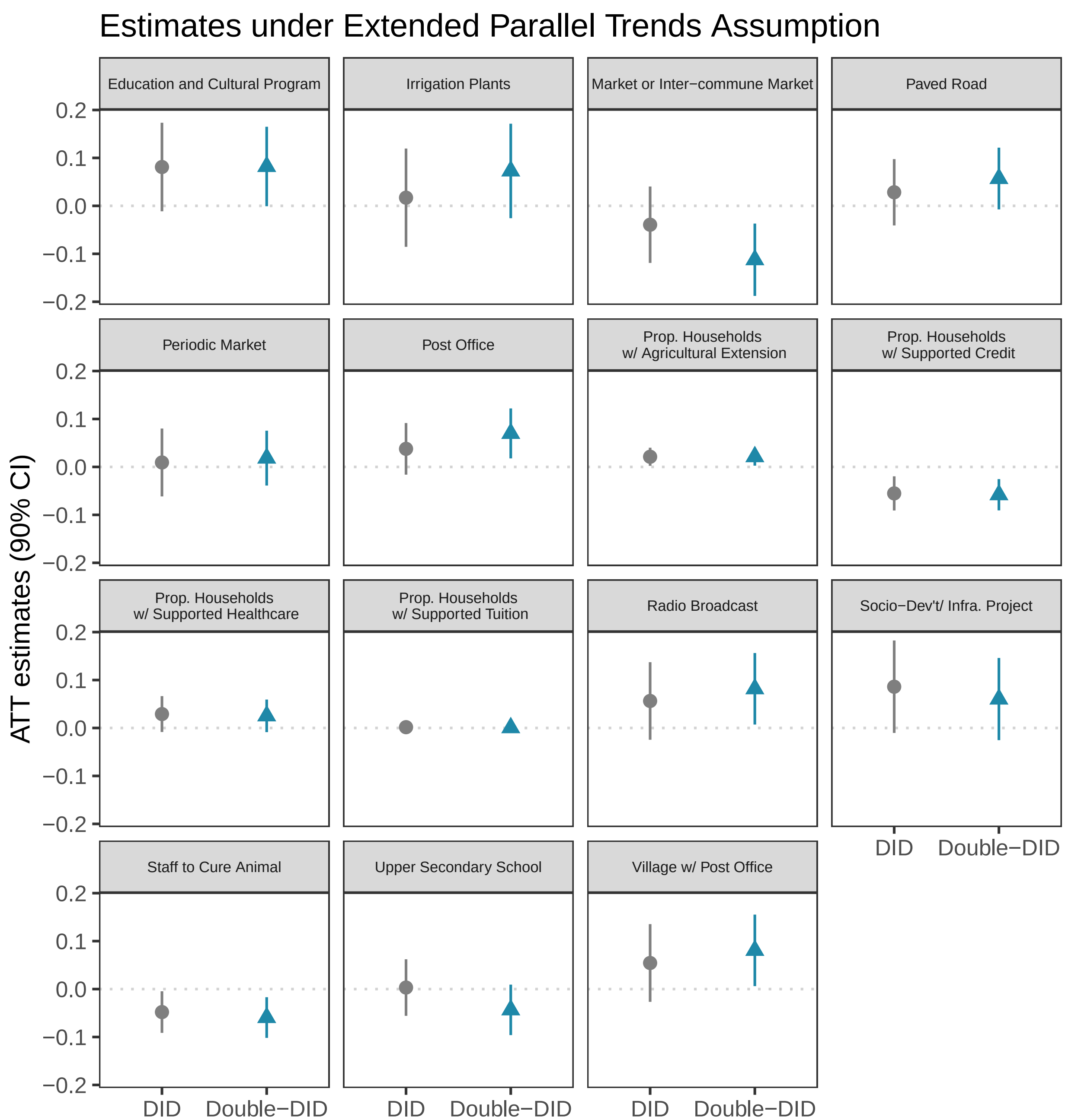

Figure A4: Comparing Standard DID and Double DID under Extended Parallel Trends Assumption. The double DID estimates are similar to those from the standard DID, and yet, standard errors are smaller because the double DID effectively uses pre-treatment periods within the GMM.

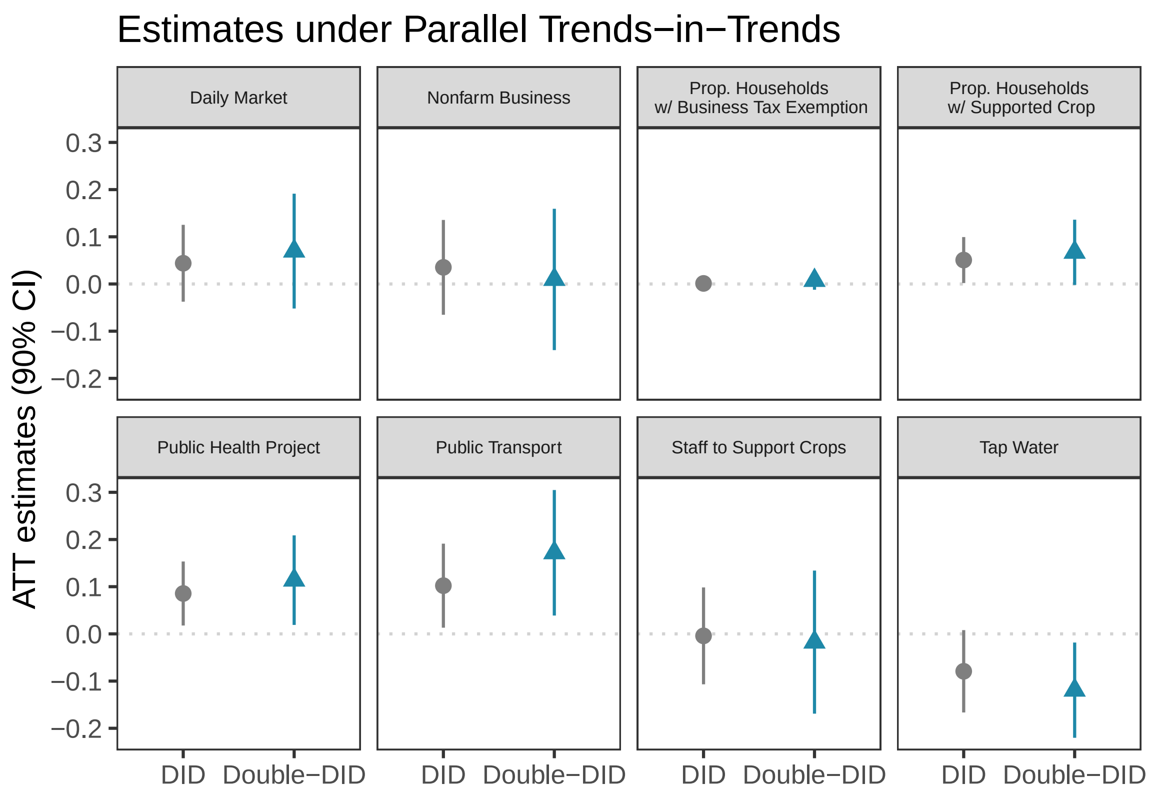

Figure A5: Comparing Standard DID and Double DID under Parallel Trends-in-Trends Assumption. The double DID estimates are often different from those of the standard DID because the extended parallel trends assumption is implausible for these outcomes.

References

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

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

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

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

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

Lee, Myoung-jae. 2016. "Generalized Difference in Differences With Panel Data and Least Squares Estimator." *Sociological Methods & Research* 45(1):134–157.

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

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