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

Naoki Egami† Soichiro Yamauchi‡

This version: September 15, 2021
First draft: December 6, 2019

Abstract

While a difference-in-differences (DID) design was originally developed with one pre- and one post-treatment period, data from additional pre-treatment periods are 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. We show that the double DID requires a weaker assumption about outcome trends and is more efficient than existing DID estimators. 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.

Word Count: 8960

*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 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. We also thank the editor and our two anonymous reviewers for providing us with valuable comments.

†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 inference from observational data. One of the most popular is a difference-in-differences (DID) design (Bertrand, Duflo, and Mullainathan 2004; Angrist and Pischke 2008). When the outcome trend of the control group would have been the same as the trend of the outcome in the treatment group in the absence of the treatment (known as the parallel trends assumption), the DID design enables scholars to estimate causal effects even in the presence of time-invariant unmeasured confounding (Abadie 2005). 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 alternative model specifications by including time-trends or leads of the treatment variable to assess possible violations of the parallel trends assumption. 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.[2] 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 (Imbens and Rubin 2015): (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 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. While our literature review finds that more than 90% of papers based on the DID design enjoy at least one of the three benefits, we also find that only 20% of the papers enjoy all three benefits.

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.

2. For each approach, we provide examples in Appendix A.

Our main contribution is to propose a new, simple estimator that achieves all 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-trends assumptions, and the sequential DID estimator (e.g., Lee 2016; Mora and Reggio 2019), which only requires that the change in the trends is the same across treatment and control groups (what we call the *parallel trends-in-trends assumption*). 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.

The proposed double DID approach makes several key methodological contributions. First, we show that the proposed method achieves better theoretical properties than widely-used DID estimators that constitute the double DID. Compared to the standard DID estimator and the two-way fixed effects regression, the double DID has smaller standard errors (i.e., more efficient) and is unbiased under a weaker assumption. While the former estimators require the parallel trends assumption, the double DID only requires the parallel trends-in-trends assumption. The double DID also improves upon the sequential DID estimator, which is inefficient when the parallel trends assumption holds. By using the GMM theory, we show that the double DID is more efficient than the sequential DID when the parallel trends assumption holds. Therefore, our proposed GMM approach enables methodological improvement both in terms of identification and estimation accuracy.

Second, and most importantly in practice, the double DID blends all the three benefits of multiple pre-treatment periods within a single framework. Therefore, instead of using different estimators for enjoying each benefit as required in existing methods, researchers can use the double DID approach to exploit all the benefits. Given that only 20% of papers based on the DID design currently enjoy all the three benefits, our proposed unified approach to optimally exploit all the three benefits of multiple pre-treatment periods is essential in practice.

We also propose three extensions of our double DID estimator. First, we develop the double DID regression, which can incorporate pre-treatment observed covariates to make the DID design more robust and efficient (Section 3.3.1). Second, we allow for any number of *pre*- and *post*-treatment periods (Section 3.3.2). While the parallel trends-in-trends assumption can allow for time-varying unmeasured confounders that linearly change over time, we show how to further relax the assumption by accounting for even more flexible forms of time-varying

unmeasured confounding when we have more *pre*-treatment periods. Because our proposed methods allow for any number of *post*-treatment periods, researchers can also estimate not only short-term causal effects but also longer-term causal effects. 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 social sciences (e.g., Ben-Michael, Feller, and Rothstein 2019; Athey and Imbens 2021; Marcus and Sant'Anna 2021).

We offer a companion R package DIDdesign that implements the proposed methods. We illustrate our proposed methods through two empirical applications. In Section 3.4, we revisit Malesky, Nguyen, and Tran (2014), which study 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. In Appendix H.2, we reanalyze Paglayan (2019), which examines the effect of granting collective bargaining rights to teacher's unions on educational expenditures and teacher's salaries. This is an example of the staggered adoption design.

Related Literature. This paper builds on the large literature of time-series cross-sectional data. 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., Strezhnev 2018; Imai and Kim 2019; Callaway and Sant'Anna 2020; Athey and Imbens 2021; Goodman-Bacon 2021; Imai and Kim 2021). 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 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.

Another class of popular methods is the synthetic control method (Abadie, Diamond, and Hainmueller 2010) and their recent extensions (e.g., Xu 2017; Ben-Michael, Feller, and Rothstein 2019; Pang, Liu, and Xu 2021) 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 is more than one pre-treatment period, and is better suited when there are a small to moderate number of pre-treatment periods.[3] However, we also show in Appendix H.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 Appendix B.

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. The basic DID design consists of treatment and control groups measured at two time periods, before and after the treatment assignment. While the basic DID design only requires data from one post- and one pre-treatment period, additional pre-treatment periods are often available. 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.

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 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 are 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. With this data, Malesky, Nguyen, and Tran (2014) aim to estimate the causal effect of abolishing elected councils on various measures of local public services.

2.1 Setup

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

3. In our literature review, we found that most DID applications have less than 10 pre-treatment periods, and the median number of pre-treatment periods is 3.5. See Appendix A for more details.

this setup here because it is sufficient for examining benefits of multiple pre-treatment periods, but we also generalize our methods to an arbitrary number of *pre-* and *post-* treatment periods (Section 3.3.2), and to the staggered adoption design (Section 4). In our 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, in which different communes are sampled at three time periods.

To define causal effects, we rely on the potential outcomes framework (Imbens and Rubin 2015). 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 $Y_{it}(1) - Y_{it}(0)$. As the treatment is assigned in the second time period, we are interested in estimating a causal effect at time $t = 2$, and a causal effect of interest is formally defined as $Y_{i2}(1) - Y_{i2}(0)$.

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 $G_i = 1$.

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 basic DID design, 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)**.**

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

The left-hand side of equation (2) 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.

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

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

When we analyze panel data, we can compute $\widehat{\tau}_{\texttt{DID}}$ nonparametrically via a linear regression with unit and time fixed effects. This numerical equivalence in the two-time-period case is often used to justify the two-way fixed effects regression as the DID design (Angrist and Pischke 2008). We discuss additional results on nonparametric equivalence between a regression estimator and the DID estimator in Appendix C.1.

2.2 Benefit 1: Assessing Parallel Trends Assumption

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

The first and the most common use of multiple pre-treatment periods is to assess the identification assumption of parallel trends. As the validity of the DID design rests on this 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 as a placebo test (Angrist and Pischke 2008).

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

$$\left(\frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right). \tag{4}$$

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 C.4, we show that a robustness check with leads effects (Angrist and Pischke 2008), which incorporates leads of the treatment variable into the two-way fixed effects regression and checks whether their coefficients are zero, is equivalent to this DID on the pre-treatment periods.

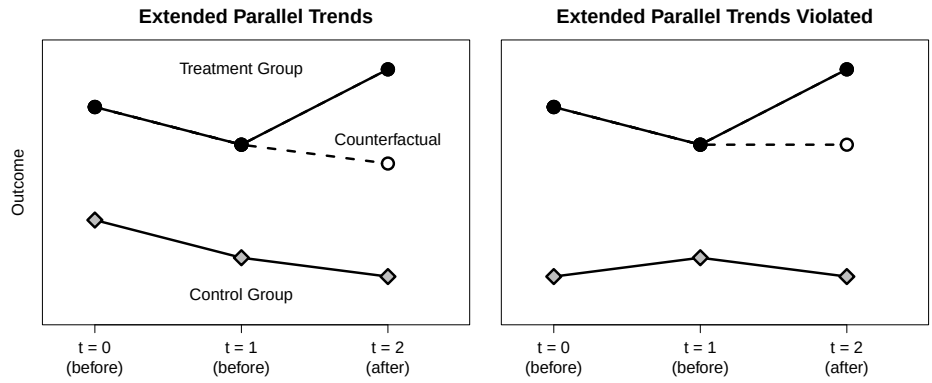

Figure 1: Parallel Pre-treatment Trends (left) and Non-Parallel Pre-treatment Trends (right).

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

Importantly, 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)**.**

$$\begin{cases} \mathbb{E}[Y_{i2}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i2}(0) \mid G_i = 0] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0] \\ \mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i0}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(0) \mid G_i = 0] - \mathbb{E}[Y_{i0}(0) \mid G_i = 0] \end{cases} \tag{5}$$

The first line of the extended parallel trends assumption is the same as the standard parallel trends assumption, and the second line is the parallel trends for pre-treatment periods. Because outcome trends are observable in pre-treatment periods, the test of pre-treatment trends (equation (4)) directly tests this assumption.

It is important to emphasize that, even if we find the DID estimate on pre-treatment periods is close to zero, we cannot confirm the extended parallel trends assumption (Assumption 2) or the parallel trends assumption (Assumption 1). This is because it is still possible that trends between $t = 1$ (pre-treatment) and $t = 2$ (post-treatment) are not parallel. Therefore, it is always important to substantively justify the parallel trends assumption in addition to using this statistical test based on pre-treatment trends.

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.

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

Under the extended parallel trends assumption, there are two natural DID estimators for the ATT.

$$\begin{aligned} \widehat{\tau}_{\texttt{DID}} &= \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i1}}{n_{11}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i1}}{n_{01}} \right), \\ \widehat{\tau}_{\texttt{DID(2,0)}} &= \left(\frac{\sum_{i:\, G_i=1} Y_{i2}}{n_{12}} - \frac{\sum_{i:\, G_i=1} Y_{i0}}{n_{10}} \right) - \left(\frac{\sum_{i:\, G_i=0} Y_{i2}}{n_{02}} - \frac{\sum_{i:\, G_i=0} Y_{i0}}{n_{00}} \right). \end{aligned} \tag{7}$$

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.

$$\widehat{\tau}_{\texttt{e-DID}} = \frac{1}{2}\widehat{\tau}_{\texttt{DID}} + \frac{1}{2}\widehat{\tau}_{\texttt{DID(2,0)}}. \tag{8}$$

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 equivalent to the two-way fixed effects estimator fitted to the three periods $t \in \{0, 1, 2\}$.

$$Y_{it} \sim \alpha_i + \delta_t + \beta D_{it}, \tag{9}$$

where α_i is a unit fixed effect, δ_t is a time fixed effect, and a coefficient of the treatment variable β is numerically equal to $\widehat{\tau}_{\texttt{e-DID}}$. We also present more general results about non-parametric relationships between the extended DID and the two-way fixed effects estimator in Appendix C.2.

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-differences 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 standard 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 $t = 1$ (equation (4)).

This sequential DID estimator requires the *parallel trends-in-trends* assumption — in the absence of the treatment, the change in the outcome trends of the treatment group is equal to the change in the outcome trends of the control group (e.g., Mora and Reggio 2019). While the parallel trends assumption requires that the outcome trends themselves are the same across the treatment and control groups, the *parallel trends-in-trends* assumption only requires the change in trends over time to be the same. Formally, the parallel trends-in-trends assumption can be written as follows.

Assumption 3 (Parallel Trends-in-Trends)**.**

$$\underbrace{\big\{\mathbb{E}[Y_{i2}(0) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1]\big\}}_{\text{Trend of the treatment group from } t=1 \text{ to } t=2} - \underbrace{\big\{\mathbb{E}[Y_{i1}(0) \mid G_i = 1] - \mathbb{E}[Y_{i0}(0) \mid G_i = 1]\big\}}_{\text{Trend of the treatment group from } t=0 \text{ to } t=1}$$

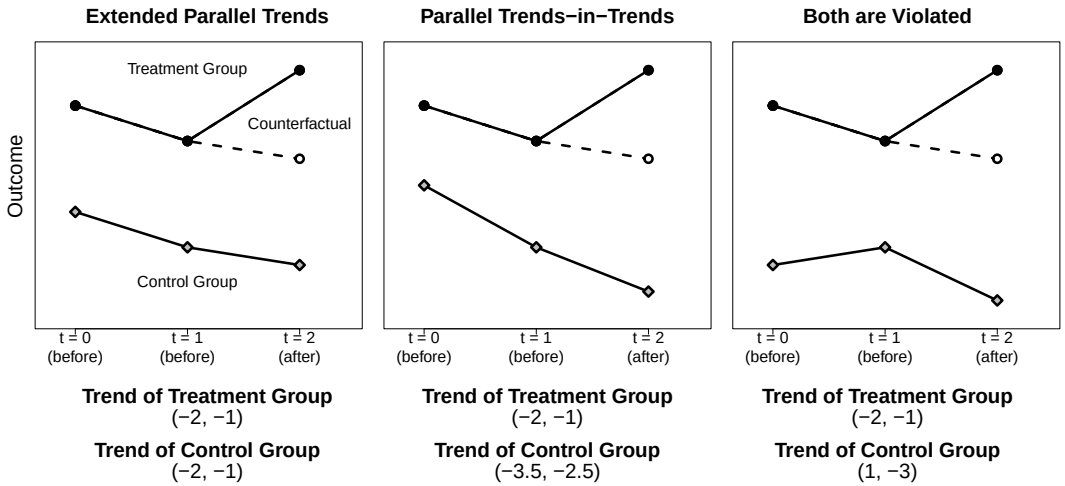

Figure 2: Comparing Extended Parallel Trends and Parallel Trends-in-Trends Assumptions. *Note:* Below each panel, we report the trends of the control potential outcomes for the treatment and control groups. The first and second elements show the outcome trends (from $t = 0$ to $t = 1$) and (from $t = 1$ to $t = 2$), respectively. The extended parallel trends assumption (left panel) means that the outcome trends are the same across the treatment and control groups for both (from $t = 0$ to $t = 1$) and (from $t = 1$ to $t = 2$). The parallel trends-in-trends assumption (middle panel) only requires its change over time is the same across the treatment and control groups; $(-1) - (-2) = (-2.5) - (-3.5) = 1$. Both assumptions are violated in the right panel.

$$= \underbrace{\big\{\mathbb{E}[Y_{i2}(0) \mid G_i = 0] - \mathbb{E}[Y_{i1}(0) \mid G_i = 0]\big\}}_{\text{Trend of the control group from } t=1 \text{ to } t=2} - \underbrace{\big\{\mathbb{E}[Y_{i1}(0) \mid G_i = 0] - \mathbb{E}[Y_{i0}(0) \mid G_i = 0]\big\}}_{\text{Trend of the control group from } t=0 \text{ to } t=1} \quad (11)$$

Here, the left-hand side represents how the outcome trends of the treatment group change between (from $t = 0$ to $t = 1$) and (from $t = 1$ to $t = 2$). The right-hand side quantifies the same change in the outcome trends for the control group.

We also emphasize an alternative way to interpret the parallel trends-in-trends assumption. Unlike the parallel trends assumption that assumes the time-invariant unmeasured confounding, the parallel trends-in-trends assumption can account for *linear time-varying* unmeasured confounding — unobserved confounding increases or decreases over time but with some constant rate. We provide examples and formal justification of this interpretation in Appendix C.3.3.

Figure 2 visually illustrates that the parallel trends-in-trends assumption holds even when the trends of the treatment and control groups are not parallel, as long as its change over time is the same. Under the parallel trends-in-trends assumption, the sequential DID estimator is unbiased and consistent for the ATT. Importantly, the extended parallel trends assumption is stronger than the parallel trends-in-trends assumption, and thus, the sequential DID estimator is unbiased and consistent for the ATT under the extended parallel trends assumption as well.

We demonstrate that a common robustness check of including group- or unit-specific time

trends (Angrist and Pischke 2008) is nonparametrically equivalent to the sequential DID estimator (see Appendix C.3). Within the potential outcomes framework, we clarified that these common techniques are justified under the parallel trends-in-trends assumption.

3 Double Difference-in-Differences

We saw in the previous section that multiple pre-treatment periods provide the three related benefits. We have clarified that each benefit requires different assumptions and estimators, and as a result, in practice, researchers tend to enjoy only a subset of the three benefits. 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 three extensions. First, we propose the double DID regression to include observed pre-treatment covariates (Section 3.3.1). Second, we generalize the proposed method to any number of *pre-* and *post*-treatment periods in the DID design (Section 3.3.2). Finally, we extend it to the staggered adoption design, where the timing of the treatment assignment can vary across units (Section 4).

3.1 Double DID via Generalized Method of Moments

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:

$$\widehat{\tau}_{\texttt{d-DID}} = \underset{\tau}{\operatorname{argmin}} \begin{pmatrix} \tau - \widehat{\tau}_{\texttt{DID}} \\ \tau - \widehat{\tau}_{\texttt{s-DID}} \end{pmatrix}^{\top} \mathbf{W} \begin{pmatrix} \tau - \widehat{\tau}_{\texttt{DID}} \\ \tau - \widehat{\tau}_{\texttt{s-DID}} \end{pmatrix} \tag{12}$$

where $\mathbf{W}$ is a 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. Table 1 illustrates that a particular choice of the weight matrix $\mathbf{W}$ recovers the standard DID, the extended DID, and the sequential DID estimators, respectively.

Using the GMM theory, we can estimate the optimal weight matrix $\widehat{\mathbf{W}}$ such that asymptotic standard errors of the double DID estimator are minimized, which we describe in detail in Section 3.1.2. Therefore, users do not need to manually pick the weight matrix $\mathbf{W}$.

We emphasize that the double DID estimator provides a unifying framework to consider identification assumptions and to estimate treatment effects within the framework of the GMM.

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

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

The double DID estimator proceeds with the following two steps.

3.1.1 Step 1: Assessing Underlying Assumptions

The first step is to assess the underlying assumptions. We use this first step to adaptively choose the weight matrix $\mathbf{W}$ in the second step. In this first step, we check the extended parallel trends assumption by applying the DID estimator on pre-treatment periods (equation (4)) and testing whether the estimate is statistically distinguishable from zero at a conventional level. To take into account correlated errors, we cluster standard errors at the level of treatment assignment.

Importantly, this step of the double DID can be viewed as the over-identification test in the GMM framework (Hansen 1982), which tests whether all the moment conditions are valid. In the context of the double DID estimator, we assume that the sequential DID estimator is correctly specified and test the null hypothesis that the standard DID estimator is correctly specified. Then, the null hypothesis of the over-identification test becomes exactly the same as testing whether an estimate of the DID estimator applied to pre-treatment periods is equal to zero.

Equivalence Approach. We note that 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 (e.g., Hartman and Hidalgo 2018) in which we evaluate the null hypothesis that trends of two groups are *not* parallel in pre-treatment periods.[4] By using 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

4. Liu, Wang, and Xu (2020) propose a similar test for a different class of estimators, what they refer to as "counterfactual estimators."

standardized interval, which can be interpreted as the standard deviation from the baseline control mean. We provide technical details in Appendix F and provide an empirical example in Section 3.4.

3.1.2 Step 2: Estimation of the ATT

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

$$\widehat{\mathbf{W}} = \begin{pmatrix} \widehat{\text{Var}}(\widehat{\tau}_{\texttt{DID}}) & \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}}) \\ \widehat{\text{Cov}}(\widehat{\tau}_{\texttt{DID}}, \widehat{\tau}_{\texttt{s-DID}}) & \widehat{\text{Var}}(\widehat{\tau}_{\texttt{s-DID}}) \end{pmatrix}^{-1} \tag{13}$$

While the double DID approach can take any weight matrix, 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. In particular, when this optimal weight matrix is used, the double DID estimator can be explicitly written as

$$\widehat{\tau}_{\texttt{d-DID}} = w_1 \widehat{\tau}_{\texttt{DID}} + w_2 \widehat{\tau}_{\texttt{s-DID}} \tag{14}$$

where $w_1 + w_2 = 1$, and

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

By pooling information from both the standard DID and sequential DID, the asymptotic variance of the double DID is smaller than or equal to variance of either the standard and sequential DIDs. This is analogous to Bayesian hierarchical models where pooling information from multiple groups makes estimation more accurate than separate estimation based on each group.

In addition, because the extended DID is a special case of the double DID (as described in Table 1), the asymptotic variance of the double DID is also smaller than or equal to variance of the extended DID. Therefore, $\text{Var}(\widehat{\tau}_{\texttt{d-DID}}) \leq \min(\text{Var}(\widehat{\tau}_{\texttt{DID}}), \text{Var}(\widehat{\tau}_{\texttt{s-DID}}), \text{Var}(\widehat{\tau}_{\texttt{e-DID}}))$. We provide the proof in Appendix D.

Following Bertrand, Duflo, and Mullainathan (2004), we estimate the variance-covariance matrix of $\widehat{\tau}_{\texttt{DID}}$ and $\widehat{\tau}_{\texttt{s-DID}}$ via block-bootstrap where the block is taken at the level of treatment

assignment. Specifically, we obtain a pair of two estimators $\{\widehat{\tau}^{(b)}_{\texttt{DID}}, \widehat{\tau}^{(b)}_{\texttt{s-DID}}\}$ for $b = 1, \ldots, B$ with B number of bootstrap iterations, and compute the empirical variance-covariance matrix. Given an estimate of the weight matrix (equation (13)), we obtain the double DID estimate as a weighted average (equation (14)). We can obtain the variance estimate of $\widehat{\tau}_{\texttt{d-DID}}$ by following the standard efficient GMM variance formula:

$$\widehat{\text{Var}}(\widehat{\tau}_{\texttt{d-DID}}) = (\mathbf{1}^\top \widehat{\mathbf{W}} \mathbf{1})^{-1},$$

where $\mathbf{1}$ is a two-dimensional vector of ones.

Remark. Under the extended parallel trends assumption, both the standard DID and the sequential DID estimator are consistent for the ATT, and thus, any weighted average is a consistent estimator. But the optimal weight matrix (equation (13)) 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 contains one moment condition $\tau - \widehat{\tau}_{\texttt{s-DID}} = 0$, and thus, it reduces 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 for the ATT without making further stringent assumptions. However, when there are more than two pre-treatment periods, researchers can also use the proposed generalized K-DID (discussed in Section 3.3.2) to further relax the parallel trends-in-trends assumption.

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 G, we use simulations to show that the double DID achieves 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 Extensions

3.3.1 Double DID Regression

Like other DID estimators, the double DID estimator has a nice connection to a regression approach. We propose the double DID regression with which researchers can include other pre-treatment covariates $\mathbf{X}_{it}$ to make the DID design more robust and efficient. We provide technical details in Appendix E.1.

3.3.2 Generalized K-Difference-in-Differences

We generalize the proposed method to *any* number of pre- and post-treatment periods in Appendix E.2, which we call K-difference-in-differences (K-DID). 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 beyond the linear time-varying unmeasured confounding under the parallel trends-in-trends assumption (Assumption 3). K-DID allows for time-varying unmeasured confounding that follows a $(K-1)$th order polynomial function when researchers have K pre-treatment periods. We can view the double DID as a special case of K-DID because in the double DID we have $K = 2$ pre-treatment periods, and it can allow for unmeasured confounding that follows $(2 - 1 = 1)$st order polynomial function (i.e., a linear function).

Second, we also allow for any number of *post*-treatment periods so that researchers can estimate not only short-term causal effects but also longer-term causal effects. This generalization can be crucial in many applications because treatment effects might not have an immediate impact on the outcome.

3.4 Empirical Application

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 thirty outcomes in Appendix H.1.

3.4.1 Visualizing and Assessing Underlying Assumptions

The first step of the DID design is to visualize trends of treatment and control groups. Figure 3 shows trends of three different outcomes: "Education and Cultural Program," "Tap Water," and "Agricultural Center."[5] 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") have 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 plausible. Based on detailed domain knowledge, Malesky, Nguyen, and Tran (2014) include

5. See Appendix H.1 for definitions.

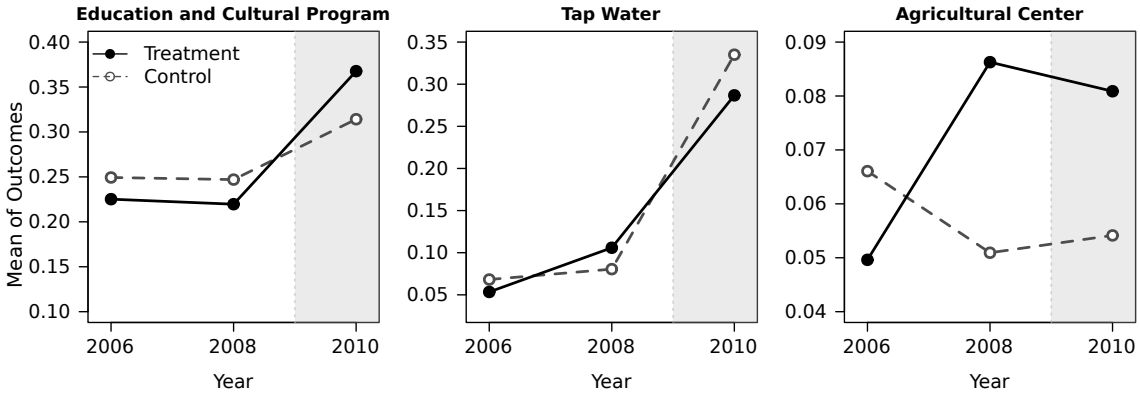

Figure 3: Visualizing Trends of Treatment and Control Groups. *Note:* We report trends for the treatment group (black solid line with solid circles) and the control group (gray dashed line with hollow circles). Two pre-treatment periods are 2006 and 2008. One post-treatment period, 2010, is indicated by the gray shaded area.

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 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 equivalence 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 3 suggests, a statistical test suggests that the extended parallel trends assumption is plausible.

For the second outcome, the test of the parallel trends reveals that the parallel trends

	Estimate	Std. Error	p-value	95% Std. Equivalence CI
Education and Cultural Program	-0.007	0.096	0.940	$[-0.166, 0.166]$
Tap Water	0.166	0.083	0.045	$[-0.302, 0.302]$
Agricultural Center	0.198	0.082	0.015	$[-0.332, 0.332]$

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.

assumption is less plausible for this outcome than for the first outcome. Finally, for the third outcome, both traditional and equivalence approaches provide little evidence for parallel trends, as graphically clear in Figure 3. 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. Upon assessing the underlying parallel trends assumptions, we now proceed to estimation of the ATT via the double DID.

3.4.2 Estimating Causal Effects

Within the double DID framework, we select appropriate DID estimators after the empirical assessment of underlying assumptions. For the first outcome, 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 4 clearly shows this pattern. In the figure, we report point estimates as well as 90% 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 $= [-0.006, 0.174]$). Using the double DID estimator, an estimate is instead 0.082 (90% CI $= [0.001, 0.163]$). 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, efficiency gain of the double DID can be even larger.

For the second outcome, 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

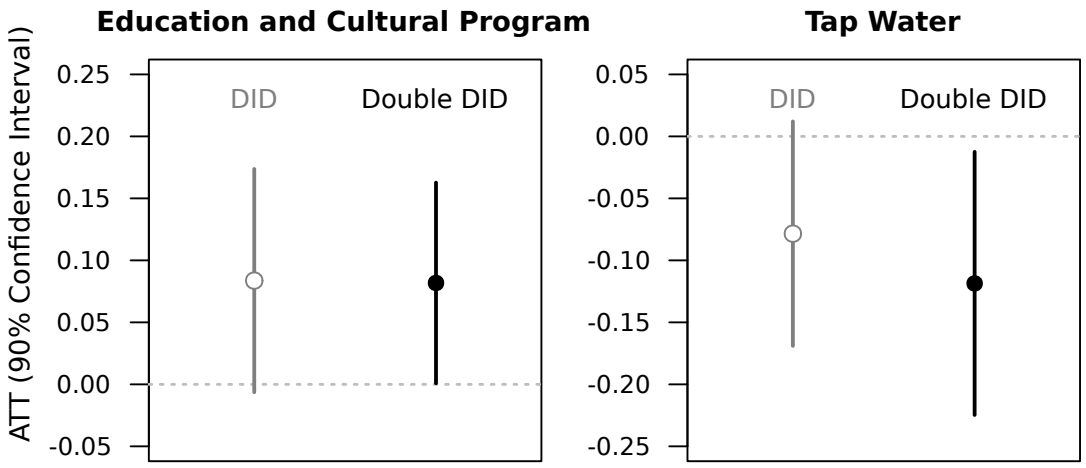

Figure 4: Estimating Causal Effects of Abolishing Elected Councils. *Note:* We compare estimates from the standard DID and the proposed double DID.

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 assumes constant unmeasured confounders. The second plot of Figure 4 shows the important difference between the two methods. While the standard DID estimates is -0.078 (90% CI $= [-0.169, 0.012]$), the double DID estimate is -0.119 (90% CI $= [-0.225, -0.012]$). 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, 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 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 H.1).

4 Staggered Adoption Design

In this section, we extend the proposed double DID estimator to the staggered adoption design where the timing of the treatment assignment can vary across units (Strezhnev 2018; Ben-Michael, Feller, and Rothstein 2019; Athey and Imbens 2021).

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 afterward. Therefore, $D_{it} = 1$ if $D_{im} = 1$ where $m < t$. We can thus summarize 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 Appendix H.2, we provide its example based on Paglayan (2019). See Figure 5 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 (Imai and Kim 2019; Athey and Imbens 2021). 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 not receive the treatment by period t. Finally, we generalize group indicator G as follows.

$$G_{it} = \begin{cases} 1 & \text{if } A_i = t \\ 0 & \text{if } A_i > t \\ -1 & \text{if } A_i < t \end{cases} \tag{15}$$

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))

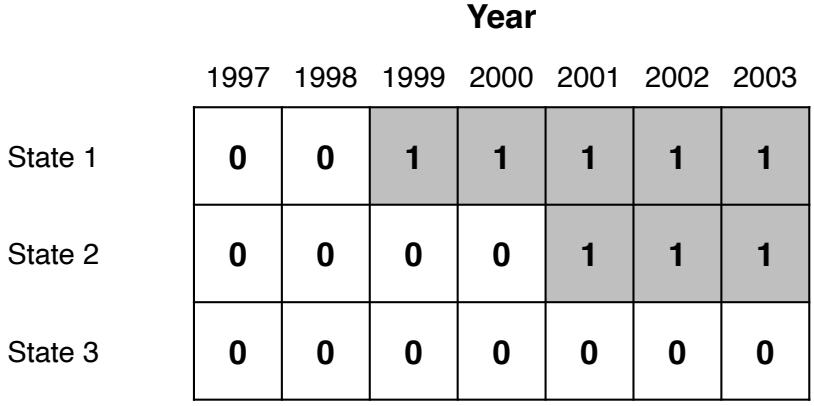

Figure 5: Example of the Staggered Adoption Design. *Note:* We use gray cells of "1" to denote the treated observation and use white cells of "0" to denote the control observation.

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 $\mathcal{T} = \{1, \ldots, T\}$. 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

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

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

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

$$- \left\{ \left(\frac{\sum_{i:\, G_{it}=1} Y_{i,t-1}}{n_{1,t-1}} - \frac{\sum_{i:\, G_{it}=1} Y_{i,t-2}}{n_{1,t-2}} \right) - \left(\frac{\sum_{i:\, G_{it}=0} Y_{i,t-1}}{n_{0,t-1}} - \frac{\sum_{i:\, G_{it}=0} Y_{i,t-2}}{n_{0,t-2}} \right) \right\},$$

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

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

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 weight matrix. Under the SA design, similar to the basic design, the standard DID and sequential DID estimators are special cases of our proposed double DID estimator with specific choices of the weight matrix. As in Section 3.1, we can estimate the optimal weight matrix $\widehat{\mathbf{W}}(t)$ (details below), and thus, users do not need to choose it manually.

Like the basic double DID estimator in Section 3.1, the double DID for the SA design also consists of two steps. The first step is to assess the underlying assumptions using the standard DID for the SA design with two points $\{t-1, t-2\}$ for units that 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 $\widehat{\mathbf{W}}(t) = \widehat{\mathrm{Var}}(\widehat{\tau}^{\mathsf{SA}}_{(1:2)}(t))^{-1}$ where $\mathrm{Var}(\cdot)$ is the variance-covariance matrix and $\widehat{\tau}^{\mathsf{SA}}_{(1:2)}(t) = (\widehat{\tau}^{\mathsf{SA}}_{\texttt{DID}}(t), \widehat{\tau}^{\mathsf{SA}}_{\texttt{s-DID}}(t))^{\top}$. 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 E.3.

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

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

where $\widehat{\overline{\tau}}^{\mathsf{SA}}_{\texttt{DID}}$ and $\widehat{\overline{\tau}}^{\mathsf{SA}}_{\texttt{s-DID}}$ are time-averages of the DID and sequential DID estimators,

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

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

5 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 uses the GMM to further improve with respect to identification and estimation accuracy. Finally, we generalize the double DID estimator to the staggered adoption design where the timing of the treatment assignment can vary across units.

References

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

Abadie, A., A. Diamond, and J. 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.

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

Athey, S., and G. W. Imbens. 2021. "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." *Journal of Econometrics.*

Ben-Michael, E., A. Feller, and J. Rothstein. 2019. "Synthetic Controls and Weighted Event Studies with Staggered Adoption." *arXiv preprint arXiv:1912.03290.*

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

Callaway, B., and P. H. Sant'Anna. 2020. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics.*

Goodman-Bacon, A. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics.*

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

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

Imai, K., and I. S. Kim. 2021. "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29 (3): 405–415.

———. 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.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

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

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

Malesky, E. J., C. V. Nguyen, and A. 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.

Marcus, M., and P. H. Sant'Anna. 2021. "The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics." *Journal of the Association of Environmental and Resource Economists* 8 (2): 235–275.

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

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

Pang, X., L. Liu, and Y. Xu. 2021. "A Bayesian Alternative to Synthetic Control for Comparative Case Studies." *Political Analysis.*

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

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