Journal of Econometrics 251 (2025) 106059

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A comparative analysis of two-way fixed effects estimators in staggered treatment designs

Jhordano Aguilar-Loyo

Department of Economics, Econometrics and Finance, University of Groningen, Nettelbosje 2, Groningen, 9747 AE, Groningen, The Netherlands

ARTICLE INFO

Keywords:
Two-way fixed effects
Difference-in-differences
Heterogeneous treatment effects
Panel data

ABSTRACT

Two-way fixed effects (TWFE) is a flexible and widely used approach for estimating treatment effects, and several TWFE estimators have been proposed for staggered treatment designs. This paper focuses on the extended TWFE estimator, introduced by Borusyak et al. (2024) and Wooldridge (2021), and compares it with alternative TWFE estimators. The main contribution is the derivation of an equation that connects the extended TWFE estimator with the difference-in-differences estimator. This equivalence provides a transparent decomposition of the components of the extended TWFE estimand. The results show that the extended TWFE estimand consists of two distinct components: one that captures meaningful comparisons and a residual term. The paper outlines the assumptions required to identify treatment effects. In line with previous literature, the findings show that the extended TWFE estimator relies on a parallel trends assumption that extends across multiple periods. Additionally, illustrative examples compare the TWFE estimators under violations of the parallel trends assumption. The results suggest that no single estimator outperforms the others. The choice of the TWFE estimator depends on the parameter of interest and the characteristics of the empirical application.

1. Introduction

Two-way fixed effects (TWFE) is a widely used method for estimating treatment effects. The TWFE regression can be described by

$$Y_{i,t} = \alpha_i + \lambda_t + \Xi_{i,t} + e_{i,t}, \tag{1}$$

where $Y_{i,t}$ is the outcome of interest for unit i at time t, α_i and λ_t are the unit and time fixed effects, respectively, $e_{i,t}$ is an unobserved error (properties specified in the following sections), and $\Xi_{i,t}$ varies depending on the specific TWFE estimator. For example, the static TWFE estimator is obtained by minimizing Eq. (1) via ordinary least squares (OLS), where $\Xi_{i,t}$ consists of a single coefficient $\beta D_{i,t}$, and the dummy variable $D_{i,t}$ equals one if unit i has received treatment at time t and zero otherwise.

In staggered treatment adoption settings, the static TWFE estimator has faced substantial criticism due to its limitations in handling heterogeneous treatment effects. As highlighted in recent work (De Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023), a causal interpretation of the static TWFE coefficient requires that the effects of treatments be constant over time and in cohorts, an assumption that is rarely met in practice. In response, researchers have developed alternative estimators that explicitly account for effect heterogeneity. For example, Borusyak et al. (2024), Wooldridge (2021) propose a TWFE estimator that minimizes Eq. (1) with $\Xi_{i,t}$ consisting of coefficients for different periods and groups, where groups are defined by the time at which units first receive treatment. This estimator is referred to as the extended TWFE estimator. Similarly, Sun and Abraham

E-mail address: jhordano.aguilar.loyo@outlook.com.

https://doi.org/10.1016/j.jeconom.2025.106059
Received 14 November 2023; Received in revised form 22 June 2025; Accepted 22 June 2025
Available online 25 July 2025

(2021) introduce an interaction-weighted estimator, referred to as the SA-TWFE estimator. However, the differences between these estimators have not yet been fully clarified.

This paper examines the extended TWFE estimator, using the SA-TWFE estimator as a reference point. To analyze the extended TWFE estimator, the paper employs a difference-in-differences (DiD) decomposition, which establishes a link between the TWFE estimator and the DiD estimator. The DiD estimator compares the changes in outcomes over time between a treated group and a control group, which offers intuitive simplicity (Cameron and Trivedi, 2005). The DiD decomposition provides a convenient framework for uncovering the structure of TWFE estimators, facilitating direct comparisons across different estimators. Prior research has derived DiD decompositions for various TWFE estimators. Goodman-Bacon (2021) established this decomposition for the static TWFE estimator, and Sun and Abraham (2021) extended it to the dynamic TWFE estimator. Alternatively, Arkhangelsky and Samkov (2024) analyze the structure of the extended TWFE estimator using a different approach. The authors establish a connection between the extended TWFE estimator and a sequential synthetic DiD estimator. This equivalence complements the DiD decomposition presented in this paper, as these two results emphasize different properties of the extended TWFE estimator.

The key contribution of this paper is the derivation of the DiD decomposition of the extended TWFE estimator. Recent literature reviews have highlighted the complexity of obtaining this decomposition in the context of staggered treatment adoption (De Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023). This decomposition demonstrates that the extended TWFE estimand consists of two types of DiDs: pre-DiD, corresponding to the period before treatment, and post-DiD, corresponding to the period after treatment.

The post-DiD term in the decomposition is a weighted average of multiple DiD terms and can be associated with a causal parameter of interest. This term enables meaningful comparisons by contrasting the target group, which undergoes treatment changes, with control groups that remain untreated. Each DiD term compares the target group to a distinct untreated control group. The weights in this average represent the conditional probability of each control group being used, given that it remains untreated. These weights reflect the relative contribution of each control group to the overall comparison. Consequently, the post-DiD term is a convex combination of sensible DiD terms.

The pre-DiD is a residual term that collects DiD terms that are unrelated to the target group and the time periods of interest. The magnitude of the pre-DiD term varies with the evaluation period. It is absent at the implementation date of the treatment, and its influence grows over time, with more components accumulating in later post-treatment periods. The pre-DiD term equals zero under a parallel trends assumption. When the assumption is violated, the pre-DiD term may accentuate the bias of the extended TWFE estimator.

The DiD decompositions facilitate a straightforward comparison between the extended TWFE estimator and the SA-TWFE estimator. This comparison can be extended to the widely used estimator proposed by Callaway and Sant'Anna (2021), referred to as the CS estimator. Previous literature (Harmon, 2022; Roth et al., 2023) establishes a connection between the SA-TWFE estimator and the CS estimator. In particular, these two estimators are equivalent when the CS estimator uses the never-treated group as a control group. Consequently, the key differences identified between the extended TWFE estimator and the SA-TWFE estimator also extend to the comparison between the extended TWFE estimator and the CS estimator.

One of the main differences between the extended TWFE estimator and the SA-TWFE estimator is the baseline of comparison. The extended TWFE estimator consists of multiple DiD terms, where the baseline of comparison spans several periods. In contrast, the SA-TWFE estimator relies on a single-period baseline. Another key distinction between these estimators is the choice of the control group. The extended TWFE estimator incorporates both the never-treated and the not-yet-treated groups as controls, whereas the SA-TWFE estimator uses only the never-treated group. Consequently, the SA-TWFE estimator relies on a smaller subset of the data compared to the extended TWFE estimator, which may lead to efficiency differences. Indeed, Borusyak et al. (2024) argue that the extended TWFE estimator is the most efficient linear unbiased estimator under the assumption of homoskedastic and serially uncorrelated errors.

Given that the SA-TWFE estimator relies on a smaller subset of the data than the extended TWFE estimator, it requires assumptions that cover a more limited portion of the data. Specifically, the SA-TWFE estimator requires a parallel trends assumption starting from the period before the implementation of the treatment. On the other hand, the extended TWFE estimator imposes restrictions for all the periods. Therefore, we can achieve potential gains in efficiency at the cost of imposing restrictions over a larger number of periods.

Although the extended TWFE estimator requires a parallel trends assumption to hold over multiple periods, this does not imply greater bias under misspecification. Prior research (such as the Online Appendix of Borusyak et al. (2024) and Harmon (2022)) has presented Monte Carlo simulation evidence indicating that there is no definitive ranking of TWFE estimators in terms of bias. This paper presents illustrative examples supporting that finding. The examples compare the estimators when the parallel trends assumption is violated. In some scenarios, the SA-TWFE estimator outperforms the extended TWFE estimator. In particular, for group-specific linear trends, a violation in the parallel trends will lead to a bias in the estimation that could be more pronounced for the extended TWFE estimator. However, there are also settings where the SA-TWFE estimator exhibits greater bias than the extended TWFE estimator. Section 6.2 analyzes nonlinear trends that violate the parallel trends assumption. The results demonstrate that for aggregate treatment effect measures – such as the average effect within a group – the extended TWFE estimator generally has lower bias than the SA-TWFE estimator. Thus, the SA-TWFE estimator is not necessarily less biased than the extended TWFE estimator.

The article is structured as follows. Section 2 introduces the event study framework and defines the key parameters of interest. Section 3 presents the extended TWFE estimator, the SA-TWFE estimator, and the CS estimator. Section 4 derives the DiD decomposition of the extended TWFE estimator. Section 5 discusses the identification assumptions required for these estimators to recover treatment effects. Section 6 provides illustrative examples. Finally, Section 7 concludes. All formal proofs are collected in the Appendix A, while additional supporting material and derivations appear in the Online Appendix.

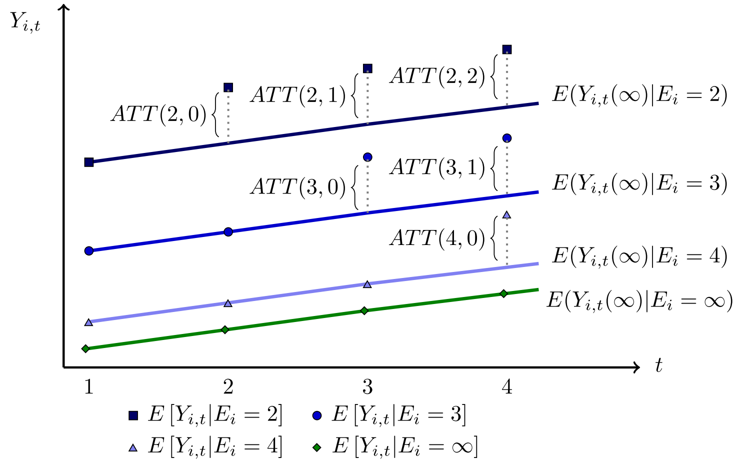

Fig. 1. Staggered Treatment and $ATT(g,\ell)$.
Note: The data points describe the expected value of the observed outcome for different groups. A line depicts the expected value if a group would not have received the treatment. The $ATT(g,\ell)$ is the difference between the data points and the line.

2. Set-up and parameters of interest

This section introduces the notation and discusses the parameter of interest. We consider a sample of N cross-sectional units that are observed over T periods, and we denote a particular unit by $i \in \{1,\dots,N\}$ and a particular time by $t \in \{1,\dots,T\}$. $D_{i,t}$ is an indicator variable that is equal to one if the unit i is treated at time t and equal to zero otherwise. We consider a staggered adoption design, where units are treated at different times, and an absorbing treatment.

Assumption 1 (*Absorbing Treatment*)**.** The treatment status is a non-decreasing sequence. At $t = 1$, $D_{i,1} = 0$ for all units. For subsequent time periods $t = 2,\dots,T$, the treatment status satisfies $D_{i,s} \le D_{i,t}$ for all $s < t$ almost surely.

Assumption 1 states that the units start receiving treatment at time 2 or later, and the treatment status is irreversible. Additionally, E_i indicates the time when unit i receives the treatment for the first time, then $E_i = \min\{t : D_{i,t} = 1\}$ and $E_i = \infty$ if unit i is untreated. Units are grouped based on the start of the treatment, and the group is denoted as g for $g \in \{2,\dots,T,\infty\}$. Concerning the distribution of the observed data, we assume a random sampling.

Assumption 2 (*Random Sampling*)**.** The observations $\{Y_{i,1}, Y_{i,2},\dots,Y_{i,T}, D_{i,1}, D_{i,2},\dots,D_{i,T}\}_{i=1}^N$ are independent and identically distributed (i.i.d.).

Assumption 2 restricts the analysis to a panel data set. To evaluate the causal effect of a treatment, we consider a potential outcome framework (Rubin, 1974; Imbens and Rubin, 2015). The potential outcome is defined as $Y_{i,t}(g)$, which corresponds to the outcome the ith subject would experience at time t if treatment starts at time $g \in \{2,\dots,T\}$, and $Y_{i,t}(\infty)$ indicates the potential outcome if the unit is not treated. The following equation relates the potential outcomes and the observed outcomes

$$Y_{i,t} = Y_{i,t}(\infty) + \sum_{g=2}^{T} \left(Y_{i,t}(g) - Y_{i,t}(\infty)\right) \mathbf{1}[E_i = g]. \tag{2}$$

A common practice is to evaluate the effect of the treatment ℓ periods after the start of the treatment. Then, the effect of the treatment is evaluated in the event period ℓ instead of the calendar time t. The parameters of interest are the average treatment effect on the treated (ATT) for different groups g and different periods ℓ.

Definition 1. The average treatment effect on the treated ℓ periods after the initial treatment for units in group g is

$$ATT(g,\ell) = \mathbb{E}\left[Y_{i,g+\ell}(g) - Y_{i,g+\ell}(\infty) | E_i = g\right]. \tag{3}$$

This definition aligns with those given by Sun and Abraham (2021) and Callaway and Sant'Anna (2021) and enables the analysis of different treatment effect dynamics. It serves as a fundamental building block for addressing policy-related questions, such as whether the impact of the treatment decays, persists, or follows more complex temporal patterns.

Fig. 1 displays the $ATT(g,\ell)$ for the first three treated groups, where treatment is assigned as follows: one group at time 2, and the remaining two groups at time 3 and 4. The data points represent the expected observed outcomes for each group, while the counterfactual line indicates the expected outcomes in the absence of treatment. Before the start of the treatment, the line and the points coincide. The $ATT(g,\ell)$ is the difference between the data points and the line, which is not observable after the treatment implementation. Estimating the $ATT(g,\ell)$ requires approximating the line that describes what the outcome would have been if the group had not received the treatment.

3. TWFE estimators and CS estimator

3.1. Extended TWFE estimator

The extended TWFE estimator was proposed by Wooldridge (2021) and Borusyak et al. (2024). This estimator is also called Imputation estimator, see De Chaisemartin and d'Haultfoeuille (2023) and Roth et al. (2023). Wooldridge (2021) discusses both balanced and unbalanced panels, but this paper focuses on balanced panels.

The extended TWFE regression includes several coefficients to model potential heterogeneity in treatment effects. The inclusion of these coefficients enhances the flexibility of the TWFE estimator by allowing for group-specific effects across different event periods ℓ. For a particular group g, the range of possible values for ℓ is $\{0,\dots,T-g\}$. The extended TWFE estimator is obtained from the OLS minimization of the following TWFE regression:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{g=2}^{T}\sum_{\ell=0}^{T-g}\beta_{g,\ell}D_{i,t}^{g,\ell} + e_{i,t}, \tag{4}$$

where $D_{i,t}^{g,\ell}$ is a dummy variable that indicates the group membership g and the number of periods ℓ relative to the treatment. Then, $D_{i,t}^{g,\ell} = D_{i,t}^{\ell} \times D_i^g$, where $D_{i,t}^{\ell} = \mathbf{1}[t - E_i = \ell]$ indicates ℓ periods relative to the treatment, and $D_i^g = \mathbf{1}[E_i = g]$ indicates the group membership. For each group g and relative period ℓ, the coefficient $\beta_{g,\ell}$ captures the dynamic treatment effect specific to group g at period ℓ. These parameters provide a flexible framework for estimating heterogeneous treatment effects across groups and periods.

3.2. SA-TWFE estimator

Sun and Abraham (2021) propose a TWFE regression specification that includes more coefficients than the extended TWFE regression in (4). In particular, the TWFE regression associated with the SA-TWFE estimator is given by:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{g=2}^{T}\sum_{\ell\neq -1}\delta_{g,\ell}D_{i,t}^{g,\ell} + \epsilon_{i,t}. \tag{5}$$

For each treated group g, the model includes leads and lags of the treatment, $\ell \in \{1-g,\dots,-2,0,1,\dots,T-g\}$, resulting in $T-1$ coefficients per group. Eq. (5) excludes dummy variables for the never-treated group and for $\ell = -1$ across all treated groups. Intuitively, this implies that the baseline comparison period corresponds to one period before treatment begins, with the never-treated group serving as the reference group.

The coefficients, $\delta_{g,\ell}$, for $\ell \geq 0$ can be used to estimate treatment effects for group g at period ℓ. Conversely, the coefficients $\delta_{g,\ell}$ for $\ell < 0$ measure differences in pre-treatment trends between group g and the never-treated group. These pre-treatment coefficients are commonly used to assess whether trends between group g and the never-treated group are parallel, which is a key condition for identifying treatment effects (see Section 5). However, testing the parallel trends assumption in this way has limitations. For a detailed discussion of this issue and potential solutions, see Roth (2022) and Rambachan and Roth (2023). Additionally, Borusyak et al. (2024) propose an approach to test the parallel trends assumption, which can be implemented in conjunction with the extended TWFE estimator. This test addresses the concerns raised in Roth (2022).

3.3. CS estimator

Callaway and Sant'Anna (2021) propose an intuitive approach to estimate $ATT(g,\ell)$. This estimator identifies $ATT(g,\ell)$ by comparing the expected change in outcomes for group g between periods $g-1$ and $g+\ell$ to the corresponding change for a control group that has not yet been treated at time $g+\ell$. This can be expressed as:

$$\hat{\phi}_{g,\ell} = \frac{1}{N_g}\sum_{i=1}^{N}(Y_{i,g+\ell} - Y_{i,g-1})\mathbf{1}[E_i = g] - \frac{1}{\sum_{h\in\mathcal{G}_{comp}} N_h}\sum_{h\in\mathcal{G}_{comp}}\sum_{i=1}^{N}(Y_{i,g+\ell} - Y_{i,g-1})\mathbf{1}[E_i = h], \tag{6}$$

where $\mathcal{G}_{comp}$ is a set that contains the comparison groups and N_g represents the number of units in group g. Callaway and Sant'Anna (2021) consider two options for comparison groups: the never-treated group or the not-yet-treated group. The existing literature has documented the equivalence between the CS estimator and the SA-TWFE estimator (see Roth et al. (2023)). Specifically, these estimators are equivalent when the CS estimator uses the never-treated group as the control group. Further details on this equivalence are provided in Section 4.2.

4. Difference-in-differences decompositions

4.1. Extended TWFE

In a simplified setting with only two periods and two groups (treated and untreated), the DiD decomposition is a straightforward result. However, in more complicated designs with multiple periods and variations in treatment timing, the relation between the TWFE estimator and the DiD estimator is more complex. The DiD decomposition of the extended TWFE estimator is given in the following theorem.

Theorem 1 (*Extended TWFE*). *Suppose that Assumptions 1 and 2 hold. The population regression coefficient $\beta_{g,\ell}$ in the extended TWFE specification (4) is a weighted average of difference-in-differences of population means, such that*

$$\beta_{g,\ell} = \sum_{h>g+\ell} w_{h,g+\ell} \frac{\sum_{t=1}^{g-1} DD_{g,h,g+\ell,t}}{g-1} - \psi_{g,\ell}, \tag{7}$$

where

$$DD_{g,h,t^*,t} = \mathbb{E}[Y_{i,t^*} - Y_{i,t} | E_i = g] - \mathbb{E}[Y_{i,t^*} - Y_{i,t} | E_i = h] \tag{8}$$

is the difference-in-differences of population means, and

$$w_{h,g+\ell} = \Pr(E_i = h | E_i > g + \ell)$$

represents the weights. For $\ell = 0$, we have $\psi_{g,\ell} = 0$. For $\ell > 0$,

$$\psi_{g,\ell} = \sum_{h>g}^{g+\ell} \tilde{w}_h \sum_{t=g}^{h-1} \left[\sum_{k>h} w_{k,h} \frac{\sum_{r=1}^{g-1} DD_{h,k,t,r}}{g-1} \right], \tag{9}$$

where $\tilde{w}_h = \frac{1}{h-1} \Pr(E_i = h | E_i \geq h)$.

The proof of Theorem 1 is discussed in Appendix A. Theorem 1 shows that $\beta_{g,\ell}$ consists of two components: a post-DiD (the first term on the right-hand side of (7)) and a pre-DiD (the second term on the right-hand side of (7)). For the post-DiD term, the time of evaluation is $g + \ell$, which is a post-treatment time for group g. For the pre-DiD $\psi_{g,\ell}$, the time of evaluation is from g to $h - 1$, which is before the start of the treatment for group h. It is important to note that $\psi_{g,\ell}$ only arises when there are multiple groups treated at different periods. In the simplified case of two groups (one treated and one untreated), $\psi_{g,\ell}$ does not appear.

The post-DiD aligns with the intuition that $\beta_{g,\ell}$ (under certain assumptions) can identify $ATT(g,\ell)$. This post-DiD term is a weighted sum of difference-in-differences, where group g is compared with group h (which was not treated at or before time $g+\ell$). The baseline comparison spans from time 1 to time $g - 1$. The weight $w_{h,g+\ell}$ represents the conditional probability that a unit is treated at time h, given that the unit was not treated at or before time $g + \ell$. These weights reflect the relative contribution of each control group to the overall comparison. Then, the post-DiD is a convex combination of DiD terms related to the target group g and period of interest ℓ.

The post-DiD term is a general expression that aligns with results reported in previous literature for simplified scenarios. For instance, consider the case of multiple periods ($T > 2$) with two groups: one never-treated group and one treated group that begins receiving treatment at time $g > 1$. In this case, the pre-DiD $\psi_{g,\ell}$ equals zero, and the difference-in-differences decomposition in (7) simplifies to $\beta_{g,\ell} = \frac{1}{g-1} \sum_{t=1}^{g-1} DD_{g,\infty,g+\ell,t}$. Wooldridge (2021), and De Chaisemartin and d'Haultfoeuille (2023) show the same expression for this simplified scenario.

The post-DiD term compares group g against both the not-yet-treated and never-treated groups. This comparison avoids the "forbidden comparison" problem identified by Borusyak and Jaravel (2017), where already-treated groups are erroneously used as controls. As ℓ increases, the size of the not-yet-treated group decreases. The left panel of Fig. 2 illustrates this inverse relationship between ℓ and the size of the not-yet-treated group. When $\ell = 0$, group g is compared against the not-yet-treated group (comprising groups $g + 1$ to T) and the never-treated group. For $\ell = 1$, the not-yet-treated group is reduced to include only groups $g + 2$ to T. At the maximum value of $\ell = T - g$, the comparison group consists solely of the never-treated group.

The pre-DiD $\psi_{g,\ell}$ is a residual term that aggregates pre-DiDs for groups treated after period g but not later than $g + \ell$. These groups cannot serve as valid comparison groups at period $g + \ell$ due to the forbidden comparison problem. The pre-DiD $\psi_{g,\ell}$ is negligible under the parallel trends assumption, as discussed in Section 5. However, if the parallel trends assumption is violated, $\psi_{g,\ell}$ may amplify the bias. Section 6.1 provides an example in which the extended TWFE estimator exhibits greater bias than the CS estimator, and the pre-DiD $\psi_{g,\ell}$ partially explains this performance discrepancy.

For $\ell = 0$, the pre-DiD $\psi_{g,\ell}$ is absent, but for $\ell > 0$, $\psi_{g,\ell}$ accumulates pre-DiD terms from different groups and comparison periods. The right panel of Fig. 2 illustrates how $\psi_{g,\ell}$ evolves as ℓ increases, with squares indicating the comparison periods for each pre-DiD term. When $\ell = 1$, the pre-DiD compares group $g + 1$ against groups $g + 2$ to T and the never-treated group. This comparison is conducted at time g, with baseline spanning from time 1 to $g - 1$. For $\ell = 2$, an additional pre-DiD term is added, comparing group $g+2$ against groups $g+3$ to T and the never-treated group. Moreover, the comparison is performed at time g and $g+1$. As ℓ increases further, $\psi_{g,\ell}$ accumulates more such terms, with the final pre-DiD comparing group T against the never-treated group across times g to $T - 1$.

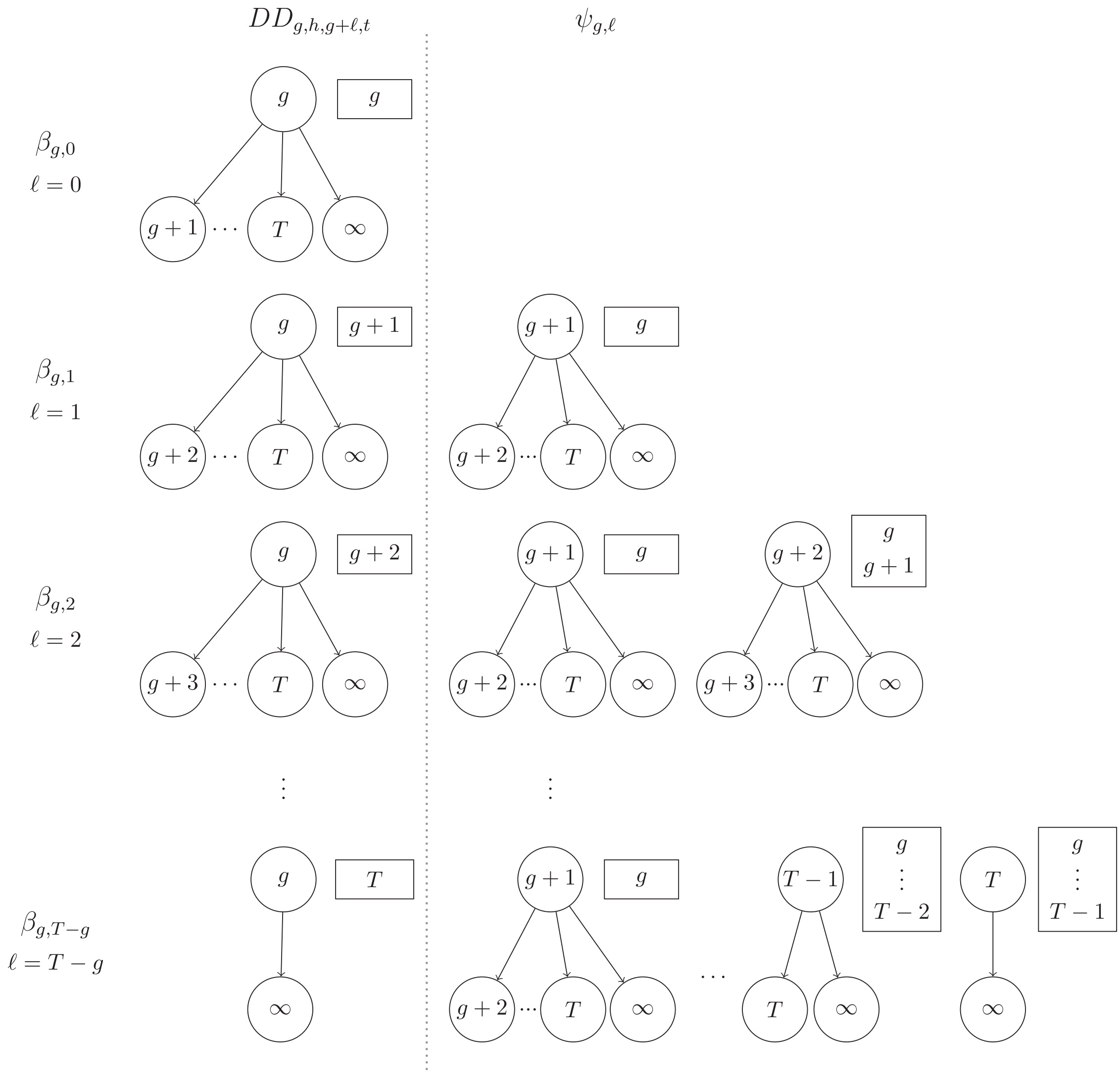

Fig. 2. Difference-in-differences decomposition of the extended TWFE.
Note: The graph illustrates Eq. (7) for different values of ℓ. Each DiD is represented as a link between two circles, where each circle corresponds to a group. All DiD comparisons share the same baseline, covering periods 1 to $g-1$. The periods in which comparisons are made are indicated inside squares. The left side of the graph represents the post-DiD term (the first term on the right-hand side of (7)), while the right side displays the pre-DiD $\psi_{g,\ell}$ (the second term on the right-hand side of (7)).

4.2. SA-TWFE and CS

Regarding the DiD decomposition of the SA-TWFE estimator, Lemma 1 provides the decomposition.

Lemma 1 (*SA-TWFE*). *Suppose that Assumptions 1 and 2 hold. The population regression coefficient $\delta_{g,\ell}$ in the SA-TWFE specification (5) can be expressed as*

$$\delta_{g,\ell} = \mathbb{E}[Y_{i,g+\ell} - Y_{i,g-1} | E_i = g] - \mathbb{E}[Y_{i,g+\ell} - Y_{i,g-1} | E_i = \infty]. \tag{10}$$

Lemma 1 shows that $\delta_{g,\ell}$ corresponds to a difference-in-differences of population means, comparing group g with the never-treated group at times $g+\ell$ and $g-1$. This result, originally discussed in Sun and Abraham (2021), is intuitive; however, the authors do not provide a formal proof. For completeness, Appendix A presents the proof of Lemma 1.

A natural approach to estimate $\delta_{g,\ell}$ is to replace expectations with their sample analogs, then $\hat{\delta}_{g,\ell} = \left(\overline{Y}_{g,g+\ell} - \overline{Y}_{g,g-1}\right) - \left(\overline{Y}_{\infty,g+\ell} - \overline{Y}_{\infty,g-1}\right)$ where $\overline{Y}_{g,t}$ denotes the sample mean of group g at time t. This result confirms that the SA-TWFE estimator and the CS estimator are equal when the CS estimator uses the never-treated group as control group.

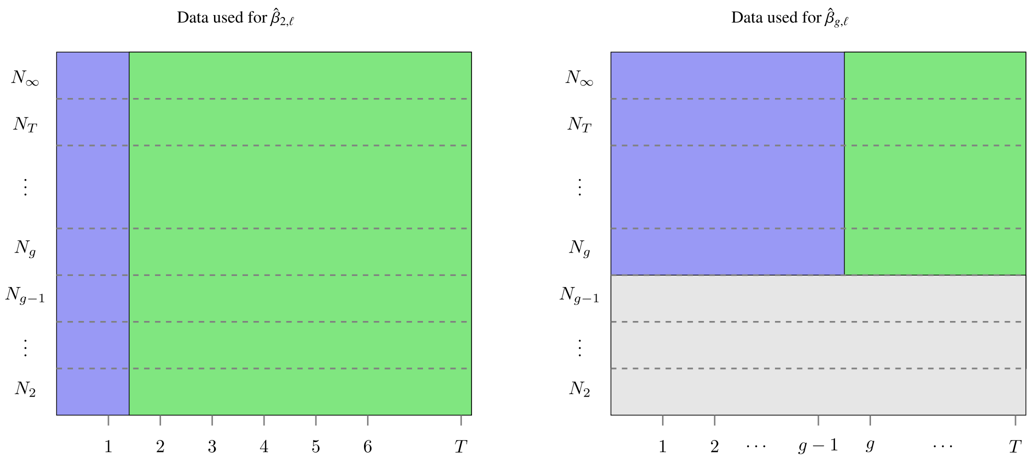

Fig. 3. Data used by the extended TWFE estimator.
Note: The graph shows the allocation of the data for the extended TWFE estimator. The blue area represents the baseline of the DiD, and the green area depicts the periods where the treatment can be evaluated. The gray area shows the fraction of the data that is not used in the computation of the DiD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. *Comparing extended TWFE, SA-TWFE and CS*

The DiD decompositions enable a straightforward comparison between the extended TWFE, the SA-TWFE, and the CS estimators. A key distinction lies in their baselines for comparison. The SA-TWFE and CS estimators consider as a baseline only the period before the implementation of the treatment. In contrast, the extended TWFE estimator considers a broader baseline, which covers from the first period up to one period before the treatment starts. Then, these estimators use different subsets of the data to estimate treatment effects, with the specific subsets varying by group.

Fig. 3 illustrates the data allocation for the extended TWFE estimator. For the group treated at time 2, the baseline for the DiD comparison corresponds to the first period, while treatment effects are estimated from time 2 onward until T. More generally, for a group treated at time g (right panel of Fig. 3), the baseline spans from time 1 to $g-1$, and treatment effects are evaluated from g to T. Moreover, information from groups treated between time 2 and $g-1$ is excluded, as these groups have already received treatment and cannot serve as control groups.

Fig. 4 presents an analogous graph of data allocation for the CS estimator. For the group treated at time 2, both the extended TWFE and CS estimators rely on the same subset of the data. However, for $g > 2$, Fig. 3 and Fig. 4 reveal diverging patterns of data utilization, with the discrepancy becoming more pronounced in later periods.

The use of a multi-period baseline for comparison in the extended TWFE estimator may lead to efficiency gains. As demonstrated by Borusyak et al. (2024), the extended TWFE estimator is the most efficient linear unbiased estimator under the assumption of homoskedastic and serially uncorrelated errors. Although efficiency gains can be expected for the extended TWFE estimator, it requires a parallel trends assumption spanning a longer time horizon than the CS estimator, as discussed in the following section.

5. Identification assumptions

This section discusses the two key identification assumptions: the no anticipatory behavior and the parallel trends assumption. The first identification assumption imposes a restriction on potential outcomes before treatment begins. In particular, it is required that individuals do not have anticipatory behavior.

Assumption 3 (*No Anticipatory Behavior*)**.** There is no treatment effect in pre-treatment periods,

$$\mathbb{E}\left[Y_{i,g+\ell}(g) - Y_{i,g+\ell}(\infty) | E_i = g\right] = 0 \text{ for all } g \text{ and all } \ell < 0. \tag{11}$$

Assumption 3 states that the treatment has no causal effect before it is implemented. For a detailed discussion of this assumption, see Callaway and Sant'Anna (2021) and De Chaisemartin and d'Haultfoeuille (2023).

The following results establish the connection between the estimand and the parameter of interest, $ATT(g,\ell)$. For the extended TWFE estimand, the following result holds:

Corollary 1. *Under Assumptions 1–3,*

$$\beta_{g,\ell} = ATT(g,\ell) + WCTD(g,\ell,g+\ell,g-1) - \sum_{h>g}^{g+\ell} \sum_{t=g}^{h-1} \tilde{\omega}_h WCTD(h,0,t,g-1), \tag{12}$$

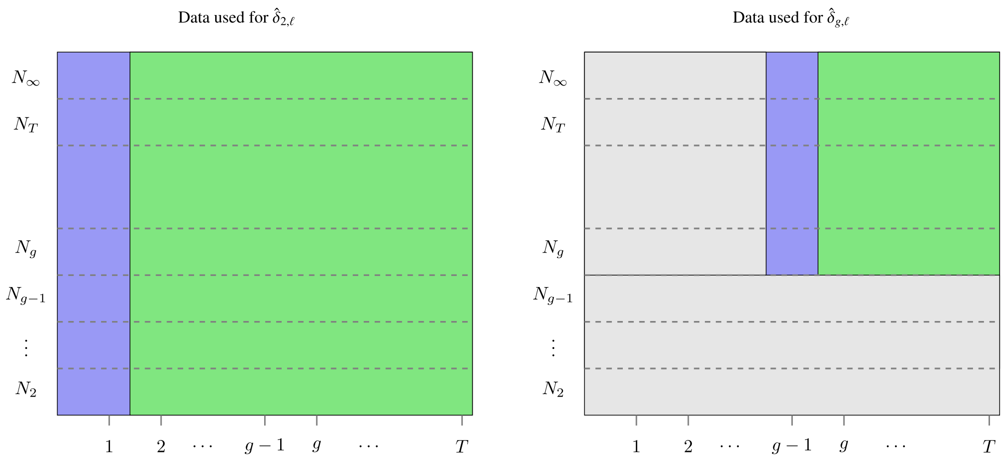

Fig. 4. Data used by the CS estimator.
Note: The graph shows the allocation of the data for the extended TWFE estimator. The blue area represents the baseline of the DiD, and the green area depicts the periods where the treatment can be evaluated. The gray area shows the fraction of the data that is not used in the computation of the DiD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where

$$WCTD(g, j, t^*, s) = \sum_{h>g+j} w_{h,g+j} \frac{\sum_{t=1}^{s} \left(\Delta Y^{(g)}_{t^*,t} - \Delta Y^{(h)}_{t^*,t} \right)}{s}, \tag{13}$$

with trends $\Delta Y^{(g)}_{t^*,t} = \mathbb{E}\left[Y_{i,t^*}(\infty) - Y_{i,t}(\infty) | E_i = g\right]$.

Corollary 1 can be obtained from Theorem 1, see the proof in the appendix. It is important to mention that Corollary 1 does not require a parallel trends assumption of any form. Corollary 1 shows that $\beta_{g,\ell}$ equals the parameter of interest, $ATT(g,\ell)$, plus the sum of two types of weighted common trend differences ($WCTD$): one associated with post-treatment periods and the other with pre-treatment periods. The first type of $WCTD$, $WCTD(g, \ell, g+\ell, g-1)$, compares the trends between group g against groups that did not receive a treatment at time $g+\ell$. This WCTD is derived from the post-DiD term in Theorem 1. The second type of $WCTD$, $\sum_{h>g}^{g+\ell} \sum_{t=g}^{h-1} \tilde{w}_h WCTD(h, 0, t, g-1)$, evaluates pre-treatment differences in trends for groups that receive the treatment after time g but not later than time $g+\ell$. This term is associated with the pre-DiD $\psi_{g,\ell}$ in Theorem 1.

Corollary 1 makes explicit the mapping between the difference in trends to bias. Unbiased estimation of $ATT(g,\ell)$ requires that all $WCTD$s in (12) are zero. Similarly, Goodman-Bacon (2021) finds the equivalence of Corollary 1 for the static TWFE estimator.

For the SA-TWFE estimator, we have the following result that connects the estimand with the parameter of interest.

Corollary 2. *Under Assumptions 1–3,*

$$\delta_{g,\ell} = ATT(g,\ell) + \Delta Y^{(g)}_{g+\ell,g-1} - \Delta Y^{(\infty)}_{g+\ell,g-1}. \tag{14}$$

Corollary 2 follows directly from Lemma 1. This corollary presents an intuitive result, and its formal derivation was originally established in Sun and Abraham (2021). A detailed discussion of Corollary 2 can be found in Sun and Abraham (2021). With respect to the CS estimator, we have the following result.

Corollary 3. *Under Assumptions 1–3,*

$$\phi_{g,\ell} = ATT(g,\ell) + \sum_{h\in\mathcal{G}_{comp}} d_h \left(\Delta Y^{(g)}_{g+\ell,g-1} - \Delta Y^{(h)}_{g+\ell,g-1} \right), \tag{15}$$

where $d_h = \Pr(E_i = h | h \in \mathcal{G}_{comp})$.

Corollary 3 can be derived from (6), establishing a connection between differences in trends and the resulting bias. This result is discussed in Callaway and Sant'Anna (2021).

The novel contribution of this paper is Corollary 1, which demonstrates that trend differences can induce bias with varying directions and magnitudes. A key feature of the WCTD terms in Corollary 1 is that the comparison baseline extends from time 1 to $g-1$, which has important implications for the required parallel trends assumption.

To identify treatment effects, researchers typically impose a parallel trends assumption, as discussed in Callaway and Sant'Anna (2021), Wooldridge (2021), Sun and Abraham (2021), and Marcus and Sant'Anna (2021). For the extended TWFE estimator, the appropriate identifying assumption is formalized in Assumption 4.

Assumption 4 (*General Parallel Trends*)**.** For all g and $t \geq 2$,

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) | E_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) | E_i \geq g\right]. \tag{16}$$

For the SA-TWFE and CS estimators, we have a parallel trends assumption that covers fewer periods.

Assumption 5 (*Parallel Trends - Post-Treatment Only*)**.** For all g and $t \geq g$,

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) | E_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) | E_i \geq g\right]. \tag{17}$$

Assumption 4 requires that the parallel trends assumption holds for $t \geq 2$, and Assumption 5 imposes the parallel trends assumption for a smaller interval ($t \geq g$). Assumptions 4 and 5 are sufficient conditions that guarantee the identification of the parameter of interest $ATT(g,\ell)$. However, these assumptions are not necessary when considering aggregate measures of treatment effects, such as the average effect within a particular group. Section 6 examines a scenario where Assumption 4 does not hold, yet the extended TWFE estimator can recover aggregated treatment effects.

In some cases, relying on Assumption 5 instead of Assumption 4 can offer greater robustness to violations of the parallel trends assumption. Section 6.1 presents an example that illustrates this situation. Given that the extended TWFE estimator may be more efficient, this leads to a trade-off between bias and efficiency. This trade-off has been discussed in the existing literature. Marcus and Sant'Anna (2021) document a 'robustness' vs. 'efficiency' trade-off in terms of the strength of the parallel trends assumption. Similarly, Roth et al. (2023) compare the imputation estimator of Borusyak et al. (2024) (the extended TWFE estimator) against the CS estimator, arguing that there is a trade-off between efficiency and the strength of the identifying assumption. The authors state that the imputation estimator is more efficient than the CS estimator; however, it could be more biased than the CS estimator since it relies on a parallel trends assumption that covers more periods. However, Assumption 4 does not necessarily imply greater sensitivity to violations of the parallel trends assumption. Section 6.2 presents a setting in which the extended TWFE estimator is less biased than the other two estimators when estimating aggregated measures of treatment effects.

6. Illustrative examples

This section compares the estimators under different trend specifications. The first subsection examines group-specific linear trends. In this setting, the extended TWFE estimator is more prone to bias. The second subsection explores nonlinear trends that violate the parallel trends assumption. The results indicate that certain coefficients of the SA-TWFE and CS estimators remain unbiased. Moreover, for aggregate measures of treatment effects, the extended TWFE estimator performs better, as it generally exhibits lower bias than the other two estimators.

6.1. Group-specific linear trends

In this and the subsequent subsection, we assume that Assumption 3 holds. This section considers group-specific linear trends, such that:

$$E\left[Y_{i,t}(\infty) | E_i = g\right] = c_g + \gamma_g t. \tag{18}$$

Each group follows a linear trend with slope γ_g. Using Corollaries 1–3, for $\ell \geq 0$, we obtain:

$$\beta_{g,\ell} = ATT(g,\ell) + \sum_{h>g+\ell} w_{h,g+\ell}\left(\frac{g}{2}+\ell\right)\left(\gamma_g - \gamma_h\right) - \sum_{h>g}^{g+\ell}\sum_{t=g}^{h-1} \tilde{w}_h \sum_{r>h} w_{r,h}\left(t - \frac{g}{2}\right)\left(\gamma_h - \gamma_r\right), \tag{19}$$

$$\delta_{g,\ell} = ATT(g,\ell) + (1+\ell)\left(\gamma_g - \gamma_h\right), \tag{20}$$

$$\phi_{g,\ell} = ATT(g,\ell) + \sum_{h\in\mathcal{G}_{comp}} d_h (1+\ell)\left(\gamma_g - \gamma_h\right). \tag{21}$$

Assumptions 4 and 5 require that $\gamma_g = \gamma_h$ for all g and h. As Eqs. (19) to (21) show, the three estimators identify $ATT(g,\ell)$ if and only if $\gamma_g = \gamma_h$.

A violation of the parallel trends assumption introduces bias in the estimation for all three estimators, with the extended TWFE estimator being the most affected. To illustrate this, consider a case where group $g+f$ has a different slope γ_{g+f} than the other groups, such that $\gamma_g > \gamma_{g+f}$. This situation represents a scenario where the parallel trends assumption holds for most groups but fails for an outlier group that evolves differently because of political or social conditions. The associated bias of $\beta_{g,\ell}$ is given by:

$$\beta_{g,\ell} - ATT(g,\ell) = \begin{cases} w_{g+f,g+\ell}\left(\gamma_g - \gamma_{g+f}\right)\frac{g}{2} & \text{for } \ell = 0 \\ w_{g+f,g+\ell}\left(\gamma_g - \gamma_{g+f}\right)\left(\frac{g}{2}+\ell\right) - \zeta_{g+\ell} & \text{for } 1 \leq \ell < f \\ \vartheta_{g+f} - \zeta_{g+f-1} & \text{for } f \leq \ell \leq T-g, \end{cases}$$

where $\zeta_{g+\ell} = \sum_{h>g}^{g+\ell} \tilde{w}_h \sum_{t=g}^{h-1} w_{g+f,h}\left(t - \frac{g}{2}\right)\left(\gamma_h - \gamma_{g+f}\right)$, and $\vartheta_{g+f} = \tilde{w}_{g+f} \sum_{t=g}^{g+f-1} \sum_{r>g+f} w_{r,g+f}\left(t - \frac{g}{2}\right)\left(\gamma_r - \gamma_{g+f}\right)$.

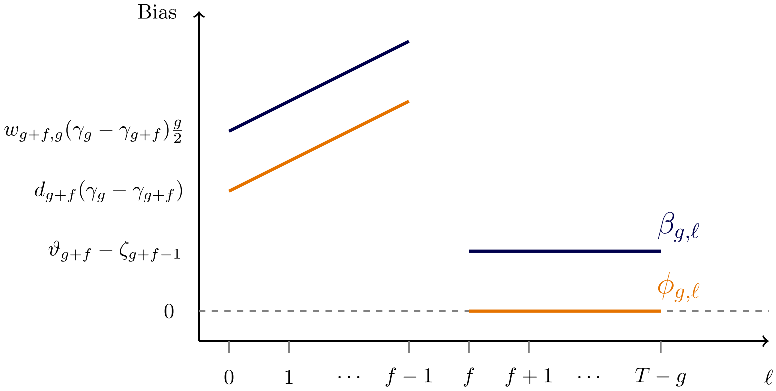

Fig. 5. Bias for group-specific linear trends.
Note: The graph displays the bias of the extended TWFE and CS estimators. The setting considers group-specific linear trends, with an outlier group $g+f$ that has a different slope, $\gamma_{g+f} < \gamma_g$. The graph shows that the extended TWFE estimator exhibits greater bias than the CS estimator across all values of ℓ.

For $1 \leq \ell < f$, the bias increases with ℓ. This bias arises mainly from the post-DiD term in Corollary 1. Additionally, for $f \leq \ell \leq T-g$, the bias is equal to $\vartheta_{g+f} - \zeta_{g+f-1}$, derived from the pre-DiD term in Corollary 1. The SA-TWFE estimator remains unbiased for all groups except $g+f$, as this estimator uses the never-treated group as the control group. Regarding the CS estimator, the bias is given by:

$$\phi_{g,\ell} - ATT(g,\ell) = \begin{cases} d_{g+f}(\gamma_g - \gamma_{g+f})(1+\ell) & \text{for } 0 \leq \ell < f \\ 0 & \text{for } f \leq \ell \leq T-g. \end{cases}$$

The results show that the extended TWFE estimator is more biased than the CS estimator for large values of g[1], which is illustrated in Fig. 5. The figure shows that the extended TWFE estimator exhibits greater bias than the CS estimator across all values of ℓ. In particular, for $f \leq \ell \leq T-g$, the CS estimator outperforms the extended TWFE estimator regardless of the value of g, since $\phi_{g,\ell}$ is unbiased and the bias of $\beta_{g,\ell}$ equals $\vartheta_{g+f} - \zeta_{g+f-1}$.

6.2. Counterexample with nonlinear trends

In this section, we examine nonlinear trends that violate the parallel trends assumption. For all treated groups, we have that

$$E\left[Y_{i,t}(\infty)|E_i = g\right] = c_g + \gamma\lceil\frac{t}{2}\rceil. \tag{22}$$

However, for the never-treated group, we assume that

$$E\left[Y_{i,t}(\infty)|E_i = \infty\right] = c_\infty + \gamma\lceil\frac{t-1}{2}\rceil. \tag{23}$$

Consequently, Assumptions 4 and 5 do not hold. However, the trends in (22) and (23) satisfy the "parallel trends at horizon h" condition described in the Online Appendix of Harmon (2022). The author emphasizes that this form of the parallel trends assumption has empirical relevance, as it accommodates settings with seasonal data. In such cases, the parallel trends condition may not hold in every period, but it does hold after h periods. In the context of this section, the parallel trends hold at even time periods.

Furthermore, we assume that the CS estimator considers the never-treated group as the control group. Under this assumption, the CS estimator and the SA-TWFE estimator are equivalent. From Corollaries 1 and 2, we derive the following result:

$$\beta_{g,\ell} = ATT(g,\ell) + w_{\infty,g+\ell}\left(\lceil\frac{g+\ell}{2}\rceil - \lceil\frac{g+\ell-1}{2}\rceil - \frac{1}{g-1}\lceil\frac{g-1}{2}\rceil\right)\gamma - \sum_{h>g}^{g+\ell}\tilde{w}_h w_{\infty,h}\left(\lceil\frac{h-1}{2}\rceil - \frac{h-1}{g-1}\lceil\frac{g-1}{2}\rceil\right)\gamma, \tag{24}$$

$$\delta_{g,\ell} = ATT(g,\ell) + \left(\lceil\frac{g+\ell}{2}\rceil - \lceil\frac{g+\ell-1}{2}\rceil - \lceil\frac{g+1}{2}\rceil + \lceil\frac{g}{2}\rceil\right)\gamma. \tag{25}$$

To simplify the analysis, we assume that units are treated in odd periods. Under this assumption, the third term on the right-hand side of Eq. (24) is equal to zero. The bias of $\beta_{g,\ell}$ fluctuates between $-0.5w_{\infty,g+\ell}\gamma$ and $0.5w_{\infty,g+\ell}\gamma$. For large values of ℓ, where

[1] For $0 \leq \ell < f$, the bias of the CS estimator is smaller than that of the extended TWFE estimator when g satisfies the condition

$$g > 2 + \sum_{h>g}^{g+\ell} Pr(E_i = h|E_i \geq h)\frac{w_{g+f,h}}{w_{g+f,g+\ell}}(h-g).$$

To derive this condition, it is assumed that the CS estimator uses both untreated and not-yet-treated groups as controls, which implies $w_{g+f,g+\ell} = d_{g+f}$.

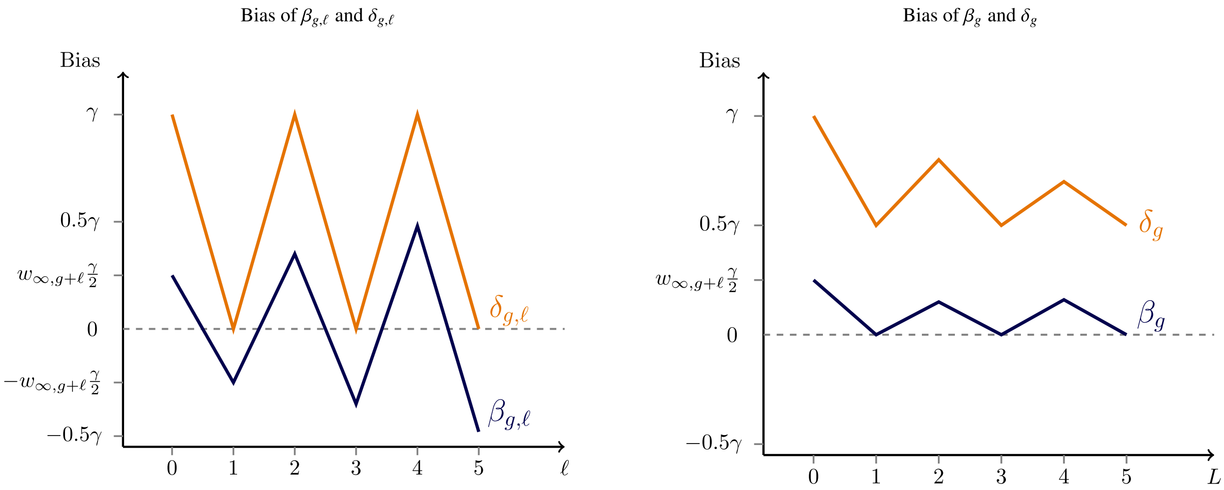

Fig. 6. Bias when parallel trends assumption is violated.
Note: The graph shows the bias of the extended TWFE and SA-TWFE estimators. The left panel shows the bias for individual estimations of the treatment effects, and the right panel shows the associated bias for aggregated estimations of the treatment effect.

$w_{\infty,g+\ell}$ approaches one, the bias fluctuates between -0.5γ and 0.5γ. The bias of $\delta_{g,\ell}$ is zero for odd values of ℓ and equal to γ for even values of ℓ.

The left panel of Fig. 6 shows the bias of $\beta_{g,\ell}$ and $\delta_{g,\ell}$ across different values of ℓ. The graph reveals that for $\ell = 0$ and even values of ℓ, the bias of $\beta_{g,\ell}$ is smaller than that of $\delta_{g,\ell}$. In contrast, for odd values of ℓ, $\delta_{g,\ell}$ remains unbiased. Thus, despite the failure of the parallel trends assumption, $\delta_{g,\ell}$ remains unbiased for certain values of ℓ.

To account for an aggregate measure of treatment effects, such as the average effect within a group, we define $\beta_g = \frac{1}{L+1}\sum_{\ell=0}^{L}\beta_{g,\ell}$ and $\delta_g = \frac{1}{L+1}\sum_{\ell=0}^{L}\delta_{g,\ell}$ where $L \in \{0, 1, \dots, T-g\}$. The corresponding bias of these aggregated measures is displayed on the right-hand side of Fig. 6. The graph shows that the extended TWFE estimator outperforms the other estimators. Moreover, for odd values of L, the extended TWFE estimator remains unbiased. This result contrasts with the findings in Section 6.1 and demonstrates that the CS and SA-TWFE estimators are not necessarily more robust against violations of the parallel trends assumption.

7. Conclusion

This paper focuses on staggered treatment adoption and evaluates the extended TWFE estimator and compares it against the SA-TWFE and CS estimators. The paper uses the DiD decomposition as the primary tool to compare these estimators. The DiD decomposition of the extended TWFE estimator reveals that its coefficients consist of two types of DiD terms: one associated with post-treatment periods and the other with pre-treatment periods. The post-DiD term is a convex combination of meaningful DiD comparisons and, under certain assumptions, can identify a parameter of interest.

The comparison of the estimators highlights a trade-off between potential efficiency gains and the scope of the parallel trends assumption. The extended TWFE estimator uses a larger subset of the data than the SA-TWFE estimator, which may enhance efficiency. However, this also requires imposing a parallel trends assumption over a longer time span. Importantly, the extended TWFE estimator is not necessarily less robust to violations of the parallel trends assumption. The examples in Section 6 demonstrate that no estimator outperforms the others in terms of bias.

Applied researchers can choose among different TWFE estimators to evaluate staggered treatments. The selection depends on the parameter of interest and the characteristics of the empirical setting. When group-specific linear trends adequately capture the underlying data structure, the SA-TWFE estimator is preferable, as it offers greater robustness to violations of the parallel trends assumption. Conversely, if the objective is to estimate an aggregate measure of treatment effects and nonlinear trends are expected, the extended TWFE estimator could be more suitable, as it can identify aggregated treatment effects when the parallel trends assumption holds after a certain time interval.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I sincerely thank the Co-Editor Aureo de Paula, the Associate Editor, and two anonymous referees for their insightful comments and suggestions. I also thank Tom Boot, Clément de Chaisemartin, Giovanni Mellace, Michaela Kesina, Damian Kozbur, Artūras Juodis, Lammertjan Dam, Aico van Vuuren, Annette Bergemann, Agnieszka Postepska, Peter van Santen, Tom Wansbeek, Hans Ligtenberg, Ioanna Tziolas and Youel Rojas Zea for comments that substantially improved the paper, as well as several seminar and conference audiences for comments and suggestions. The usual disclaimer applies.

Appendix A. Proofs of main results

Notation

Additional notation is introduced that simplifies the discussion of the proofs. The mean for a specific unit i and the mean for a particular t are given by $\overline{Y}_{i,\cdot} = \frac{1}{T}\sum_{t=1}^{T} Y_{i,t}$ and $\overline{Y}_{\cdot,t} = \frac{1}{N}\sum_{i=1}^{N} Y_{i,t}$ respectively. The double-demeaned of $Y_{i,t}$ is denoted as $\ddot{Y}_{i,t} = Y_{i,t} - \overline{Y}_{i,\cdot} - \overline{Y}_{\cdot,t} + \overline{\overline{Y}}$. The mean of $\ddot{Y}_{i,t}$ for a particular group g and specific t is referred to as $\overline{\ddot{Y}}_{g,t} = \frac{1}{N_g}\sum_{i=1}\ddot{Y}_{i,t}\mathbf{1}[E_i = g]$. The difference between means at time t and t' is given by $\Delta\overline{Y}_{g,t,t'} = \overline{Y}_{g,t} - \overline{Y}_{g,t'}$ and $\widehat{DD}_{g,h,t,t'} = \Delta\overline{Y}_{g,t,t'} - \Delta\overline{Y}_{h,t,t'}$. The number of units not included in groups 1 to g is denoted as $[N]_{(g)} = N - \sum_{s=1}^{g} N_s$.

Auxiliary results

The lemmas in this section support the proof of Theorem 1. The proofs of Lemmas 2 and 3 are discussed in an online appendix.

Lemma 2. *We consider three cases.*

First Case. *For $t \geq g+1$ and $d \geq 0$, we have*

$$\overline{\ddot{Y}}_{g,t} + \frac{1}{g}\sum_{r=g+1}^{T}\overline{\ddot{Y}}_{g,r} + \sum_{h=1}^{g+d}\frac{N_h}{[N]_{(g+d)}}\left(\overline{\ddot{Y}}_{h,t} + \frac{1}{g}\sum_{r=g+1}^{T}\overline{\ddot{Y}}_{h,r}\right) = \frac{1}{g}\sum_{r=1}^{g}\left(\Delta\overline{Y}_{g,t,r} - \frac{N_\infty\Delta\overline{Y}_{\infty,t,r}}{[N]_{(g+d)}} - \sum_{h=g+d+1}^{G}\frac{N_h\Delta\overline{Y}_{h,t,r}}{[N]_{(g+d)}}\right). \tag{A.1}$$

Second Case. *For $t \geq g+1$, we have*

$$\frac{[N]_{(g-1)}}{[N]_{(g)}}\left(g\overline{\ddot{Y}}_{g,t} + \sum_{r=g+1}^{T}\overline{\ddot{Y}}_{g,r}\right) + \sum_{h=1}^{g-1}\frac{N_h}{[N]_{(g)}}\left(g\overline{\ddot{Y}}_{h,t} + \sum_{r=g+1}^{T}\overline{\ddot{Y}}_{h,r}\right) = \sum_{r=1}^{g}\left(\Delta\overline{Y}_{g,t,r} - \frac{N_\infty\Delta\overline{Y}_{\infty,t,r}}{[N]_{(g)}} - \sum_{h=g+1}^{G}\frac{N_h\Delta\overline{Y}_{h,t,r}}{[N]_{(g)}}\right). \tag{A.2}$$

Third Case. *For $d \geq 1$, and $t \leq g+d$, we have*

$$g\overline{\ddot{Y}}_{g+d,t} + \sum_{r=g+1}^{T}\overline{\ddot{Y}}_{g+d,r} + \sum_{h=1}^{g+d-1}\frac{N_h\left(g\overline{\ddot{Y}}_{h,t} + \sum_{r=g+1}^{T}\overline{\ddot{Y}}_{h,r}\right)}{[N]_{(g+d-1)}} = \sum_{r=1}^{g}\left(\frac{[N]_{(g+d)}\Delta\overline{Y}_{g+d,t,r} - N_\infty\Delta\overline{Y}_{\infty,t,r} - \sum_{h=g+d+1}^{G}N_h\Delta\overline{Y}_{h,t,r}}{[N]_{(g+d-1)}}\right). \tag{A.3}$$

Lemma 3. *Let*

$$\mathbf{D}'\mathbf{M_Z}\mathbf{D} = \begin{pmatrix} N_1\mathbf{I}_{T-1} - \mathbf{A}_{11} & -\mathbf{A}_{12} & \dots & -\mathbf{A}_{1G} \\ -\mathbf{A}_{21} & N_2\mathbf{I}_{T-2} - \mathbf{A}_{22} & \dots & -\mathbf{A}_{2G} \\ \vdots & \vdots & \vdots & \vdots \\ -\mathbf{A}_{G1} & -\mathbf{A}_{G2} & \dots & N_G - \mathbf{A}_{GG} \end{pmatrix}, \tag{A.4}$$

where $G = T-1$,

$$\mathbf{A}_{ii} = \frac{N_i(N-N_i)}{NT}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}'_{T-i} + \frac{N_i^2}{N}\mathbf{I}_{T-i},$$

$$\textit{for } i<j \quad \mathbf{A}_{ij} = -\frac{N_iN_j}{NT}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}'_{T-j} + \frac{N_iN_j}{N}\begin{pmatrix}\mathbf{0}_{(j-i)\times(T-j)} \\ \mathbf{I}_{T-j}\end{pmatrix},$$

$$\textit{for } i>j \quad \mathbf{A}_{ij} = -\frac{N_iN_j}{NT}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}'_{T-j} + \frac{N_iN_j}{N}\begin{pmatrix}\mathbf{0}_{(T-i)\times(i-j)} & \mathbf{I}_{T-i}\end{pmatrix}.$$

The inverse of the matrix $\mathbf{D}'\mathbf{M_Z}\mathbf{D}$ is given by

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \dots & \mathbf{B}_{1G} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \dots & \mathbf{B}_{2G} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{B}_{G1} & \mathbf{B}_{G2} & \dots & \mathbf{B}_{GG} \end{pmatrix}, \tag{A.5}$$

where

$$\begin{aligned}\mathbf{B}_{ii} =& \frac{[N]_{(i-1)}}{N_i[N]_{(i)}}\mathbf{I}_{T-i} + \frac{[N]_{(i-1)}}{iN_i[N]_{(i)}}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}'_{T-i} + \sum_{r=1}^{T-i-1}\frac{N_{r+i}}{(r+i)[N]_{(i+r-1)}[N]_{(i+r)}}\begin{pmatrix}\mathbf{0}_r\\ \boldsymbol{\iota}_{T-i-r}\end{pmatrix}\begin{pmatrix}\mathbf{0}_r\\ \boldsymbol{\iota}_{T-i-r}\end{pmatrix}' \\ &+ \sum_{r=1}^{T-i-1}\frac{N_{r+i}}{[N]_{(i+r-1)}[N]_{(i+r)}}\begin{pmatrix}\mathbf{0}_{r\times r} & \mathbf{0}_{r\times T-i-r}\\ \mathbf{0}_{T-i-r\times r} & \mathbf{I}_{T-i-r}\end{pmatrix},\end{aligned} \tag{A.6}$$

$$\textit{for } i<j \textit{ and } j<G \qquad \mathbf{B}_{ij} = \begin{pmatrix}\mathbf{0}_{j-i\times T-j}\\ \mathbf{H}_j\end{pmatrix}, \tag{A.7}$$

$$\textit{for } i<j \textit{ and } j=G \qquad \mathbf{B}_{ij} = \begin{pmatrix}\mathbf{0}_{j-i\times T-j}\\ \frac{T}{(T-1)[N]_{(G)}}\end{pmatrix}, \tag{A.8}$$

$$\textit{for } i>j \textit{ and } i<G \qquad \mathbf{B}_{ij} = \begin{pmatrix}\mathbf{0}_{T-i\times i-j} & \mathbf{H}_i\end{pmatrix}, \tag{A.9}$$

$$\textit{for } i>j \textit{ and } i=G \qquad \mathbf{B}_{ij} = \begin{pmatrix}\mathbf{0}_{T-i\times i-j} & \frac{T}{(T-1)[N]_{(G)}}\end{pmatrix}, \tag{A.10}$$

and

$$\begin{aligned}\mathbf{H}_i =& \frac{1}{i[N]_{(i)}}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}'_{T-i} + \frac{1}{[N]_{(i)}}\mathbf{I}_{T-i} + \sum_{r=1}^{T-1-i}\frac{N_{r+i}}{[N]_{(i+r)}[N]_{(i+r-1)}}\begin{pmatrix}\mathbf{0}_{r\times r} & \mathbf{0}_{r\times T-i-r}\\ \mathbf{0}_{T-i-r\times r} & \mathbf{I}_{T-i-r}\end{pmatrix} \\ &+ \sum_{r=1}^{T-1-i}\frac{N_{r+i}}{(i+r)[N]_{(i+r)}[N]_{(i+r-1)}}\begin{pmatrix}\mathbf{0}_r\\ \boldsymbol{\iota}_{T-i-r}\end{pmatrix}\begin{pmatrix}\mathbf{0}_r\\ \boldsymbol{\iota}_{T-i-r}\end{pmatrix}'.\end{aligned} \tag{A.11}$$

Proof of Theorem 1

Assume that the treatment is applied in all periods $t = 2, \dots, T$. For notational simplicity and a cleaner proof, I adopt separate indices for the group g and the time when the group was treated, t_g. The values of the group index g are $[1, 2, 3, \dots, G]$, where the total number of groups is $G = T - 1$ and $t_g = g + 1$.

The equation associated with the extended TWFE estimator is given by:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{g=1}^{G}\sum_{\ell=0}^{T-t_g}\beta_{g,\ell}D_{i,t}^{g,\ell} + e_{i,t}.$$

Let us collect the unit and time-fixed effects in the matrix $\mathbf{Z}$,

$$\mathbf{Z} = \begin{bmatrix}\mathbf{I}_N \otimes \boldsymbol{\iota}_T & \boldsymbol{\iota}_N \otimes \begin{pmatrix}\mathbf{0}\\ \mathbf{I}_{T-1}\end{pmatrix}\end{bmatrix}. \tag{A.12}$$

Additionally, we define a matrix $\mathbf{D}$ that contains all the dummy variables,

$$\begin{aligned}\mathbf{D} &= \begin{bmatrix}\mathbf{D}_1 \otimes \mathbf{f}_{t_1} & \dots & \mathbf{D}_1 \otimes \mathbf{f}_T & \mathbf{D}_2 \otimes \mathbf{f}_{t_2} & \dots & \mathbf{D}_G \otimes \mathbf{f}_T\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{D}_1 \otimes \mathbf{f}_2 & \dots & \mathbf{D}_1 \otimes \mathbf{f}_T & \mathbf{D}_2 \otimes \mathbf{f}_3 & \dots & \mathbf{D}_G \otimes \mathbf{f}_T\end{bmatrix}.\end{aligned} \tag{A.13}$$

where $\mathbf{D}_g$ is an $N \times 1$ vector that indicates group membership. $\mathbf{f}_t$ is a $T \times 1$ vector that contains zeros except in the t-th position.

Using the Frisch–Waugh–Lovell (FWL) theorem, we obtain

$$\hat{\beta} = \left(\mathbf{D}'(\mathbf{I} - \mathbf{P_Z})\mathbf{D}\right)^{-1}\left(\mathbf{D}'(\mathbf{I} - \mathbf{P_Z})\mathbf{Y}\right), \tag{A.14}$$

where $\mathbf{P_Z} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$, $\hat{\beta} = \begin{bmatrix}\hat{\beta}'_2 & \hat{\beta}'_3 & \dots & \hat{\beta}'_G\end{bmatrix}'$, and $\hat{\beta}_g = \begin{bmatrix}\hat{\beta}_{g,0} & \hat{\beta}_{g,1} & \dots & \hat{\beta}_{g,T-t_g}\end{bmatrix}'$.

First, we focus on $\left(\mathbf{D}'(\mathbf{I} - \mathbf{P_Z})\mathbf{D}\right)^{-1}$. After some matrix algebra, we obtain

$$\mathbf{D}'\mathbf{M_Z}\mathbf{D} = \begin{pmatrix}N_1\mathbf{I}_{T-1} - \mathbf{A}_{11} & -\mathbf{A}_{12} & \dots & -\mathbf{A}_{1G}\\ -\mathbf{A}_{21} & N_2\mathbf{I}_{T-2} - \mathbf{A}_{22} & \dots & -\mathbf{A}_{2G}\\ \vdots & \vdots & \vdots & \vdots\\ -\mathbf{A}_{G1} & -\mathbf{A}_{G2} & \dots & N_G - \mathbf{A}_{GG}\end{pmatrix}, \tag{A.15}$$

where

$$\mathbf{A}_{ii} = \frac{N_i(N-N_i)}{NT}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}'_{T-i} + \frac{N_i^2}{N}\mathbf{I}_{T-i},$$

for $i < j$

$$\mathbf{A}_{ij} = -\frac{N_iN_j}{NT}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}'_{T-j} + \frac{N_iN_j}{N}\begin{pmatrix}\mathbf{0}_{(j-i)\times(T-j)}\\ \mathbf{I}_{T-j}\end{pmatrix},$$

for $i > j$

$$\mathbf{A}_{ij} = -\frac{N_i N_j}{NT}\boldsymbol{\iota}_{T-i}\boldsymbol{\iota}_{T-j}' + \frac{N_i N_j}{N}\begin{pmatrix}\mathbf{0}_{(T-i)\times(i-j)} & \mathbf{I}_{T-i}\end{pmatrix}.$$

The inverse of $\mathbf{D}'\mathbf{M_Z}\mathbf{D}$ is given in Lemma 3. The inverse is equal to

$$\left(\mathbf{D}'\mathbf{M_Z}\mathbf{D}\right)^{-1} = \mathbf{B} = \begin{pmatrix}\mathbf{B}_{11} & \mathbf{B}_{12} & \dots & \mathbf{B}_{1G}\\ \mathbf{B}_{21} & \mathbf{B}_{22} & \dots & \mathbf{B}_{2G}\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{B}_{G1} & \mathbf{B}_{G2} & \dots & \mathbf{B}_{GG}\end{pmatrix}. \tag{A.16}$$

The matrices $\mathbf{B}_{ij}$ are described in (A.6)–(A.10). From (A.14) and (A.16), we have

$$\hat{\boldsymbol{\beta}} = \mathbf{B}\left(\mathbf{D}'(\mathbf{I}-\mathbf{P_Z})\mathbf{Y}\right) = \mathbf{B}\mathbf{D}'\ddot{\mathbf{Y}}, \tag{A.17}$$

where the elements of $\ddot{\mathbf{Y}}$ are equal to $\ddot{Y}_{i,t}$. We have

$$\mathbf{D}'\ddot{\mathbf{Y}} = \left[N_1\overline{\ddot{\mathbf{Y}}}_1' \quad N_2\overline{\ddot{\mathbf{Y}}}_2' \quad \dots \quad N_G\overline{\ddot{\mathbf{Y}}}_G'\right]', \tag{A.18}$$

where $\overline{\ddot{\mathbf{Y}}}_g = \left[\overline{\ddot{Y}}_{g,g+1} \quad \overline{\ddot{Y}}_{g,g+2} \quad \dots \quad \overline{\ddot{Y}}_{g,T}\right]'$. For a given group g, the estimated coefficients $\hat{\beta}_g$ can be obtained from (A.16)–(A.18),

$$\hat{\boldsymbol{\beta}}_g = \sum_{h=1}^{g-1}\mathbf{B}_{gh}N_h\overline{\ddot{\mathbf{Y}}}_h + \mathbf{B}_{gg}N_g\overline{\ddot{\mathbf{Y}}}_g + \sum_{h=g+1}^{G}\mathbf{B}_{gh}N_h\overline{\ddot{\mathbf{Y}}}_h. \tag{A.19}$$

Using Eq. (A.6) for $\mathbf{B}_{gg}$ and Eqs. (A.7)–(A.10) to obtain the different values of $\mathbf{B}_{gh}$, we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_g =& \frac{[N]_{(g-1)}}{[N]_{(g)}}\left(\overline{\ddot{\mathbf{Y}}}_g + \frac{\sum_{t=g+1}^T\overline{\ddot{Y}}_{g,t}}{g}\boldsymbol{\iota}_{T-g}\right) + \sum_{h=1}^{g-1}\frac{N_h}{[N]_{(g)}}\begin{pmatrix}\overline{\ddot{Y}}_{h,g+1} + \frac{\sum_{r=1}^{T-g}\overline{\ddot{Y}}_{h,g+r}}{g}\\ \vdots\\ \overline{\ddot{Y}}_{h,T} + \frac{\sum_{r=1}^{T-g}\overline{\ddot{Y}}_{h,g+r}}{g}\end{pmatrix} + \sum_{h=g+1}^{G}\frac{N_h}{[N]_{(h)}}\begin{pmatrix}\mathbf{0}_{h-g}\\ \overline{\ddot{\mathbf{Y}}}_h + \frac{\sum_{r=1}^{T-h}\overline{\ddot{Y}}_{h,h+r}}{h}\boldsymbol{\iota}_{T-h}\end{pmatrix}\\ &+ \sum_{h=1}^{g}\sum_{r=1}^{T-1-g}\frac{N_hN_{r+g}}{[N]_{(g+r)}[N]_{(g+r-1)}}\begin{pmatrix}\mathbf{0}_r\\ \overline{\ddot{Y}}_{h,g+1+r} + \frac{\sum_{t=g+1+r}^T\overline{\ddot{Y}}_{h,t}}{g+r}\\ \vdots\\ \overline{\ddot{Y}}_{h,T} + \frac{\sum_{t=g+1+r}^T\overline{\ddot{Y}}_{h,t}}{g+r}\end{pmatrix} + \sum_{h=g+1}^{G-1}\sum_{r=1}^{T-1-h}\frac{N_{r+h}N_h}{[N]_{(h+r)}[N]_{(h+r-1)}}\begin{pmatrix}\mathbf{0}_{h-g+r}\\ \overline{\ddot{Y}}_{h,h+1+r} + \frac{\sum_{t=h+1+r}^T\overline{\ddot{Y}}_{h,t}}{h+r}\\ \vdots\\ \overline{\ddot{Y}}_{h,T} + \frac{\sum_{t=h+1+r}^T\overline{\ddot{Y}}_{h,t}}{h+r}\end{pmatrix}.\end{aligned} \tag{A.20}$$

In the last part of the proof, we inspect the elements of $\hat{\beta}_g$. First, we focus on $\hat{\beta}_{g,0}$, and after we analyze $\beta_{g,\ell}$ for $T - t_g \geq \ell > 1$. Using (A.20), we can find the first element of $\hat{\boldsymbol{\beta}}_g$, we have

$$\begin{aligned}\hat{\beta}_{g,0} =& \frac{1}{[N]_{(g)}}\left([N]_{(g-1)}\overline{\ddot{Y}}_{g,g+1} + \sum_{h=1}^{g-1}N_h\overline{\ddot{Y}}_{h,g+1}\right) + \frac{1}{g[N]_{(g)}}\sum_{t=g+1}^{T}\left([N]_{(g-1)}\overline{\ddot{Y}}_{g,t} + \sum_{h=1}^{g-1}N_h\overline{\ddot{Y}}_{h,t}\right)\\ =& \frac{1}{g}\sum_{t=1}^{g}\left(\Delta\overline{Y}_{g,g+1,t} - \frac{N_\infty}{[N]_{(g)}}\Delta\overline{Y}_{\infty,g+1,t} - \sum_{h=g+1}^{G}\frac{N_h}{[N]_{(g)}}\Delta\overline{Y}_{h,g+1,t}\right).\end{aligned} \tag{A.21}$$

For the second equality, we use the second case of Lemma 2. Given that $t_g = g + 1$ and $[N]_{(g)} = N_\infty + \sum_{h=g+1}^G N_h = \sum_{h>g} N_h$, we have

$$\hat{\beta}_{g,0} = \sum_{h>g}\frac{N_h}{\sum_{s>g}N_s}\left(\frac{1}{t_g - 1}\sum_{t=1}^{t_g-1}\widehat{DD}_{g,h,t_g,t}\right). \tag{A.22}$$

For $\hat{\beta}_{g,\ell}$ where $T - t_g \geq \ell > 1$, we have

$$\begin{aligned}\hat{\beta}_{g,\ell} =& \frac{[N]_{(g-1)}}{[N]_{(g)}}\left(\overline{\ddot{Y}}_{g,g+1+\ell} + \frac{\sum_{t=g+1}^T\overline{\ddot{Y}}_{g,t}}{g}\right) + \sum_{h=1}^{\ell}\frac{N_{g+h}}{(g+h)[N]_{(g+h)}}\left((g+h)\overline{\ddot{Y}}_{g+h,g+1+\ell} + \sum_{t=g+h+1}^{T}\overline{\ddot{Y}}_{g+h,t}\right)\\ &+ \sum_{h=1}^{g-1}\frac{N_h}{[N]_{(g)}}\overline{\ddot{Y}}_{h,g+1+\ell} + \sum_{r=1}^{\ell}\frac{N_{r+g}N_g\left((g+r)\overline{\ddot{Y}}_{g,g+1+\ell} + \sum_{t=g+1+r}^T\overline{\ddot{Y}}_{g,t}\right)}{(g+r)[N]_{(g+r)}[N]_{(g+r-1)}}\\ &+ \sum_{h=1}^{g-1}\sum_{r=1}^{\ell}\frac{N_{r+g}N_h\left((g+r)\overline{\ddot{Y}}_{h,g+1+\ell} + \sum_{t=g+1+r}^T\overline{\ddot{Y}}_{h,t}\right)}{(g+r)[N]_{(g+r)}[N]_{(g+r-1)}} + \sum_{h=1}^{\ell-1}\sum_{r=1}^{\ell-h}\frac{N_{g+h+r}N_{g+h}\overline{\ddot{Y}}_{g+h,g+1+\ell}}{[N]_{(g+h+r)}[N]_{(g+h+r-1)}}\\ &+ \sum_{h=1}^{g-1}\frac{N_h}{g[N]_{(g)}}\sum_{t=g+1}^{T}\overline{\ddot{Y}}_{h,t} + \sum_{h=1}^{\ell-1}\sum_{r=1}^{\ell-h}\frac{N_{g+h+r}N_{g+h}}{(g+h+r)[N]_{(g+h+r)}[N]_{(g+h+r-1)}}\sum_{t=g+h+1+r}^{T}\overline{\ddot{Y}}_{g+h,t}.\end{aligned} \tag{A.23}$$

The above expression can be rewritten as

$$\begin{aligned}\hat{\beta}_{g,\ell} =&\frac{[N]_{(g+\ell)}+N_g}{[N]_{(g+\ell)}}\bar{\ddot{Y}}_{g,g+1+\ell}+\frac{[N]_{(g+\ell)}+N_g}{g[N]_{(g+\ell)}}\sum_{t=g+1}^{T}\bar{\ddot{Y}}_{g,t}-\sum_{r=1}^{\ell}\frac{N_{g+r}N_g}{(g+r)[N]_{(g+r)}[N]_{(g+r-1)}}\left(\frac{r}{g}\sum_{t=g+1}^{T}\bar{\ddot{Y}}_{g,t}+\sum_{t=g+1}^{g+r}\bar{\ddot{Y}}_{g,t}\right)\\
&+\frac{1}{[N]_{(g+\ell)}}\sum_{h=1}^{g-1}N_h\bar{\ddot{Y}}_{h,g+1+\ell}+\frac{1}{g[N]_{(g+\ell)}}\sum_{h=1}^{g-1}\sum_{t=g+1}^{T}N_h\bar{\ddot{Y}}_{h,t}-\sum_{r=1}^{\ell}\frac{rN_{g+r}}{g(g+r)[N]_{(g+r)}[N]_{(g+r-1)}}\sum_{h=1}^{g-1}\sum_{t=g+1}^{T}N_h\bar{\ddot{Y}}_{h,t}\\
&-\sum_{r=1}^{\ell}\frac{N_{g+r}}{(g+r)[N]_{(g+r)}[N]_{(g+r-1)}}\sum_{h=1}^{g-1}\sum_{t=g+1}^{g+r}N_h\bar{\ddot{Y}}_{h,t}+\sum_{h=1}^{\ell}\frac{N_{g+h}}{[N]_{(g+\ell)}}\bar{\ddot{Y}}_{g+h,g+1+\ell}\\
&+\sum_{h=1}^{\ell}\frac{N_{g+h}}{(g+h)[N]_{(g+\ell)}}\sum_{t=g+h+1}^{T}\bar{\ddot{Y}}_{g+h,t}-\sum_{h=1}^{\ell-1}\sum_{r=1}^{\ell-h}\frac{rN_{r+g+h}N_{g+h}\sum_{t=g+h+1}^{T}\bar{\ddot{Y}}_{g+h,t}}{(g+h)(g+h+r)[N]_{(g+h+r)}[N]_{(g+h+r-1)}}\\
&-\sum_{h=1}^{\ell-1}\sum_{r=1}^{\ell-h}\frac{N_{r+g+h}N_{g+h}}{(g+h+r)[N]_{(g+h+r)}[N]_{(g+h+r-1)}}\sum_{t=g+h+1}^{g+h+r}\bar{\ddot{Y}}_{g+h,t}.\end{aligned} \tag{A.24}$$

After some algebraic manipulations, we have

$$\begin{aligned}\hat{\beta}_{g,\ell} =&\bar{\ddot{Y}}_{g,g+1+\ell}+\frac{1}{g}\sum_{t=g+1}^{T}\bar{\ddot{Y}}_{g,t}+\sum_{h=1}^{g+\ell}\frac{N_h\left(g\bar{\ddot{Y}}_{h,g+1+\ell}+\sum_{t=g+1}^{T}\bar{\ddot{Y}}_{h,t}\right)}{g[N]_{(g+\ell)}}-\sum_{r=1}^{\ell}\sum_{h=1}^{g}\frac{rN_{g+r}N_h}{g(g+r)[N]_{(g+r)}[N]_{(g+r-1)}}\sum_{t=g+1}^{T}\bar{\ddot{Y}}_{h,t}\\
&-\sum_{r=1}^{\ell}\sum_{h=1}^{g}\frac{N_{g+r}N_h}{(g+r)[N]_{(g+r)}[N]_{(g+r-1)}}\sum_{t=g+1}^{g+r}\bar{\ddot{Y}}_{h,t}-\sum_{h=1}^{\ell}\frac{N_{g+h}}{g[N]_{(g+\ell)}}\left(\sum_{t=g+1}^{g+h}\bar{\ddot{Y}}_{g+h,t}+\frac{h}{g+h}\sum_{t=g+h+1}^{T}\bar{\ddot{Y}}_{g+h,t}\right)\\
&-\sum_{h=1}^{\ell-1}\sum_{r=1}^{\ell-h}\frac{rN_{r+g+h}N_{g+h}\sum_{t=g+h+1}^{T}\bar{\ddot{Y}}_{g+h,t}}{(g+h)(g+h+r)[N]_{(g+h+r)}[N]_{(g+h+r-1)}}-\sum_{h=1}^{\ell-1}\sum_{r=1}^{\ell-h}\frac{N_{r+g+h}N_{g+h}\sum_{t=g+h+1}^{g+h+r}\bar{\ddot{Y}}_{g+h,t}}{(g+h+r)[N]_{(g+h+r)}[N]_{(g+h+r-1)}}\\
=&\frac{1}{g}\sum_{r=1}^{g}\left(\Delta\bar{Y}_{g,g+1+\ell,r}-\frac{N_\infty}{[N]_{(g+\ell)}}\Delta\bar{Y}_{\infty,g+1+\ell,r}-\sum_{h=g+1+\ell}^{G}\frac{N_h\Delta\bar{Y}_{h,g+1+\ell,r}}{[N]_{(g+\ell)}}\right)\\
&-\sum_{h=1}^{\ell}\sum_{t=g+1}^{g+h}\frac{N_{g+h}}{(g+h)[N]_{(g+h)}}\left[\bar{\ddot{Y}}_{g+h,t}+\frac{1}{g}\sum_{r=g+1}^{T}\bar{\ddot{Y}}_{g+h,r}+\sum_{r=1}^{g+h-1}\frac{N_r}{[N]_{(g+h-1)}}\left(\bar{\ddot{Y}}_{r,t}+\frac{\sum_{t=g+1}^{T}\bar{\ddot{Y}}_{r,t}}{g}\right)\right].\end{aligned} \tag{A.25}$$

For the second equality, we use the first case of Lemma 2. We apply the third case of Lemma 2, and we use $t_g = g+1$, we have

$$\hat{\beta}_{g,\ell} = \sum_{h>g+\ell}\frac{N_h}{\sum_{s>g+\ell}N_s}\left(\frac{1}{t_g-1}\sum_{t=1}^{t_g-1}\widehat{DD}_{g,h,t_g+\ell,t}\right)-\sum_{h>g}^{g+\ell}\sum_{t=t_g}^{t_h-1}\frac{N_{g+h}}{(t_{g+h}-1)[N]_{(g+h-1)}}\left[\sum_{k>h}\frac{N_k\sum_{r=1}^{t_g-1}\widehat{DD}_{h,k,t,r}}{\left(N_u+\sum_{s>h}N_s\right)(t_g-1)}\right]. \tag{A.26}$$

Using Assumption 2, the sample means converge to population means, which gives the result $\widehat{DD}_{t^*,t,g,h}\xrightarrow{p}DD_{t^*,t,g,h}$, and

$$\begin{aligned}\frac{N_h}{\sum_{h>g+\ell}N_h}&\xrightarrow{p}\frac{Pr(E_i=t_h)}{Pr(E_i>t_g+\ell)}=Pr(E_i=t_h|E_i>t_g+\ell)=w_{h,g+\ell},\\
\frac{N_h}{(t_h-1)\sum_{s\geq h}N_h}&\xrightarrow{p}\frac{Pr(E_i=t_h)}{(t_h-1)Pr(E_i\geq t_h)}=\frac{Pr(E_i=t_h|E_i\geq t_h)}{t_h-1}=\tilde{w}_h.\end{aligned}$$

We obtain

$$\beta_{g,\ell} = \sum_{h>g+\ell}\frac{w_{h,g+\ell}}{t_g-1}\sum_{t=1}^{t_g-1}DD_{g,h,g+\ell,t}-\sum_{h>g}^{g+\ell}\sum_{t=t_g}^{t_h-1}\tilde{w}_h\left[\sum_{k>h}\frac{w_{k,h}}{t_g-1}\sum_{r=1}^{t_g-1}DD_{h,k,t,r}\right]. \tag{A.27}$$

This shows (7) and ends the proof of Theorem 1. ■

Proof of Corollary 1

First we analyze $\psi_{g,\ell}$, for $\ell>0$, we have

$$\psi_{g,\ell} = \sum_{h>g}^{g+\ell}\sum_{t=g}^{h-1}\frac{\tilde{w}_h}{g-1}\sum_{r=1}^{g-1}\left(\sum_{k>h}w_{k,h}DD_{h,k,t,r}\right). \tag{A.28}$$

Then given Assumption 3, we have

$$\psi_{g,\ell} = \sum_{h>g}^{g+\ell}\sum_{t=g}^{h-1}\tilde{w}_h\left[\sum_{k>h}\frac{w_{k,h}}{g-1}\sum_{r=1}^{g-1}\left(\Delta Y_{t,r}^{(h)}-\Delta Y_{t,r}^{(k)}\right)\right]=\sum_{h>g}^{g+\ell}\sum_{t=g}^{h-1}\tilde{w}_hWCTD(h,0,t,g-1), \tag{A.29}$$

where $\Delta Y^{(g)}_{t^*,t} = \mathbb{E}\left[Y_{i,t^*}(\infty) - Y_{i,t}(\infty)|E_i = g\right]$ and

$$WCTD(g, j, t^*, s) = \sum_{h>g+j} \frac{w_{h,g+j}}{s} \sum_{t=1}^{s} \left(\Delta Y^{(g)}_{t^*,t} - \Delta Y^{(h)}_{t^*,t}\right). \tag{A.30}$$

From Theorem 1, we have

$$\beta_{g,\ell} = \sum_{h>g+\ell} \frac{w_{h,g+\ell}}{g-1} \sum_{t=1}^{g-1} \left(\mathbb{E}\left[Y_{i,g+\ell}(g) - Y_{i,t}(g)|E_i = g\right] - \mathbb{E}\left[Y_{i,g+\ell}(h) - Y_{i,t}(h)|E_i = h\right]\right) - \psi_{g,\ell}. \tag{A.31}$$

Then, given Assumption 3, we have

$$\begin{aligned}\beta_{g,\ell} =& ATT(g,\ell) + \sum_{h>g+\ell} \frac{w_{h,g+\ell}}{g-1} \sum_{t=1}^{g-1} \left[\Delta Y^{(g)}_{g+\ell,t} - \Delta Y^{(h)}_{g+\ell,t}\right] - \psi_{g,\ell} \\ =& ATT(g,\ell) + WCTD(g,\ell,g+\ell,g-1) - \sum_{h>g}^{g+\ell} \sum_{t=g}^{h-1} \tilde{w}_h WCTD(h,0,t,g-1).\end{aligned} \tag{A.32}$$

This shows (12) and ends the proof. ■

Proof of Lemma 1

The proof of Lemma 1 is straightforward. It relies on the fact that the equation associated with the SA-TWFE estimator corresponds to a saturated TWFE specification. In particular, we have the following equation:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{g=2}^{T} \sum_{\ell \neq -1} \delta_{g,\ell} D^{g,\ell}_{i,t} + e_{i,t}.$$

We take conditionals probabilities

$$\begin{aligned}\mathbb{E}[Y_{i,t}|E_i = \infty, t = g-1] &= \alpha_\infty + \lambda_{g-1}, \\ \mathbb{E}[Y_{i,t}|E_i = \infty, t = g+\ell] &= \alpha_\infty + \lambda_{g+\ell}, \\ \mathbb{E}[Y_{i,t}|E_i = g, t = g-1] &= \alpha_g + \lambda_{g-1}, \\ \mathbb{E}[Y_{i,t}|E_i = g, t = g+\ell] &= \alpha_g + \lambda_{g+\ell} + \delta_{g,\ell}.\end{aligned}$$

From the previous equations, we obtain

$$\delta_{g,\ell} = \mathbb{E}\left[Y_{i,g+\ell}|E_i = g\right] - \mathbb{E}[Y_{i,g-1}|E_i = g] - \left(\mathbb{E}[Y_{i,g+\ell}|E_i = \infty] - \mathbb{E}[Y_{i,g-1}|E_i = \infty]\right).$$

This shows (10) and ends the proof. ■

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2025.106059.

References

Arkhangelsky, D., Samkov, A., 2024. Sequential synthetic difference in differences. arXiv preprint arXiv:2404.00164.
Borusyak, K., Jaravel, X., 2017. Revisiting event study designs. Working Paper.
Borusyak, K., Jaravel, X., Spiess, J., 2024. Revisiting event-study designs: robust and efficient estimation. Rev. Econ. Stud. 91 (6), 3253–3285.
Callaway, B., Sant'Anna, P.H., 2021. Difference-in-differences with multiple time periods. J. Econometrics 225 (2), 200–230.
Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: Methods and applications. Cambridge University Press.
De Chaisemartin, C., d'Haultfoeuille, X., 2023. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. Econom. J. 26 (3), C1–C30.
Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. J. Econometrics 225 (2), 254–277.
Harmon, N.A., 2022. Difference-in-differences and efficient estimation of treatment effects. Work. Pap..
Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.
Marcus, M., Sant'Anna, P.H., 2021. The role of parallel trends in event study settings: An application to environmental economics. J. Assoc. Environ. Resour. Econ. 8 (2), 235–275.
Rambachan, A., Roth, J., 2023. A more credible approach to parallel trends. Rev. Econ. Stud. 90 (5), 2555–2591.
Roth, J., 2022. Pretest with caution: Event-study estimates after testing for parallel trends. Am. Econ. Rev.: Insights 4 (3), 305–322.
Roth, J., Sant'Anna, P.H., Bilinski, A., Poe, J., 2023. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. J. Econometrics 235 (2), 2218–2244.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66 (5), 688.
Sun, L., Abraham, S., 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J. Econometrics 225 (2), 175–199.
Wooldridge, J., 2021. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. Available At SSRN 3906345.