Online Appendix for:

'Difference-in-Differences and Efficient Estimation of Treatment Effects'

Content:

- Section A generalizes the theoretical framework from the main paper to allow for unbalanced panels.
- Section B states and proves a general version of the main paper's Theorem 1 allowing for unbalanced panels.
- Section C discusses various cases where SGDD is equivalent to SWDD and is thus efficient (including the 'never-treated' approach of Callaway and Sant'Anna (2021) and Sun and Abraham (2021))
- Section D goes through a simple example illustrating how SWDD provides efficiency gains over SGDD.
- Section E discusses extensions of the SWDD estimator to condition on predetermined covariates and to examine pretrends.
- Section F provides details and additional results from the Brenøe *et al.* (2024) data.
- Section G shows that SGDD can be viewed as an efficient RI estimator under random walk errors. In particular, SGDD is the best unbiased estimator of treatment effects at horizon h under the weaker assumption that parallel trends hold *only* at this horizon.

A Framework and assumptions allowing for unbalanced panels

Below I present a generalized version of the framework and assumptions from the main paper which allows for unbalanced panels. To make the presentation self-contained, I restate notation and assumptions when applicable.

A.1 Basic setup and notation

The data set contain a number of units, indexed by i, observed over several periods, indexed by t. We are interested in the causal effect of a particular treatment on some outcome. $Y_{i,t}$ is the outcome for unit i in period t, while $D_{i,t}$ is an indicator for whether i is treated in period t. Different from the main text, the data may be arbitrarily unbalanced meaning that some units may be unobserved in some periods. As a convention, I let $t = 1$ denote the first period where data is available on some unit and $T \geq 2$ denote the last period where data is available on some unit.

Treatment is an absorbing state so for each unit there is some period E_i when treatment occurs. In any period $t < E_i$ where unit i is observed we thus have $D_{i,t} = 0$, while in any period $t \geq E_i$ where i is observed we have $D_{i,t} = 1$. $E_i = \infty$ corresponds to units that are never treated. The analysis will treat the observed data as containing a fixed set of observations $(i,t) \in \Omega$, with the treatment timing being non-stochastic (meaning that $D_{i,t}$ and E_i are non-stochastic). I let $\Omega^N = \{i : (i,t) \in \Omega \text{ for some } t\}$ denote the set of observed units and use N to refer to the total number of units ever observed.

For expositional convenience, I also define some additional variables and notation. I let $K_{i,t} = t - E_i$ denote the number of periods since unit i was treated. I let $\bar{K}_i = \max_{t:(i,t)\in\Omega} K_{i,t}$ denote the maximum number of additional post-treatment periods for which it is possible to observe i and $\bar{K} = \max_{(i,t)\in\Omega} K_{i,t}$ be the maximum number of such post-treatment periods observed for any unit in the data. Finally, for $k = 0, 1, ..., \bar{K}$, I define $H^k_{i,t}$ as a dummy variable for whether at time t, the unit i is treated and has experienced the treatment for exactly k previous periods:

$$H^k_{i,t} = \begin{cases} 1 & \text{if} \quad K_{i,t} = k \\ 0 & \text{otherwise} \end{cases}$$

I also adopt some simplifying notation regarding averages over units satisfying various conditions.

For a statement $\mathcal{A}_i$ that depends on i, I let $\frac{\mathbf{1}}{\mathbf{N}}\sum_{i:\mathcal{A}_i}$ denote the average over those units i for which $\mathcal{A}_i$ evaluates as true.[13] Because I allow for some forms of missing data and unbalanced panels, I will here adopt the convention that a statement $\mathcal{A}_i$ always evaluate to false if it involves an expression that is undefined because of missing data. As an example, the expression below corresponds to the the average period t outcome for units who are untreated at both time t and time $t+k$ *and are observed* in the data at both times:

$$\frac{\mathbf{1}}{\mathbf{N}}\sum_{\substack{i:\ D_{i,t}=0,\\ D_{i,t+k}=0}} Y_{i,t}$$

A.2 Potential outcomes, treatment effects and estimands

Let $Y_{i,t}^0$ denote the (unobserved) potential outcome for unit i in period t in a situation where i never receives the treatment. Estimands of interests will build on individual treatment effects measured at different time horizons relative to the onset of treatment. I let $\gamma_{i,h} = E\left[Y_{i,E_i+h} - Y_{i,E_i+h}^0\right]$ denote unit i's treatment effect, at the time when they have experienced the treatment for h previous periods. I will refer to this as the horizon h treatment effect for i.

One general estimand of interest is then a weighted sum of treatment effects at a specific horizon h: $\gamma_h^w = \sum_{i:\bar{K}_i \geq h} w_i \gamma_{i,h}$ for some set of weight $\{w_i\}_{i:\bar{K}_i \geq h}$ that may depend on observed treatment timing. For the main theoretical result, I consider a more general estimand which possibly also sums over effects at different horizons: $\gamma^w = \sum_{i,h:\bar{K}_i \geq h} w_{i,h}\gamma_{i,h}$ for some set of weights $\{w_{i,h}\}_{i,h:\bar{K}_i \geq h}$.

A.3 Subgroup Difference-in-Differences estimators

Next, I define Subgroup Difference-in-Differences (SGDD) estimators for the estimands, γ_h^w and γ^w, in the case where the data may be an unbalanced panel:

Definition. *The Subgroup Difference-in-Differences estimators for the weighted sum of horizon h treatment effects, γ_h^w, and the weighted sum of arbitrary treatment effects, γ^w, are defined as*

$$\widehat{\gamma_h^w}^{SGDD} = \sum_{i:\bar{K}_i \geq h} w_i \hat{\gamma}_{i,h}^{SGDD} \tag{8}$$

[13] The notation $\frac{\mathbf{1}}{\mathbf{N}}\sum_{i:\mathcal{A}_i}$ is thus equivalent to the longer notation $\frac{1}{\#\{i\in\Omega^n:\mathcal{A}_i \text{ is true}\}}\sum_{i\in\Omega^N:\mathcal{A}_i \text{ is true}}$.

$$\widehat{\gamma^w}^{SGDD} = \sum_{i,h:\bar{K}_i \geq h} w_{i,h} \hat{\gamma}^{SGDD}_{i,h} \tag{9}$$

where $\left\{\hat{\gamma}^{SGDD}_{i,h}\right\}_{i,h:\bar{K}_i \geq h}$ *are individual-level treatment effect estimators, defined as*

$$\hat{\gamma}^{SGDD}_{i,h} = (Y_{i,E_i+h} - Y_{i,E_i-1}) - \sum_{\substack{j:\, D_{j,E_i}=0,\\ D_{j,E_i+h}=0}} \frac{\mathbf{1}}{\mathbf{N}} (Y_{j,E_i+h} - Y_{j,E_i-1}) \tag{10}$$

Relative to the defintion of SGDD in the main text, the difference here is only that the last sum in (10) now requires the relevant controls to be *observed* as untreated in *both* period E_i and $E_i + h$.

A.4 Restrictions on observed and missing data

The setup so far imposes no restrictions on what types of units exist and no restriction on the time periods in which the different units are observed. Some restrictions are required so that treatment effects are identified and that the SGDD estimators are well defined.

First, throughout the analysis, I will assume that the SGDD estimators under study are well-defined. The following assumption is necessary and sufficient for this to hold:

Assumption 6. *SGDD Estimator is Defined for Horizon h: If* $K_{it} = h$ *for some* $(i,t) \in \Omega$ *then* $(i, E_i - 1) \in \Omega$ *and there exists some other unit* j *with* $(j,t), (j, E_i - 1) \in \Omega$ *and* $D_{j,t} = 0$.

If this assumption fails, then there is some treated unit for which data is missing on the period just before treatment onset or where no untreated control units are observed in the corresponding time periods. Either of these possibilites makes it impossible to form an individual-level SGDD estimator for this unit. Conditional on considering SGDD estimators, the assumption is thus innocuous; any treated units for which the assumption fails would mechanically have to be excluded from an SGDD analysis. At the same time, I note that there are cases where the assumption above fails but where treatment effects are in fact identified. This highlights that standard difference-in-difference estimators may fail to be defined even when all treatment effects are identified (see Bellégo *et al.* (2024) for a discussion of estimation in such settings).

For the main theorem, I will focus on the case where the SGDD estimator is defined for all the treatment horizons considered in the data:

Assumption 7. *SGDD Estimator is Defined for the Relevant Horizons: For all $h = 1, 2, ..., \bar{K}$, the SGDD Estimator is Defined for Horizon h.*

Additionally, I will maintain an additional restriction that units in the data do not drop in and out of the sample but are observed continously for some number of periods. I refer to this as assuming no holes in the data:

Assumption 8. *No Holes in the Data: For each unit i there exists a first and last observed period, $\underline{T_i}, \overline{T_i}$ such that $(i,t) \in \Omega$ if and only if $t \in \{\underline{T_i}, \underline{T_i} + 1..., \overline{T_i}\}$.*

This is substantially weaker than assuming a balanced panel: each unit may be missing for an arbitrary number of periods both at the beginning and end of the sample period. While the assumptions is also likely to be satisfied in most applications, I note that it does play an important role for the efficiency results presented later. With arbitrary patterns of missing data, even the Stepwise Difference-in-Differences estimator presented later may fail to leverage all relevant information.

The assumptions above will be invoked for the main theorem. For some results presented later in this appendix, I will further strengthen the assumptions however. For some results, I will require the data to be a balanced panel:

Assumption 9. *Balanced Panel: For each unit i and each $t \in \{1, 2..., \overline{T}\}$, we have $(i,t) \in \Omega$.*

Additionally, for some results I will require non-staggered adoption:

Assumption 10. *Non-Staggered Adoption: For any pair of units (i,j) such that $E_i, E_j \neq \infty$, we have $E_i = E_j$.*

A.5 Identifying assumptions

Identification will rest on two standard assumptions throughout. The first is a no anticipation assumption:

Assumption 11. *No Anticipation: $Y_{i,t} = Y_{i,t}^0$ whenever $D_{i,t} = 0$.*

The second assumption will be a parallel trends assumption, imposing that in the absence of treatment, outcomes move in parallel across all observations included in the data:

Assumption 12. *Parallel Trends: For any two periods t and t', $E\left[Y_{i,t}^0 - Y_{i,t'}^0\right]$ is constant across all units i that are observed at both t and t'.*

As discussed in the main text, because the framework here allows for quite general forms of unbalanced panels, the parallel trends assumption above (and the corresponding SGDD) estimator can be made equivalent to various different approaches considered in the literature simply by appropriately restricting the data: If the data is restricted to only include observations from one period before the first unit experiences treatment (e.g. $\min_{i,t\in\Omega} E_i = 2$) both the parallel trends assumptions and the SGDD estimators defined above become equivalent to the 'not-yet-treated' approach of Callaway and Sant'Anna (2021). If the data is restricted to only include data on never-treated units and on treated units only starting from period before they receive treatment (e.g. $E_i = \infty$ or $\underline{T_i} = E_i - 1$ for all i, where $\underline{T_i}$ is the first period in which i is observed), the defined parallel trends assumptions and the SGDD estimators become equivalent to the 'never-treated' approach of Sun and Abraham (2021) and Callaway and Sant'Anna (2021).

A.6 Persistent error benchmark

Define $\varepsilon_{i,t} = Y_{i,t} - E[Y_{i,t}]$ to be the error for unit i in time period t. The efficiency results in this paper pertains to the case where these errors follow a random walk. To cover the case where data is missing for some periods, I formulate this assumption explicitly in terms of shocks: For all units i and *all* time periods t, I let $\eta_{i,t}$ denote the corresponding shock to the outcome variable. Letting η be the NT-dimensional vector of all shocks I then consider the following assumption:

Assumption 13. *Random Walk Errors: For any $k \geq 1$, the errors satisfy*

$$\varepsilon_{i,t+k} - \varepsilon_{i,t} = \sum_{t'=t+1}^{t+k} \eta_{i,t'} \tag{11}$$

whenever unit i is observed at both t and $t+k$.

The shocks η are mean zero, homoskedastic and uncorrelated over time and units: $E(\eta) = 0$, $Var(\eta) = \mathbb{I}_{NT}\sigma^2$.

B Statement and proof of Theorem 1 in the general case of unbalanced panels1 and 1

In the context of the framework from Section A, I now restate and prove the general version of Theorem 1 allowing for unbalanced panels. To avoid confusion, I refer to this general version as Theorem 2:

Theorem 2. *Assume that the SGDD Estimator is Defined for the Relevant Horizons, that there are No Holes in the Data, there is No Anticipation, there is Parallel Trends and there are Random Walk Errors. Then the best unbiased estimator of any treatment effect γ^w is the Stepwise Difference-in-Differences estimator, $\widehat{\gamma^w}^{SWDD}$, which is defined as:*

$$\widehat{\gamma^w}^{SWDD} = \sum_{i,h:\bar{K}_i \geq h} w_{i,h} \hat{\gamma}^{SWDD}_{i,h}$$

where $\left\{\hat{\gamma}^{SWDD}_{i,h}\right\}_{i,h:\bar{K}_i \geq h}$ *are individual-level treatment effect estimators, defined as*

$$\hat{\gamma}^{SWDD}_{i,h} = \sum_{k=0}^{h} \left[(Y_{i,E_i+k} - Y_{i,E_i+k-1}) - \underset{\substack{j:\, D_{j,E_i}=0,\\ D_{j,E_i+k}=0}}{\frac{\mathbf{1}}{\mathbf{N}} \sum} (Y_{j,E_i+k} - Y_{j,E_i+k-1}) \right] \tag{12}$$

.

B.1 Detailed proof of Theorem 2

For clarity, I split the proof in three parts. First I show that the efficient estimator can be viewed as an OLS estimator from a particular first-differenced regression equation. Second, I show that the Regression Imputation Theorem of Borusyak *et al.* (2024) applies to this first-differenced regression equation. Finally, I use the Regression Impuation Theorem to show that the OLS estimator in question is in fact the Stepwise Difference-in-Differences estimator.

B.1.1 The efficient estimator can be viewed as an OLS estimator from a first-differenced regression equation

As noted for example by BJS, the assumptions of *No Anticipation* and *Parallel Trends* are equivalent to assuming that the data satisfy a linear regression equation of the following form (see Section G.4 for proof in a more general case):

$$Y_{i,t} = \alpha_i + \beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{13}$$

In this equation, $\{\alpha_i\}_i$ and $\{\beta_t\}_t$ are sets of (generally non-unique) fixed effects. Moreover, for i and k such that $\bar{K}_i < k$, $\gamma_{i,k}$ is an arbitrary constant, which is introduced only for notational convenience (it is a treatment effect for unit i at a time horizons where i is never observed). For i and k such that $\bar{K}_i \geq k$, the coefficient $\gamma_{i,k}$ appearing in (13) is a treatment effect of interest.

The red thread of this proof will be to consider OLS estimation of the treatment effect coefficients of interest in a version of this linear regresion. As a first step, however, we need to deal with the non-uniquenes noted above; many of the coefficients in the model are not uniquely determined by the data and assumptions. To deal with this, it will be convenient to first reparameterize the equation so that for $t > 1$, the time fixed effect β_t is replaced by a first-differenced version $\Delta\beta_t$. This can be done by rewriting the regression equation in the following cumbersome form (where $\mathbf{1}[\cdot]$ is the indicator function):

$$Y_{i,t} = \alpha_i + \beta_1 + \sum_{j=1}^{T} \mathbf{1}[t > j] \Delta\beta_j + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{14}$$

Now as noted, some coefficients in this equation will be unidentified. We can always pick some normalization however, that sets a subset of the coefficients β_1,$\{\Delta\beta_t\}_t$ and $\{\gamma_{i,k}\}_{i,k}$ equal to zero and renders all remaining coefficients identified. Importantly, the assumption that the *SGDD Estimator is Defined for the Relevant Horizons*, guarantees that for units i with $K_i \geq h$, the horizon h treatment effect $\gamma_{i,h}$ *is* identified and will not be affected by the normalization (e.g. the treatment effects of interest $\{\gamma_{i,h}\}_{i,h:\bar{K}_i \geq h}$ are not affected by the normalization). The same assumption also implies identification of any coefficient $\Delta\beta_t$ that corresponds to a period t in which some unit is observed as treated for some number of periods (i.e. where $K_{i,t} > 0$ for some i). This guarantees that

these coefficients will also not be normalized. After picking some normalization, I let θ denote the vector of non-normalized coefficients in (14), not including the fixed effects $\{\alpha_i\}_i$ or the coefficient β_1.

Next, I let Ω^D denote the data that one obtains after applying first differencing. Under *No Holes in the Data*, such differencing will remove the first observations for each unit so $\Omega^D = \{(i,t) \in \Omega : t \geq \underline{T_i} + 1\}$. Applying first-differencing now means that for $(i,t) \in \Omega^{D,1}$ the following regression equation is satisfied:

$$\Delta Y_{i,t} = \Delta\beta_t + \sum_{k=0}^{\bar{K}} \Delta H_{i,t}^k \gamma_{i,k} + \Delta\varepsilon_{i,t} \tag{15}$$

Under the normalization chosen above, the non-normalized coefficient vector θ is identified in this differenced regression and applying OLS will yield an unbiased estimator for it. Let $\hat{\theta}^{OLS}$ denote the corresponding estimator. Since the vector θ includes the treatment effect coefficents $\{\gamma_{i,h}\}_{i,h:\bar{K_i} \geq h}$, note that this OLS estimator will provide estimates of all the treatment effects of interest. Let $\hat{\gamma}_{i,h}^{OLS}$ denote the corresponding estimator of $\gamma_{i,h}$. Under *Random Walk Errors*, a standard application of the Gauss-Markov theorem to panel data now implies that the subvector of $\hat{\theta}^{OLS}$ that estimates $\{\gamma_{i,h}\}_{i,h:\bar{K_i} \geq h}$ is the best unbiased estimator for these parameters and that the same applies if one forms weighted sums of these estimators to estimate any weighted sum of treatment effects γ^w (see Section G.5 for a detailed argument in a more general case).

B.1.2 The Regression Imputation Theorem of BJS applies to the first-differenced regression equation

To complete the proof, it needs to be shown that the efficient estimator $\hat{\gamma}_{i,h}^{OLS}$ is equivalent to the individual-level Stepwise DID estimator, $\hat{\gamma}_{i,h}^{SWDD}$. As it turns out, this can be done using the Regression Imputation Theorem of BJS. This has the added benefit of establishing that SWDD estimators can be viewed as efficient Regression Imputation estimators, so that they are covered by the other results in BJS.

To see that the Regression Imputation Theorem applies here, I consider a linear reparameterization of the regression equation (15). Specifically, I will reparameterize so that for $k > h$, instead of the treatment effect coefficient $\gamma_{i,k}$ appearing in the regression, a first differenced version, $\Delta\gamma_{i,k}$,

appears which is defined by $\Delta\gamma_{i,k} = \gamma_{i,k} - \gamma_{i,k-1}$.

First I rewrite (15) as:

$$\Delta Y_{i,t} = \Delta\beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} - \sum_{k=0}^{\bar{K}} H_{i,t-1}^k \gamma_{i,k} + \Delta\varepsilon_{i,t}$$

Now I note two things: First, by definition we have $H_{i,t-1}^k = H_{i,t}^{k+1}$. Second, since no unit is observed after having been treated for $\bar{K}$ previous periods we must have $H_{i,t-1}^{\bar{K}} = 0$ for all $(i,t) \in \Omega^{D,1}$. Using this I can rewrite the second sum as:

$$\Delta Y_{i,t} = \Delta\beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} - \sum_{k=0}^{\bar{K}-1} H_{i,t}^{k+1} \gamma_{i,k} + \Delta\varepsilon_{i,t}$$

Then splitting up the first sum and shifting the index in the second sum yields:

$$\Delta Y_{i,t} = \Delta\beta_t + H_{i,t}^0 \gamma_{i,0} + \sum_{k=1}^{\bar{K}} H_{i,t}^k \gamma_{i,k} - \sum_{k=1}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \Delta\varepsilon_{i,t}$$

Combining the two sums then completes the reparameterization:

$$\Delta Y_{i,t} = \Delta\beta_t + H_{i,t}^0 \gamma_{i,0} + \sum_{k=1}^{\bar{K}} H_{i,t}^k \Delta\gamma_{i,k} + \Delta\varepsilon_{i,t} \tag{16}$$

Now consider applying OLS to (16). This will produce OLS estimators for the parameters $\gamma_{i,0}$ and $\{\Delta^1\gamma_{i,k}\}_{i,k:\bar{K}_i \geq k \geq 1}$. The OLS estimator of $\gamma_{i,0}$ is unaffected by the reparameterization so will directly be equal to the estimator of interest $\hat{\gamma}_{i,0}^{OLS}$. Additionally, letting $\widehat{\Delta\gamma_{i,k}}^{OLS}$ denote the OLS estimator of $\Delta\gamma_{i,k}$ from (16), note that we can of course recover the OLS estimators of interest simply by reversing the definition of $\Delta\gamma_{i,k}$:

$$\hat{\gamma}_{i,h}^{OLS} = \hat{\gamma}_{i,0}^{OLS} + \sum_{k=1}^{h} \widehat{\Delta^1\gamma_{i,k}}^{OLS} \tag{17}$$

Additionally, the regression equation (16) is of the same form as the ones considered by BJS.[14]

[14]To see this most clearly note that across the the sum $H_{i,t}^0\gamma_{i,0} + \sum_{k=1}^{\bar{K}} H_{i,t}^k \Delta\gamma_{i,k}$, we have $H_{i,t}^k = 1$ for at most one value of k and

$$H_{i,t}^0 \gamma_{i,0} + \sum_{k=1}^{\bar{K}} H_{i,t}^k \Delta\gamma_{i,k} = D_{i,t}\tau_{i,t}$$

where $\tau_{i,t}$ is defined by

Accordingly, the Regression Imputation Theorem thus imply that the efficient OLS estimator, $\hat{\gamma}_{i,h}^{OLS}$, can be obtained in the following steps:

1. Estimate equation (16) using only untreated observations. That is, estimate the following equation using only observations $(i,t) \in \Omega^{D,1}$ such that $D_{i,t} = 0$:

$$\Delta Y_{i,t} = \Delta\beta_t + \Delta\varepsilon_{i,t} \tag{18}$$

2. Compute the predicted values $\widehat{\Delta Y_{i,t}}$ based on the estimated model from step 1.

3. Compute the estimators $\hat{\gamma}_{i,0}^{OLS}$ and $\widehat{\Delta\gamma_{i,k}}^{OLS}$ as:

$$\hat{\gamma}_{i,0}^{OLS} = \Delta Y_{i,E_i} - \widehat{\Delta Y_{i,E_i}} \tag{19}$$

$$\widehat{\Delta\gamma_{i,k}}^{OLS} = \Delta Y_{i,E_i+k} - \widehat{\Delta Y_{i,E_i+k}} \tag{20}$$

4. Obtain the OLS estimator $\hat{\gamma}_{i,h}^{OLS}$ for $h > 0$ by applying (17).

B.1.3 Regression imputation on the first-differenced regression yields the Stepwise Difference-in-Differences estimator

Finally I show that applying the steps 1-4 from the previous section yields the SWDD estimator.

The prediction $\widehat{\Delta Y_{i,t}}$ from (18) is simple to characterize here as it will simply equal the OLS estimate of $\Delta\beta_t$ from 18 (note that this coefficient is guaranteed to not have been affected by the normalization applied earlier). This estimate is simply:

$$\frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: D_{j,t-1}=0, \\ D_{j,t}=0}} \Delta Y_{j,t}$$

Plugging into (19) and (20) we then get:

$$\tau_{i,t} = \begin{cases} 0 & \text{for } t < E_i \\ \gamma_{i,0} & \text{for } t = E_i \\ \Delta\gamma_{i,t-E_i} & \text{for } t > E_i \end{cases}$$

$$\hat{\gamma}_{i,0}^{OLS} = \Delta Y_{i,E_i} - \frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: D_{j,E_i-1}=0, \\ D_{j,E_i}=0}} \Delta Y_{j,E_i}$$

$$\widehat{\Delta\gamma_{i,k}}^{OLS} = \Delta Y_{i,E_i+k} - \frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: D_{j,E_i+k-1}=0, \\ D_{j.E_i+k}=0}} \Delta Y_{j,t}$$

Plugging into (17) then completes the proof by showing that $\hat{\gamma}_{i,h}^{OLS} = \hat{\gamma}_{i,h}^{SWDD}$.

C Cases where SGDD is efficient under random walk errors

As discussed in the main text, there are a number of relevant cases where SGDD is equivalent to SWDD, meaning that SGDD is efficient with random walk errors (cf. Theorem 2). Using the general framework from Section A, this section goes through such cases (including restating/generalizing results from the main text).

C.1 SGDD is efficient for contemporaneous treatment effects

As noted in the main text, one case where SGDD and SWDD is always equivalent is when estimating contemporaneous treatment effect at the time of treatment onset. This holds true even when adoption is staggered and when the data is an unbalanced panel, as long as the data satisfies the 'no holes' restriction. The following generalization of Corollary 1 summarizes:

Corollary. *Assume that the SGDD Estimator is Defined for the Relevant Horizons, that there are No Holes in the Data, there is No Anticipation, there is Parallel Trends and there are Random Walk Errors. Then the best unbiased estimator of any horizon 0 treatment effect ,γ_0^w, is the Subgroup Difference-in-Differences estimator, $\widehat{\gamma_0^w}^{SGDD}$.*

C.2 SGDD is efficient with non-staggered adoption and a balanced panel

Additionally, as emphasized in the main text, SGDD is equivalent to SWDD when the data is a balanced panel and adoption is non-staggered. This is contained in the following restatement of Proposition 1 from the main text:

Proposition. *Assume that the Subgroup Difference-in-Differences Estimator is Defined for the Relevant Horizons, there is No Anticipation, there is Parallel Trends, there are Random Walk Errors, the data is a Balanced Panel and there is Non-Staggered Adoption. Then the best unbiased estimator of any treatment effect γ^w is the Subgroup Difference-in-Differences Estimator, $\widehat{\gamma^w}^{SGDD}$.*

C.3 SGDD is efficient under the 'never-treated' approach

Another case where SGDD and SWDD is equivalent is if the data includes no additional preperiods for eventually treated units but there is otherwise no missing data. In this case there are never any untreated that are included in SWDD but drop out of SGDD. We thus have:

Corollary. *Assume that for any unit i where $E_i \neq \infty$, we have $(i,t) \in \Omega$ if and only $t \geq E_i - 1$. Also assume that for any unit i where $E_i = \infty$, we have $(i,t) \in \Omega$ for all $t = 1, 2, ..., \bar{T}$. If the SGDD Estimator is Defined for the Relevant Horizons, there is No Anticipation, there is Parallel Trends and there are Random Walk Errors, then the best unbiased estimator of any treatment effect γ^w is the Subgroup Difference-in-Differences estimator, $\widehat{\gamma^w}^{SGDD}$.*

The case considered in this lemma occurs if the identifying assumption for eventually treated units are assumed to *only* hold from the baseline period and onwards, so that earlier preperiods are excluded. This is exactly the 'never treated' approach proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021).

C.4 SGDD is efficient with sufficiently spaced adoption and a balanced panel

An additional case in which SGDD and SWDD is equivalent occurs in balanced panels if the treatment events are sufficiently spaced in time. In particular if there are at least h periods in between the periods where some units get treated, the SGDD estimator is efficient for treatment effects up to horizon h:

Corollary. *Assume that for all i, i' where $E_i, E_{i'} \neq \infty$, we have either $E_i = E_{i'}$ or $|E_i - E_{i'}| > h$. If the SGDD Estimator is Defined for the Relevant Horizons, the data is a Balanced Panel there is No Anticipation, there is Parallel Trends and there are Random Walk Errors, then the best unbiased estimator of any horizon h treatment effect γ_h^w is the Subgroup Difference-in-Differences estimator, $\widehat{\gamma_h^w}^{SGDD}$.*

D Illustrative example of SWDD efficiency gains over SGDD

As discussed in Section of the main text, the difference between SWDD estimator and the SGDD estimator can be summarized as follows: when estimating horizon h treatment effects for a unit treated at time E_i, SGDD uses as controls only those units that are observed untreated at both $E_i - 1$ and $E_i + h$. These are the only control units for whom it is possible to compute the total change in the outcome from $E_i - 1$ to $E_i + h$ as is done when using SGDD. In contrast, because SWDD instead uses a series of one-period-ahead comparisons it will include information from units that are observed as untreated in only some of the periods in-between $E_i - 1$ and $E_i + h$.

The two estimators thus differ only when the set of observed untreated units changes between period $E_i - 1$ and $E_i + h$. This can occur for two reasons: As noted in the main text, it will occur under staggered adoption whenever there is an untreated unit at $E_i - 1$ that becomes treated in between $E_i - 1$ and $E_i + h$. Additionally, when allowing for unbalanced panels and missing data, it can occur if an untreated observation enters or leaves the sample between $E_i - 1$ and $E_i + h$. In either case, SWDD is able to average over more untreated units in periods after E_i. Under *Random Walk Errors* this is guaranteed to improve efficiency.

Figure 1 provides a simple illustration of the above poins in a setting with 4 units: Unit A gets treated in period 2 and we are interested in estimating the horizon 1 treatment effect for this unit, $\gamma_{A,1}$. Unit B is never treated. Unit C gets treated later, in period 3. Finally, unit D is observed untreated in periods 1 and 2 but then drops out of the sample due to missing data. The SGDD estimator for $\gamma_{A,1}$ compares unit A's total change in the outcome between period 1 and 3 only to the corresponding change in the outcome for unit B (included observations marked by boxes in the figure). Units C and D are not leveraged at all because they are not observed as untreated in period 3. In contrast, the SWDD estimator for $\gamma_{A,1}$is based on summing over two one-period-ahead comparisons: The first involves changes in the outcome between period 1 and 2 and second involves changes in the outcome between period 2 and 3. Because units C and D *are* observed as untreated both in period 1 and 2, the SWDD estimator includes these units in the first comparison, (included observations marked by boldface in the figure).[15]

[15]Casual inspection of the boldfaced numbers in Figure 1 might suggest that another efficiency gain for SWDD comes from the fact that the period 2 observations for unit A and B are also directly included in the computation of the SWDD estimator. This is not the case, however. If units C and D were excluded from the figure, the data is a balanced panel and has non-staggered adoption so SWDD and SGDD is equivalent. This is because for a fixed set

Figure 1: Stepwise Difference-in-Differences leverages additional untreated observations

$t:$	1	2	3	4	5	6	7
$i:$	Treatment status ($\cdot$ = missing)						
A	$\boxed{\mathbf{0}}$	**1**	$\boxed{\mathbf{1}}$	1	1	1	1
B	$\boxed{\mathbf{0}}$	**0**	$\boxed{\mathbf{0}}$	0	0	0	0
C	**0**	**0**	1	1	1	1	1
D	**0**	**0**	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$

Observations included when estimating the horizon 2 treatment effect for unit 1 ($\gamma_{1,1}$) :

boxed: Observation included in SGDD
boldface: Observation included in SWDD

E Stepwise Difference-in-Differences Extensions: Covariates and Pretrends

Below I present two simple extensions to the SWDD estimator, which are relevant in many practical applications. First I consider the use of predetermined covariates and conditional parallel trends. Second I discuss examination of pretrends.

E.1 Covariates and conditional parallel trends

In many practical applications, it may be unreasonable to assume that parallel trends hold across all units but rather only across units which are similar in terms of some predetermined characteristics. To accomodate this, I extend the framework and setup by assuming that for each unit i, we observe some vector of predetermined covariates X_i (typically this will include a constant). Analagous to the approach in the main text, X_i will be treated as non-stochastic.

The assumption that parallel trends hold only across units with similar characteristics then corresponds to assuming that $E\left[Y^0_{i,t} - Y^0_{i,t'}\right]$ is constant only across units with the same value of X_i (e.g. parallel trends holds conditional on X_i). In estimation, however, this is often strengthened to include a linearity assumption. Following BJS, a simple way to collapse the identifying assumptions in this case is to assume that the data satisfies a linear regression model of the form:

of units (e.g. A and/or B), adding up a series of one-period-ahead changes is obviously equivalent to computing the long difference directly (the sum telescopes).

$$Y_{i,t} = \alpha_i + X_i'\beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{21}$$

Assuming that the data contains sufficient variation that this equation is identified, a simple modification of the SWDD estimator then turns out to be best unbiased under the same error and data assumptions as in the previous section (Section E.2 below provides the the proof):

Theorem 3. *Assume that the data satisfy (21) and that all coefficients in this equation are identifed. If there are No Holes in the Data and there are Random Walk Errors. then the best unbiased estimator of any treatment effect γ^w is the Stepwise Difference-in-Differences estimator, $\widehat{\gamma^w}^{SWDD}$, which here is defined as:*

$$\widehat{\gamma^w}^{SWDD} = \sum_{i,h:\bar{K}_i \geq h} w_{i,h} \hat{\gamma}_{i,h}^{SWDD}$$

where $\left\{\hat{\gamma}_{i,h}^{SWDD}\right\}_{i,h:\bar{K}_i \geq h}$ are individual-level treatment effect estimators, defined as

$$\hat{\gamma}_{i,h}^{SWDD} = \sum_{k=0}^{h} \left[\left(Y_{i,E_i+k} - Y_{i,E_i+k-1}\right) - \frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: D_{j,E_i+k-1}=0, \\ D_{j,E_i+k}=0}} \kappa_{i,E_i+k,j} \left(Y_{j,E_i+k} - Y_{j,E_i+k-1}\right) \right] \tag{22}$$

with weights $\{\kappa_{i,t,j}\}_{i,t,j}$ defined by

$$\kappa_{i,t,j} = X_i' \left(\frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{l: D_{l,t-1}=0, \\ D_{l,t}=0}} X_l X_l' \right)^{-1} X_j$$

Comparing 22 to the corresponding expression for the SWDD estimator without covariates in the main text, the only difference is that when computing the one-period-ahead average change for untreated units, the expression in 22 applies a set of weights $\{\kappa_{i,t,j}\}$ to the untreated units. For computing the individual-level treatment effect for unit i, the weight put on some untreated unit j depend on the characteristics of the this unit, X_j, vis-a-vis the treated unit, X_i. In the case where

the covariate vector X_i only contains a constant, the weights collapse to always equal one and the estimator becomes equivalent to the one from main text without covariates.

For the purpose of computation (and for conducting inference), it is useful to note that instead of applying (22) above, the SWDD estimator with covariates above can also be computed by doing Regression Imputation using the following regression model:

$$\Delta Y_{i,t} = X_i' \Delta \beta_t + \sum_{k=0}^{\bar{K}} \Delta H_{it}^k \gamma_{i,k} + \Delta \varepsilon_{i,t}$$

Equivalenty, the SWDD estimator with covariates can also be computed by first estimating this regression equation on untreated observation, then using this estimated equation to residualize the one-period-ahead change in the outcome for all observations, and then finally applying the simple SWDD formula from the main text to the residualized data. This residualization approach is equivalent to what is proposed for example by de Chaisemartin and D'Haultfœuille (2022).

Finally, note that the expression for the SWDD estimator above also suggests a natural way to compute SWDD estimators after reweighting untreated observations according to some other weighting scheme (where possibly the weights $\kappa_{i,t,j}$ does not depend on i and/or t).

E.2 Proof of Theorem 3

The proof proceeds almost identical to the proof of Theorem 2 so I limit the exposition to sketching the main steps:

The efficient estimator is equivalent to applying OLS to a regression equation that can be written in the following form:

$$\Delta Y_{i,t} = X_i' \Delta \beta_t + H_{it}^0 \gamma_{i,0} + \sum_{k=1}^{\bar{K}} H_{it}^k \Delta \gamma_{i,k} + \Delta \varepsilon_{i,t}$$

The Regression Imputation Theorem of BJS implies that for $\gamma_{i,h}$ this OLS estimator for can be computed as

$$\hat{\gamma}_{i,h}^{OLS} = \sum_{k=0}^{h} \left[\Delta Y_{i,t} - \frac{1}{\mathbf{N}} \sum_{\substack{j: K_{j,E_i+k-1}<0, \\ K_{j,E_i+k}<0}} \widehat{\Delta Y_{i,t}} \right] \tag{23}$$

where $\widehat{\Delta Y_{i,t}}$ is the predicted value from the following regression estimated only on untreated observations (observations with $D_{i,t} = 0$):

$$\Delta Y_{i,t} = X_i' \Delta \beta_t + \Delta \varepsilon_{i,t}$$

.

Writing out these predictions we have:

$$\begin{aligned}
\widehat{\Delta Y_{i,t}} = X_i' \widehat{\Delta \beta_t} &= X_i' \left(\sum_{\substack{j:\, K_{j,t}<0, \\ K_{j,t-1}<0}} \tfrac{1}{N} X_j X_j' \right)^{-1} \left(\sum_{\substack{j:\, K_{j,t}<0, \\ K_{j,t-1}<0}} \tfrac{1}{N} X_j \Delta Y_{j,t} \right) \\
&= \sum_{\substack{j:\, K_{j,t}<0, \\ K_{j,t-1}<0}} \tfrac{1}{N} \left(X_i' \left(\sum_{\substack{j:\, K_{j,t}<0, \\ K_{j,t-1}<0}} \tfrac{1}{N} X_j X_j' \right)^{-1} X_j \right) \Delta Y_{j,t}
\end{aligned}$$

Plugging for $\widehat{\Delta Y_{i,t}}$ in (23) then completes the proof.

E.3 Pretrends

Another common extension is to supplement difference-in-difference estimates of treatment effects by so-called pretrend estimates, which measure differences in the evolution of outcomes prior to treatment. Since these differences should be zero under the identifying assumption, they are often used as a validity check.

While there are several ways to construct pretrend estimates a natural approach here is to simply modify the definition of the SWDD estimator to consider treatment effects at negative horizons, e.g. treatment effects at horizon $-h$, where h is some positive integer:

$$\hat{\gamma}_{i,-h}^{SWDD} = (Y_{i,E_i-h} - Y_{i,E_i-1}) - \sum_{k=0}^{h} \left(\sum_{\substack{j:K_{j,E_i+k-1}<0, \\ K_{j,E_i+k}<0}} \tfrac{1}{N} \kappa_{i,t,j} \left(Y_{j,E_i+k} - Y_{j,E_i+k-1} \right) \right)$$

Under Parallel Trends and No Anticipation holds, the expected value of $\hat{\gamma}_{i,-h}^{SWDD}$ is zero. Accordingly inspecting the values of $\hat{\gamma}_{i,-h}^{SWDD}$ for different h can be used as check of the identifying

assumptions as usual.

F Evidence on practical relevance using Brenøe *et al.* (2024)

To provide evidence on the practical relevance of the two error benchmarks and the extent of possible efficiency gains from estimator choice, I analyze the Danish administrative data from Brenøe *et al.* (2024) (BCHH from now on). BCHH uses a non-staggered difference-in-differences design to estimate the causal effect on firms when one of their female employees gives birth and goes on parental leave. The main motivation for focusing on BCHH is the diverse set of outcome variables studied, including some that (ex ante) should approximately satisfy spherical errors, as well as some that instead appear more likely to exhibit near random walk errors. Closely related data and research designs have been used by e.g. Jäger and Heining (2022), Bertheau *et al.* (2022) and Schmutte and Skira (2023)

The main analysis in BCHH differs from the setup in this paper by applying weights and using an event-based sampling scheme that de facto allows a given firm-year to appear in the sample several times. For the analysis here, I use an adapted version of the data that instead fits this papers theoretical framework: With sampling probabilites proportional to the original weights, I randomly sample from the original data to arrive at a balanced panel of unique firms and women, which I then treat as the raw data. As I show further below, the relative performance of the estimators is virtually unchanged if I instead apply RI and SGDD/SWDD estimators directly to the original data and design.

After modifying it to fit the theoretical framework in the main text, the data from BCHH are as follows: Time periods are years and units are pairs of unique workers and firms. In each worker-firm pair, the worker is a woman satisfying some sampling criteria and the firm is her (baseline) employer. The absorbing treatment is defined as the woman becoming pregnant (and thus later giving birth). Treatment adoption is non-staggered by construction and the main analysis is carried out on a balanced panel with 7 time periods where roughly half the units get treated in period 4, while the rest remain untreated. The outcome variables of interest are a range of firm outcomes related to firm performance, as well as total employee leave-taking and fertility. The original analysis in BCHH uses an SGDD estimator to estimate average treatment effects at different horizons

(ATT_h). Below, I compare results using both the RI estimator and the SGDD/SWDD estimators (SGDD/SWDD is equivalent here because of non-staggered adoption). Following standard practice and recommendations in the literature, I use clustering at the unit level when estimating the standard error/variance of the estimates.[16]

The first column of Table 2 lists the different outcome variables of interest. I note that these include both some that (ex ante) should approximately satisfy spherical errors, as well as some that instead appear more likely to exhibit random walk errors: Total employee fertility for example is likely to reflect mostly idiosyncratic and transitory shocks suggesting that it should fit the spherical errors assumption well. Conversely, theories of sticky wages and frictional labor adjustment suggest that total employees or the firm's total wage bill is subject to very persistent shocks and thus might be well approximated by random walk errors.

The next two columns provides empirical evidence on the error persistence in the different outcome variables, based on data for 13 years of available data for untreated units: The second column shows the raw autocorrelation in residuals from a TWFE effect model (a consistent estimator of error autocorrelation as $T \longrightarrow \infty$) , while the third columns shows the Nickell (1981)-corrected autocorrelation (a consistent estimator as $N \longrightarrow \infty$ if errors are AR(1)). As expected, we see a large spread in the degree of error persistence across the different outcomes, confirming the empirical relevance of both the spherical error benchmark (corresponding to a true autocorrelation of 0) and the random walk benchmark (true autocorrelation of 1)

The last columns compares the estimated variance of treatment effects at different horizons using the RI and SGDD/SWDD estimators. For each outcome and each horizon, the table shows the variance relative to the best of the two alternatives. Results accord with the theoretical and numerical results provided earlier. For outcomes with impersistent errors, RI estimators typically have a markedly lower estimated variance, while the reverse is true for outcomes with persistent errors. The efficiency gains are also sizeable. For a range of outcomes, picking the best estimator leads to reductions in the estimated variance that are equivalent to as much as 50 percent more data .

[16] I implement the RI estimator via the `did_imputation` Stata package and implement SGDD/SWDD via my own `did_stepwise` package which relies on `did_imputation` for computation of standard errors. Results are similar using other implementations of SGDD (and numerically equivalent if using `csdid` with analytical, pointwise standard errors)

Table 2: Reanalyzing data from Brenøe *et al.* (2024)

	Residual Autocorr.	Nickell-corrected AR(1) coef.	Horizon	Estimated variance relative to best shown estimator:	
				Reg. Imputation	SGDD/SWDD
Total births at firm	0.087	0.187	0	1.000	1.530
			1	1.000	1.352
			2	1.000	1.396
			3	1.000	1.363
Total leave days at firm	0.218	0.334	0	1.000	1.354
			1	1.000	1.308
			2	1.000	1.342
			3	1.000	1.379
Number of employees	0.645	0.833	0	1.383	1.000
			1	1.226	1.000
			2	1.179	1.000
			3	1.161	1.000
New hires	0.295	0.419	0	1.000	1.008
			1	1.000	1.000
			2	1.000	1.008
			3	1.000	1.007
Turnover	0.193	0.305	0	1.000	1.107
			1	1.000	1.137
			2	1.000	1.137
			3	1.000	1.133
Hours at firm	0.714	0.923	0	1.495	1.000
			1	1.233	1.000
			2	1.167	1.000
			3	1.138	1.000
Wage bill	0.766	0.995	0	1.442	1.000
			1	1.203	1.000
			2	1.140	1.000
			3	1.109	1.000
Wage bill, excluding leave	0.766	0.995	0	1.437	1.000
			1	1.207	1.000
			2	1.139	1.000
			3	1.109	1.000
Total variable costs	0.701	0.906	0	1.531	1.000
			1	1.179	1.000
			2	1.077	1.000
			3	1.079	1.000
Total sales	0.712	0.920	0	1.510	1.000
			1	1.201	1.000
			2	1.119	1.000
			3	1.080	1.000
Profits	0.560	0.727	0	1.000	1.349
			1	1.000	1.464
			2	1.000	1.405
			3	1.000	1.358
Firm still active	0.726	0.939	0	1.311	1.000
			1	1.077	1.000
			2	1.036	1.000
			3	1.019	1.000

The table analyzes data from Brenøe *et al.* (2024), modified to match the theoretical framework from the main text. Column one reports the autocorrelation of the regression residuals from a two-way fixed effect model fit to the sample of untreated firms observed over 13 years. Column two applies the Nickell(1981)-correction to the autocorrelation from column 1 to provide a consistent estimate of the AR(1) autocorrelation in the data. The last two column compares the estimated variance when applying RI or SGDD/SWDD to the data and using clustering at the unit level (adoption is non-staggered so SGDD and SWDD are equivalent).

F.1 Alternative application BCHH data

As noted above, the main specification in Brenøe *et al.* (2024) uses a more complicated research design involving weighting and using an event-based sampling scheme that de facto allows a given firm-year to appear in the sample several times. The latter mechanically introduced correlation in outcomes (errors) across units that pertain to the same firm, which Brenøe *et al.* (2024) address by using standard errors clustered on firm. The results above in Table 2 used a modified version of the data fitting the theoretical setup of this paper. In Table 3 I insteead show results using the same sample and weigthing as BCHH, and using standard errors clustered on firm.[17] Comparing to results in the main text we see that the relative performance of the estimators is virtually identical.

G Parallel trends at horizon h and the equivalence of SGDD and Regression Imputation

In this section, I consider the case where a researcher is interested in estimating of treatment effects at some particular horizon h under a weaker identifying assumption that parallel trends assumption holds only at this horizon. As it turns out, this case is helpful both for placing SGDD estimators relative to the efficiency frontier, for clarifying the relationship between SGDD estimators and RI estimators and for further understanding the robustness properties of SGDD. In particular, with Random Walk Errors SGDD turns out to be the best unbiased estimator under the weaker parallel trends assumption, regardless of whether treatment adoption is non-staggered. Moreover, SGDD is in this case equivalent to doing efficient Regression Imputation using a particular regression model. Finally, the standard RI estimator as well as the SWDD estimator will generally be biased for all treatment effects if one only imposes this weaker parallel trends assumption thus establishing a robustness property of SGDD relative to the other estimators.

G.1 Robustness of SGDD under a weaker parallel trend assumption

Maintaining the setup and assumptions from the main text, I now consider replacing the standard parallel trends assumption with an alternative assumption that parallel trends hold only when

[17] Again I implement the RI estimator using via the `did_imputation` Stata package and implement SGDD/SWDD via my own `did_stepwise` package which relies `did_imputation` for computation of estimates and standard errors.

Table 3: Reanalyzing data from Brenøe *et al.* (2024), alternative implementation

	Residual Autocorr.	Nickell-corrected AR(1) coef.	Horizon	Estimated variance relative to best shown estimator:	
				Reg. Imputation	SGDD/SWDD
Total births at firm	0.072	0.172	0	1.000	1.543
			1	1.000	1.391
			2	1.000	1.414
			3	1.000	1.371
Total leave days at firm	0.221	0.337	0	1.000	1.354
			1	1.000	1.330
			2	1.000	1.340
			3	1.000	1.355
Number of employees	0.660	0.852	0	1.387	1.000
			1	1.231	1.000
			2	1.186	1.000
			3	1.155	1.000
New hires	0.281	0.404	0	1.000	1.000
			1	1.002	1.000
			2	1.000	1.007
			3	1.000	1.020
Turnover	0.179	0.289	0	1.000	1.123
			1	1.000	1.136
			2	1.000	1.115
			3	1.000	1.128
Hours at firm	0.729	0.944	0	1.556	1.000
			1	1.278	1.000
			2	1.194	1.000
			3	1.153	1.000
Wage bill	0.782	1.018	0	1.479	1.000
			1	1.228	1.000
			2	1.157	1.000
			3	1.115	1.000
Wage bill, excluding leave	0.782	1.018	0	1.475	1.000
			1	1.233	1.000
			2	1.158	1.000
			3	1.117	1.000
Total variable costs	0.732	0.947	0	1.566	1.000
			1	1.185	1.000
			2	1.104	1.000
			3	1.092	1.000
Total sales	0.730	0.945	0	2.364	1.000
			1	1.623	1.000
			2	1.421	1.000
			3	1.282	1.000
Profits	0.542	0.705	0	1.000	1.527
			1	1.000	1.558
			2	1.000	1.409
			3	1.000	1.377
Firm still active	0.737	0.954	0	1.322	1.000
			1	1.086	1.000
			2	1.040	1.000
			3	1.022	1.000

The table analyzes the original data from Brenøe *et al.* (2024), using reweighting and allowing firms to enter the sample several times as in the original analysis. Column one reports the autocorrelation of the regression residuals from a two-way fixed effect model fit to the sample of untreated firms observed over 13 years. Column two applies the Nickell(1981)-correction to the autocorrelation from column 1 to provide a consistent estimate of the AR(1) autocorrelation in the data. The last two column compares the estimated variance when applying RI or SGDD/SWDD to the data and using clustering at the firm level (adoption is non-staggered so SGDD and SWDD are equivalent).

looking h time periods ahead:

Assumption 14. *Parallel Trends at Horizon h: For any period t, $E\left[Y^0_{i,t+h} - Y^0_{i,t}\right]$ is constant across all units i that are observed at both t and $t+h$.*

Two remarks are in order here: First, note that if parallel trends hold horizon at h then it automatically also holds at horizons $2h, 3h$, etc. An implication of this is that *Parallel Trends at Horizon* 1 is equivalent to the standard *Parallel Trends* assumption but that *Parallel Trends at Horizon h* is strictly weaker as long as $h > 1$.

Second, as noted, I consider this weaker version of parallel trends partly for illustrative purposes. The assumption is empirically relevant, however, if outcomes happen to be subject to unit-specific seasonality (or other periodic variation). In quarterly data, for example, unit-specific seasonality would mean that the standard version of *Parallel Trends* fail, but *Parallel Trends at Horizon* 4 hold.

As is easy to verify, if one assumes *No Anticipation, Parallel Trends at Horizon h,* and that t*he SGDD Estimator is Defined for Horizon h*, then the SGDD estimator is unbiased for any treatment effect at horizon h: $E\left[\widehat{\gamma^w_h}^{SGDD}\right] = \gamma^w_h$. Since both RI and SWDD estimators will generally be biased under these assumptions, this highlights a particular robustness property of SGDD relative to these other estimators. Whether this robustness property is relevant in practice depends on the specific application. In particular, I note that if one only assumes *Parallel Trends at Horizon h* then SGDD is generally also biased for treatment effects at horizons other than h. A researcher who reports SGDD estimates at a wide range of horizons would thus generally report at least some biased estimates. Conversely, if a researcher knows with certainty that only *Parallel Trends at Horizon h* for some specific horizon h, then it is possible to construct alternative estimators which are in fact unbiased for all relevant treatment effects.

G.2 SGDD as an efficient Regression Imputation estimator

As noted, one reason for considering the weaker assumption of *Parallel Trends at Horizon h* is that the SGDD estimator for any horizon h treatment effects, γ^w_h, can be shown to be best unbiased under this assumption. Moreover, it can in fact be shown to be equivalent to an efficient Regression Imputation estimator which is based on the following regression model (see Section G.3 for derivations):

$$\Delta^{h+1} Y_{i,t} = \Delta^{h+1} \beta_t + \sum_{k=0}^{\bar{K}} \Delta^{h+1} H_{i,t}^k \gamma_{i,k} + \Delta^{h+1} \varepsilon_{i,t}$$

Here Δ^x is the x-periods-back difference operator (i.e. $\Delta^x Y_{i,t} = Y_{i,t} - Y_{i,t-x}$). Under Random Walk Errors, OLS estimation of this differenced regression equation will be efficient and as it turn out the OLS estimator for all horizon h treatment effects is in fact equivalent to the SGDD estimator. We thus have following efficiency property of SGDD in this case (see G.3 for the proof):

Theorem 4. *Assume that the SGDD Estimator is Defined for Horizon h, there are No Holes in the Data, there is No Anticipation, there is Parallel Trends at Horizon $h+1$ and there are Random Walk Errors. Then the best unbiased estimator of any horizon h treatment effect, γ_h^w, is the SGDD estimator, $\widehat{\gamma_h^w}^{DID}$.*

G.3 Proof of Theorem 4

The proof of Theorem 4 proceeds similar to the proof of Theorem 2: First I show that the efficient estimator can be viewed as an OLS estimator from a particular differenced regression equation. Second, I show that the Regression Imputation Theorem applies to this first-differenced regression equation. Finally, I use the Regression Impuation Theorem to show that the OLS estimator in question is in fact the Stepwise Difference-in-Differences estimator. Some steps in the proof rely on two auxiliary results, which I derive separately in Sections G.4 and G.5 for clarity.

G.3.1 The efficient estimator can be viewed as an OLS estimator from a differenced regression equation

Parallel Trends at Horizon $h+1$ implies that there a unit-specific periodical pattern in outcomes, with period $h+1$. In what follows it will therefore be convenient to define the function $c(t)$ as:

$$c(t) = mod(t-1, h+1)$$

To see the utility of this function, note that it reproduces the assumed periodicity, i.e. $c(1) = 1$,$c(2) = 2, ..., c(h+1) = h+1$,$c(h+2) = 1, ...$

With this definition, the assumptions of *No Anticipation* and *Parallel Trends at Horizon $h+1$*,

are equivalent to assuming that the data satisfy a linear regression equation of the following form (see Section G.4 for a full derivation):

$$Y_{i,t} = \alpha_{i,c(t)} + \beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{24}$$

In this equation, $\{\alpha_{i,c}\}_{i,c}$ and $\{\beta_t\}_t$ are sets of (generally non-unique) fixed effects. Moreover, for i and k such that $\bar{K}_i < k$, $\gamma_{i,k}$ is an arbitrary constant, which is introduced only for notational convenience (it is a treatment effect for unit i at a time horizons where i is never observed). For i and k such that $\bar{K}_i \geq k$, the coefficient $\gamma_{i,k}$ appearing in 24 is a treatment effect of interest.

The main part of this proof will be to consider OLS estimation of the treatment effect coefficients of interest in (a version of) this linear regresion. As a first step, however, we need to deal with the non-uniquenes noted above; many of the coefficients in the model are not uniquely determined by the data and assumptions. To deal with this, it will be convenient to first reparameterize the equation so that for $t = h+2, h+3, ..., T$, the time fixed effect β_t is replaced by an $h+1$-back differenced version $\Delta^{h+1}\beta_t$. This can be done by rewriting the regression equation in the following cumbersome form (where $\mathbf{1}[\cdot]$ is the indicator function):

$$\begin{aligned} Y_{i,t} = \alpha_{i,c(t)} + \sum_{c'=1}^{h+1} \mathbf{1}[c(t) = c'] \left(\beta_{c'} + \sum_{j=1}^{T} \mathbf{1}\left[t > j(h+1)\right] \Delta \beta_{j(h+1)+c'}^{h+1} \right) \\ + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \end{aligned}$$

Now as noted, some coefficients in this equation will be unidentified. We can always pick some normalization however, that sets a subset of the coefficients $\{\beta_t\}_t$,$\{\Delta^{h+1}\beta_t\}_t$ and $\{\gamma_{i,k}\}_{i,k}$ equal to zero and renders all remaining coefficients identified. Importantly, the assumption that the *SGDD Estimator Is Defined at Horizon* h, guarantees that for units i with $K_i \geq h$, the horizon h treatment effect $\gamma_{i,h}$ *is* identified and will not be affected by the normalization. The same assumption also implies identification of any coefficient $\Delta^{h+1}\beta_t$ that corresponds to a period t in which some unit is observed as having been treated for h periods (i.e. where $K_{it} = h$ for some i). This guarantees that none of these coefficients will be normalized either.

After picking some normalization, I let θ denote the vector of non-normalized coefficients in (24), not including the fixed effects $\{\alpha_{i,c}\}_{i,c}$ or the coefficients $\beta_1, \beta_2, ..., \beta_{h+1}$.

Next, I let $\Omega^{D,h+1}$ denote the data that one obtains after applying $h+1$ back differencing. Under *No Holes in the Data*, such differencing will remove the first $h+1$ observations for each unit so $\Omega^{D,h+1} = \{(i,t) \in \Omega : t > \underline{T_i} + h + 1\}$.

Now applying $h+1$ back differencing to equation 24 means that for all $(i,t) \in \Omega^{D,h+1}$ the following regression equation holds:

$$\Delta^{h+1} Y_{i,t} = \Delta^{h+1} \beta_t + \sum_{k=0}^{\bar{K}} \Delta^{h+1} H_{i,t}^k \gamma_{i,k} + \Delta^{h+1} \varepsilon_{i,t} \tag{25}$$

Under the normalization chosen above, the non-normalized coefficient vector θ is identified in this differenced regression and applying OLS will yield an unbiased estimator for it. Let $\hat{\theta}^{OLS}$ denote the corresponding estimator. Since the vector θ includes the treatment effect coefficents $\{\gamma_{i,h}\}_{i:\bar{K_i} \geq h}$, note that this OLS estimator will provide estimates of all the horizon h treatment effects of interest. Let $\hat{\gamma}_{i,h}^{OLS}$ denote the corresponding estimator of $\gamma_{i,h}$.

We next show that this OLS estimator is efficient. Under *Random Walk Errors*, the regression equation (25) has spherical errors so it follows from the Gauss-Markov Theorem that on the differenced data, $\Omega^{D,h+1}$ the OLS estimator, $\hat{\theta}^{OLS}$, is the best unibased estimatorfor θ (as well as any linear combination of its components). When the full data Ω obey an equation like (24) with an i-by-c-specific fixed effect, however, any linear unbiased estimator for θ on the full data Ω can be written as a linear estimator using only the differenced data $\Omega^{D,h+1}$ (for $h = 0$ this is a standard result used when applying Gauss-Markov to first-differenced panel data models; for completeness, Appendix G.5 contains a proof for the general case). for.[18] It follows that the best unbiased property of the OLS estimator, $\hat{\theta}^{OLS}$, holds also on the full data. This shows that the subvector of $\hat{\theta}^{OLS}$ that estimates $\{\gamma_{i,h}\}_{i:K_i \geq h}$ is the best unbiased estimator for these parameters and that the same applies if one forms weighted sums of these estimators to estimate any weighted sum of treatment effects γ_h^w.

[18] For $h = 0$ this is a standard result used when applying Gauss-Markov to first-differenced panel data models. For completeness, Appendix G.5 contains a proof for the general case.

G.3.2 The Regression Imputation Theorem of BJS applies to the differenced regression equation

All that remains is to show that $\hat{\gamma}_{i,h}^{OLS}$ corresponds to the individual level DID estimator, $\hat{\gamma}_{i,h}^{DID}$. As it turns out, this can be done using the Regression Imputation Theorem of BJS. This has the added benefit of establishing that standard SGDD estimators can be viewed as efficient Regression Imputation estimators, so that they are covered by the other results in BJS.

To use the Regression Imputation Theorem, I first apply a linear reparameterization to the regression equation (25). Specifically, I will reparameterize so that for $k > h$, the treatment effect coefficient $\gamma_{i,k}$ is replaced with an $h+1$ differenced version, $\Delta^{h+1}\gamma_{i,k}$, that is defined by $\Delta^{h+1}\gamma_{i,k} = \gamma_{i,k} - \gamma_{i,k-h-1}$.

To reparameterize, I first rewrite (25) as:

$$\Delta^{h+1}Y_{i,t} = \Delta^{h+1}\beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} - \sum_{k=0}^{\bar{K}} H_{i,t-h-1}^k \gamma_{i,k} + \Delta^{h+1}\varepsilon_{i,t}$$

Then I note two things: First, by definition we have $H_{i,t-h-1}^k = H_{i,t}^{k+h+1}$. Second, since no unit is observed is after having been treated for $\bar{K}$ previous periods, for $k > \bar{K} - h - 1$ we must have $H_{i,t-h-1}^k = 0$ for all $(i,t) \in \Omega^{D,h+1}$. Using this we can write:

$$\Delta^{h+1}Y_{i,t} = \Delta^{h+1}\beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} - \sum_{k=0}^{\bar{K}-h-1} H_{i,t}^{k+h+1} \gamma_{i,k} + \Delta^{h+1}\varepsilon_{i,t}$$

Splitting the first sum and shifting the index in the second sum this becomes:

$$\Delta^{h+1}Y_{i,t} = \Delta^{h+1}\beta_t + \sum_{k=0}^{h} H_{i,t}^k \gamma_{i,k} + \sum_{k=h+1}^{\bar{K}} H_{i,t}^k \gamma_{i,k} - \sum_{k=h+1}^{\bar{K}} H_{i,t}^k \gamma_{i,k-h-1} + \Delta^{h+1}\varepsilon_{i,t}$$

Collecting terms in the second and third sum then yields:

$$\Delta^{h+1}Y_{i,t} = \Delta^{h+1}\beta_t + \sum_{k=0}^{h} H_{i,t}^k \gamma_{i,k} + \sum_{k=h+1}^{\bar{K}} H_{i,t}^k \Delta^{h+1}\gamma_{i,k} + \Delta^{h+1}\varepsilon_{i,t} \tag{26}$$

This regression equation is of the same form as the ones considered by BJS.[19] Their Regression Imputation Theorem thus imply that the OLS estimator, $\hat{\gamma}_{i,h}^{OLS}$, can be obtained in the following

[19] To see this most clearly note that across the two sums $\sum_{k=0}^{h} H_{i,t}^k \gamma_{i,k} + \sum_{k=h+1}^{\bar{K}} H_{i,t}^k$, we have $H_{i,t}^k = 1$ for at most one value of k and

steps:

1. Estimate equation (26) using only untreated observations. That is, estimate the following equation using only observations $(i,t) \in \Omega^{D,h+1}$ such that $D_{i,t} = 0$:

$$\Delta^{h+1} Y_{i,t} = \Delta^{h+1} \beta_t + \Delta^{h+1} \varepsilon_{i,t} \tag{27}$$

2. Compute the predicted values $\widehat{\Delta^{h+1} Y_{i,t}}$ based on the estimated model from step 1.

3. Compute the OLS estimator as:

$$\hat{\gamma}_{i,h}^{OLS} = \Delta^{h+1} Y_{i,E_i+h} - \widehat{\Delta^{h+1} Y_{i,E_i+h}} \tag{28}$$

G.3.3 Regression imputation on the differenced regression yields the Subgroup Difference-in-Differences estimator

Next note that, in contrast to the general models studies by BJS, the prediction $\widehat{\Delta^{h+1} Y_{i,E_i+h}}$ is easy to characterize here. The prediction equals the OLS estimate of $\Delta^{h+1}\beta_{E_i+h}$ from (27) (note that the coefficient $\Delta^{h+1}\beta_{E_i+h}$ is not affected by any of the normalizations applied earlier). This estimate simply equals:

$$\frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: D_{j,E_i-1}=0, \\ D_{j,E_i+h}=0}} \Delta^{h+1} Y_{j,E_i+h}$$

.

Plugging into 28 we get:

$$\hat{\gamma}_{i,h}^{OLS} = \Delta^{h+1} Y_{i,E_i+h} - \frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: D_{j,E_i-1}=0, \\ D_{j,E_i+h}=0}} \Delta^{h+1} Y_{j,E_i+h}$$

$$\sum_{k=0}^{h} H_{i,t}^{k} \gamma_{i,k} + \sum_{k=h+1}^{\bar{K}} H_{i,t}^{k} \Delta^{h+1} \gamma_{i,k} = D_{i,t} \tau_{i,t}$$

where $\tau_{i,t}$ is defined by

$$\tau_{i,t} = \begin{cases} 0 & \text{for } t < E_i \\ \gamma_{i,t-E_i} & \text{for } E_i \le t < E_i + h + 1 \\ \Delta^{h+1} \gamma_{i,t-E_i} & \text{for } E_i + h + 1 \le t \end{cases}$$

This shows that $\hat{\gamma}_{i,h}^{OLS} = \hat{\gamma}_{i,h}^{SGDD}$ and completes the proof.

G.4 Proof that the identifying assumption are equivalent to the linear regression model

It is easy to verify that if the data sastifies an equation like 24 from Appendix G.3, then both *Parallel Trends at Horizon* $h+1$ and *No Anticipation* holds. The following procedure establishies the converse by showing that if *Parallel Trends at Horizon* $h+1$ and *No Anticipation* holds, we can choose values for all the relevant constants so that the data satisfy (24):

Starting with period $t = 1$, define $\alpha_{i,1} = E\left[Y_{i,1}^0\right]$ for all units i observed at $t = 1$. Then define $\beta_1 = 0$.

Now sequentially go through the periods $t = 2$, $t = 3, ..., t = T$ and do the following for each t: If there are any units i observed in the current period t that were also observed at $t-h-1$ then pick one of these units i and define $\beta_t = \beta_{t-h-1} + E\left[Y_{i,t}^0\right] - E\left[Y_{i,t-h-1}^0\right]$. Otherwise define $\beta_t = 0$. Then for any unit i observed at t for which $\alpha_{i,c(t)}$ has not yet been defined define $\alpha_{i,c(t)} = E\left[Y_{i,t}^0\right] - \beta_t$.

For the resulting set of constants, the data then satisfies (24).

G.5 Proof that unbiased estimators depend only on differences

Consider the version of equation 24 from Appendix G.3 where a suitable normalization has been applied and continue to let θ denote the identified vector of coefficients. I now introduce notation to rewrite this equation in general matrix notation.

Let p denote the number of rows in θ. Let ε be the vector of error terms in the data and let Y be the vector of outcomes. For $j = 1, 2, ..., h+1$, let $\boldsymbol{\alpha}^j$ be a vector the fixed effect coefficients, $\alpha_{i,c}$, and adopt the convention that α^j contains the coefficient pertaining to the jth period where each unit is observed. This implies that α^j is an N_j-dimensional vector, where N_j is the number of units that are oberved for at least j periods. Letting M be the total number of observations in the data, we have $M = \sum_{j=1}^{T} N_j$. Let $\boldsymbol{S}^j$ be the M-by-n_j dimensional matrix of zeros and ones that assigns the elements of α^j appropriately across the rows of observations. For an approriately defined M-by-p-dimensional matrix $\boldsymbol{Z}$ (consisting of time dummies and dummies of the form $H_{i,t}^k$), equation (24) can then be written in matrix-form as:

$$Y = \sum_{j=1}^{h+1} \boldsymbol{S}^j \alpha^j + \boldsymbol{Z}\theta + \varepsilon$$

Now let $\boldsymbol{\Delta}^{h+1}$ denote the $h+1$-back differencing matrix, let $\boldsymbol{F}^j$ be the N_j-by-M-dimensional matrix that picks out the jth observation for each unit and let $\hat{\theta}$ be some linear estimator for θ. Assuming there are *No Holes in the Data*, any such estimator, $\hat{\theta}$, can be expressed as a weighted sum of the $h+1$-back differenced outcomes and the outcomes from each of the first $h+1$ periods that each unit is in the data:

$$\hat{\theta} = \boldsymbol{\Pi}\left(\boldsymbol{\Delta}^{h+1} Y\right) + \sum_{j=1}^{h+1} \boldsymbol{\pi}_j \left(\boldsymbol{F}^j Y\right)$$

Here $\boldsymbol{\Pi}$ is the matrix of weights applied to the differenced outcomes, while $\boldsymbol{\pi}_j$ is the matrix of weights applied to outcomes from each unit's jth period in the data.

Now, if $\hat{\theta}$ is an unbiased estimator, we must have $E\left[\hat{\theta}\right] = \theta$ for any value of the parameter θ and the nuisance parameters $\alpha^1, \alpha^2, ..., \alpha^{h+1}$. Since $E\left[\varepsilon\right] = 0$, the expecation of the estimator can be written:

$$E\left[\hat{\theta}\right] = \boldsymbol{\Pi}\left(\boldsymbol{\Delta}^{h+1}\left(\sum_{j=1}^{h+1} \boldsymbol{S}^j \alpha^j + \boldsymbol{Z}\theta\right)\right) + \sum_{j=1}^{h+1} \boldsymbol{\pi}_j \boldsymbol{F}^j \left(\sum_{j=1}^{h+1} \boldsymbol{S}^j \alpha^j + \boldsymbol{Z}\theta\right)$$

Now, the fact that $h+1$ back differencing eliminates the i-by-c-specific fixed effects means $\boldsymbol{\Delta}^{h+1} \boldsymbol{S}^j \alpha^j = 0$ in matrix notation. Moreover simple index accounting implies $\boldsymbol{F}^j \boldsymbol{S}^j = \mathbb{I}_{N_j}$. We can therefore further evaluate:

$$E\left[\hat{\theta}\right] = \left(\boldsymbol{\Pi}\boldsymbol{\Delta}^{h+1}\boldsymbol{Z} + \sum_{j=1}^{h+1} \boldsymbol{\pi}_j \boldsymbol{F}^j \boldsymbol{Z}\right)\theta + \sum_{j=1}^{h+1} \boldsymbol{\pi}_j \alpha^j$$

But clearly this implies that unbiasedness can only hold if $\boldsymbol{\pi}_j = 0$ for all j (and also $\boldsymbol{\Pi}\boldsymbol{\Delta}^{h+1}\boldsymbol{Z} = \mathbb{I}_p$). This implies that $\hat{\theta}$ can be expressed as a linear combination of only the $h+1$-back differenced outcomes, $\hat{\theta} = \boldsymbol{\Pi}\left(\boldsymbol{\Delta}^{h+1} Y\right)$, which completes the proof.