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

Nikolaj A. Harmon

University of Copenhagen

First version: November 2022

This version: December 2023

Abstract

Doubts have been raised about the efficiency of modern difference-in-difference estimators in the vein of de Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2021). I show that these estimators in fact have attractive efficiency properties under a benchmark 'strong persistence' assumption for errors: With non-staggered adoption and a balanced panel, they are best unbiased estimators for any treatment effect. With staggered adoption, the estimators remain efficient for estimating effects immediately at treatment onset but a novel adjusted 'Stepwise Difference-in-Differences' estimator is the best unbiased estimator for effects at longer horizons. The results provide a simple guide to estimator choice.

*E-mail: nikolaj.harmon@econ.ku.dk. The paper has benefited from discussions with Kirill Borusyak, Thomas Jørgensen, Anders Munk-Nielsen, Clement de Chaisemartin, Jesper Riis-Vestergaard, as well participants at the Conference of the European Association of Labor Economics 2023 and the Applied Econometrics Reading Group at University of Copenhagen. The work was supported by a grant from the Independent Research Fund Denmark.

A string of recent papers have developed treatment effect estimators that apply to difference-in-differences designs with hetereogeneous treatment effects and possible staggered treatment. While this has been a boon for applied research, it has also placed a new burden of choice on applied researchers. When analyzing a difference-in-differences design, researchers now need to make a potentially very important choice of estimator. Moreover, statistical theory demands that the choice be made ex ante; while prudent researchers often report results from many different estimators for transparency, this practice leads to problems if the preferred estimator is not specified in advance.

For the canonical difference-in-differences design with an absorbing discrete treatment and no covariates, applied researchers face a choice between two broad groups of estimators: The first group is what I term 'Subgroup Difference-in-Differences' (SGDD). SGDD estimators selects particular subgroups of treated and untreated observations and forms direct difference-in-differences comparisons beetween these groups (e.g. de Chaisemartin and D'Haultfœuille (2020, 2022); Sun and Abraham (2021); Callaway and Sant'Anna (2021); see also Dube *et al.* (2022)). The second group is what I refer to as 'Regression Imputation' (RI) estimators. RI estimators compare actual outcomes for treated units with imputed counterfactual outcomes from a particular linear regression model (e.g. Borusyak *et al.* (forthcoming); Gardner (2022); see also Wooldridge (2021)).[1]

In choosing between SGDD and RI estimators, efficiency considerations should play an important role. In a given sample, both groups of estimators are unbiased under the same assumptions so picking the efficient alternative means getting more precise estimates for the same data. Based on existing results, however, efficiency comparisons between SGDD and RI are lopsided and incomplete. While Borusyak *et al.* (forthcoming) (BJS from now on) has established that Regression Imputation is the best unbiased estimator under spherical errors, the efficiency properties of Subgroup Difference-in-Differences estimators are largely unknown.[2] A particular concern is that SGDD estimators may have generally poor efficiency properties because they only use data on the single period just before treatment, thus ignoring information from additional pretreatment periods.

This paper provides efficiency results for SGDD estimators. To do this, I use the same framework as BJS but consider an alternative benchmark assumption on errors. While BJS's spherical errors assumption is a common benchmark, it is also an extreme benchmark in the sense that it imposes

[1] I use the *Regression* Imputation for clarity. Borusyak *et al.* (forthcoming) has shown that any linear unbiased treatment effect estimator can be viewed as doing some form of imputation of counterfactual outcomes.

[2] An exception is the discussion of efficiency in Marcus and Sant'Anna (2021).

no correlation in errors over time. Here I consider the opposite benchmark where errors are strongly correlated over time. Specifically, I assume that errors follow a random walk, as will be the case if errors reflect that units in the data are subject to permanent shocks. Since many microeconomic processes are in fact modelled as being subject to permanent or strongly persistent shocks, efficiency properties under this benchmark should also be empirically relevant in many settings.

I first consider the case where the data is a balanced panel and where treatment adoption is non-staggered, e.g. all eventually treated units get treated in the same period. Under these assumptions, I show that SGDD estimators are the best unbiased estimators for any weighted sum of treatment effects. This efficiency turns out to hold exactly *because* SGDD estimators only rely on the last period prior to treatment: when errors reflect persistent shocks, using data from any additional preperiod will only add additional noise stemming from the shocks occuring in between this preperiod and treatment onset. This result thus also provides a rigorous theoretical justification for the standard practice of choosing the last untreated period as the baseline for SGDD estimators. A simple simulation study adapted from BJS show that the efficiency gains of SGDD can be large when errors are persistent: Relative to RI estimators, SGDD provides efficiency gains that are equivalent to as much as 60 percent more data when errors follow a random walk.

Next I consider the general case of staggered treatment adoption and also allow for some forms of unbalanced panels. In this setting, SGDD estimators remain the best unbiased estimators for any weighted sum of contemperaneous treatment effects at the time of treatment onset. For treatment effects at longer time horizons, however, I show that the best unbiased estimator is a novel adjuisted estimator that I term 'Stepwise Difference-in-Differences' (SWDD). Instead of considering long differences across several time periods as the SGDD estimator does, the SWDD estimator computes treatment effect estimates step-by-step in a series of one-period ahead comparisons. I clarify how this stepwise estimation leads to efficiency gains: with missing data or under staggered treatmet adoption, the SWDD estimator is able to leverage data on more untreated units when estimating treatment effects at longer horizons. A simple simulation show that the associated efficiency gains can be large: Relative to SGDD, using SWDD under staggered adoption provides efficiency gains equivalent to as much as 35 percent more data when estimating treatment effects 4 periods after treatment onset.

Combined with previous work, these results suggest a simple principle for estimator choice in

practice: If the errors in outcome variable can be expected to exhibit low serial correlation - as is the case if the errors reflect transitory shocks or idiosyncratic measurement error - Regression Imputation should perform well. Conversely, if the errors in outcome variable can be expected to exhibit high serial correlation - as is the case if they reflect mostly persistent shocks - SGDD should perform well when estimating short-run treatment effects or when adoption is non-staggered, while SWDD will improve efficiency when estimating longer-run treatment effects under staggered adoption. Using data from the difference-in-difference design in Brenøe *et al.* (forthcoming) I confirm the practical relevance of this approach. The behavior of the outcome variables in Brenøe *et al.* (forthcoming) range from nearly uncorrelated errors to near-random walk errors. Accordingly, choosing the right estimator for each outcome delivers reductions in the width of confidence intervals equivalent to as much as 50 percent more data.

This paper builds on and contributes to the growing literature on difference-in-differences and event study designs. In addition to the implications for practice, the paper also makes some technical contributions, particularly underscoring the versatilify of the methodology developed in BJS. In addition to relying on the BJS framework, the proofs presented in the appendix imply that any SGDD estimator (and the SWDD estiamator) can be viewed as an efficient Regression Imputation estimator based on a particular linear regression model. In addition to clarifying the relationship between these groups of estimators, this also implies that all tools provided by BJS can be directly applied also to SGDD and SWDD estimators, including their approach to computation, their approach to (fixed sample) inference as well as their approach to addressing the pretrend test problems pointed out by Roth (2022).

Besides efficiency, an alternative approach to estimator choice is to consider the extent of bias in the two estimators under violations of the identifying assumptions (see e.g. Roth *et al.* (2023)), although this will depend on the exact way in which the assumptions fails (see Borusyak *et al.* (forthcoming)). Appendix E of this paper provides some results in this direction by establishing robustness of the SGDD estimator to a certain violation of parallel trends.

Finally, since the original circulation of the present paper, Bellégo *et al.* (2023) has proposed a class of 'Chained Difference-in-Differences' estimators that are closely related to the Stepwise Difference-in-Differences estimator derived here. The aim and results in their paper are fundamentally different, from mine however. While Bellégo *et al.* (2023) address the complications arising

when balanced panels are not available, the efficiency results in this paper apply also in the case of a balanced panel.

1 Framework and assumptions

With the exception of notational changes, I adopt the same fixed sample framework as BJS, treating the realized sample and treatment timing as non-stochastic.[3]

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. At this point, I allow the data to 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.

I consider the standard case where treatment is an absorbing state meaning that for each unit there is some period E_i when treatment occurs and $D_{i,t}$ switches from zero to one. The possibility that some units are never treated is allowed for and corresponds to $E_i = \infty$. Note that the framework covers both the case of staggered treatment adoption and the case where treatment happens at the same time for all units that ever receive treatment.

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. For a given period t and unit i, the condition $K_{i,t} < 0$ thus means that the unit is not yet treated in this period. Conditions of this form will appear frequently in the analysis.

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

[3]See BJS for discussion and results that links the framework to a (super)population framework with random sampling.

dummy variable for whether at time t, the unit i is treated and has experienced the treatment for exactly k previous periods:

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

In the analysis presented later, averages over units that satisfy certain conditions will play a prominent role. I therefore adopt some simplifying notation here. 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.[4] In a slight abuse of notation, I will 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* at both times:

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

1.1 Potential outcomes, treatment effects and estimands

Treatment effects are defined relative to a situation where units never experience the treatment. Accordingly, $Y_{i,t}^0$ denotes 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. Note that the treatment effect at horizon $h = 0$ thus corresponds to the contemporaneous effect at the initial onset of treatment.

With this as the building block, I will consider the case where the estimand of interest is some 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. This flexible formulation covers most standard estimands in the literature. For example, one candidate for γ_h^w is the conditional average

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

horizon h treatment effect for units first treated at time t, written as $CATT_{t,h} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{i:K_{i,t+h}=h} \gamma_{i,h}$. Such cohort-by-horizon specific average effects are often used as building block estimands in causal inference (see e.g. Sun and Abraham (2021); Callaway and Sant'Anna (2021); de Chaisemartin and D'Haultfœuille (2022)).

Another example of an estimand that may serve as γ_h^w is the average treatment effect at horizon h across all units in the sample observed at horizon h. We can write this as $ATT_h = \sum_t \omega_t CATT_{t,h}$ for appropriately defined sample share weights $\{\omega_t\}_{t=1,2,..,T}$.[5] This estimand is a common target parameter in applied work.[6]

For most theoretical results, however, I consider the more general case where a researchers may also be interested in averaging treatment effects across different time horizons. This corresponds to the having the estimand $\gamma^w = \sum_{i,h:\bar{K}_i \geq h} w_{i,h}\gamma_{i,h}$ for some set of weights $\{w_{i,h}\}_{i,h:\bar{K}_i \geq h}$. Note that any weighted sum of treatment effects at a particular horizon, γ_h^w, is a special case of the more general weighted sum γ^w (with $w_{i,h'} = 0$ for all $h' \neq h$). It follows trivially that an estimator that is efficient for any γ^w will also be efficient for any γ_h^w.

1.2 Subgroup Difference-in-Differences estimators

Next, I define Subgroup Difference-in-Differences (SGDD) estimators for the estimands, γ_h^w and γ^w:

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*

$$\widehat{\gamma_h^w}^{SGDD} = \sum_{i:\bar{K}_i \geq h} w_i \hat{\gamma}_{i,h}^{SGDD} \tag{1}$$

$$\widehat{\gamma^w}^{SGDD} = \sum_{i,h:\bar{K}_i \geq h} w_{i,h} \hat{\gamma}_{i,h}^{SGDD} \tag{2}$$

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

[5] That is $\omega_t = \frac{\#\{i:K_{it+h}=h\}}{\#\{i,t':K_{it'+h}=h\}}$.

[6] Estimates of ATT_h are reported by common Stata packages (e.g. `did_multiplegt` with the option `dynamic_robust`, `did_imputation` with `allhorizons` and `csdid` with `agg(event)`).

$$\hat{\gamma}_{i,h}^{SGDD} = (Y_{i,E_i+h} - Y_{i,E_i-1}) - \underset{\substack{j:K_{j,E_i-1}<0,\\ K_{j,E_i+h}<0}}{\frac{\mathbf{1}}{\mathbf{N}} \sum} (Y_{j,E_i+h} - Y_{j,E_i-1}) \tag{3}$$

To unpack this definition start by considering the individual-level treatment effect estimates in 3. The first term is the change in i's outcome from the last period before i is treated ($t = E_i - 1$) until the period when i has been treated for h periods ($t = E_i + h$) . The second term subtracts off the average corresponding change for all units that are observed as untreated both at $t = E_i - 1$ and $t = E_i + h$.[7] To arrive at estimators for the weighted sums, the relevant weights are simply applied to the individual-level estimates as shown in 1 and 2.

To see that this definition indeed corresponds to the standard Subgroup DID estimators in the literature, note that if we set the estimand γ_h^w to be $CATT_{t,h}$, we arrive at the familiar expression:

$$\widehat{CATT_{t,h}}^{DID} = \underset{i:K_{i,E_i+h}=k}{\frac{\mathbf{1}}{\mathbf{N}} \sum} (Y_{i,E_i+h} - Y_{i,E_i-1}) - \underset{\substack{j:K_{j,E_i-1}<0,\\ K_{j,E_i+h}<0}}{\frac{\mathbf{1}}{\mathbf{N}} \sum} (Y_{j,E_i+h} - Y_{j,E_i-1})$$

1.3 Restrictions on observed and missing data

The setup so far imposes no restrictions on what types of units exists 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 1. *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 i' with $(i',t), (i', E_i - 1) \in \Omega$ and both $K_{i',E_i-1} < 0$ and $K_{i',t} < 0$.*

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

[7] The restriction in the second term to require both $K_{j,E_i-1} < 0$ and $K_{j,E_i+h} < 0$ is necessary because the framework allows for an untreated unit at $E_i + k$ to not have been observed at $E_i - 1$. For this reason, the expression for the individual DID estimators may actually also be undefined if no units satisfy the restriction in second term. In Section 1.3 below I impose conditions to ensure that rule out this possibility.

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.[8]

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

Assumption 2. *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 3. *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 maintained throughout the main analysis. In some parts of the analysis, I will further strengthen the assumptions however. For some results, I will require the data to be a balanced panel:

Assumption 4. *Balanced Panel: For each unit i and each $t \in \{1, 2..., \overline{T}\}$, we have $(i,t) \in \Omega$.*

[8] A relevant example is the following: There is a unit i first treated at time t, which is observed at both $t-1$ and $t+1$. There is also a unit i' which is observed as untreated at both $t-1$ and t and another unit i'' that is observed as untreated both at time t and $t+1$. Now if there are no units that are observed as untreated at both $t-1$ and $t+1$ then the DID estimator for i's treatment effect at horizon 1 is undefined. Under the *No Anticipation* and *Full Parallel Trends Assumption* presented below, however, this treatment effects is identified from appropriate comparisons between i, i' and i'' over the time periods $t-1, t$ and $t+1$. Bellégo *et al.* (2023) discusses estimation in such case where standard difference-in-differences estimators are undefined due to missing data.

Finally, for some results I additionally require that treatment adoption is non-staggered so that all units that are eventually treated receive the treatment at the same time:

Assumption 5. *Non-Staggered Adoption: For any pair of units (i, i') such that $E_i, E_{i'} \neq \infty$, we have $E_i = E_{i'}$.*

1.4 Identifying assumptions

Identification will rest on two standard assumptions throughout. The first is a no anticipation assumption, restricting the timing of how treatment affects outcomes:

Assumption 6. *No Anticipation: $Y_{i,t} = Y^0_{i,t}$ whenever $K_{i,t} < 0$.*

As written this assumption imposes that eventual treatment does not affect outcomes in periods before the treatment occurs. As is well known, however, a simple relabeling of treatment variables makes it trivial to cover cases where outcomes might be affected some known number of periods before the treatment occurs.

The second assumption will be a parallel trends assumption, imposing that in the absence of treatment, outcomes move in parallel.

Assumption 7. *Parallel Trends: For any two periods t and t', $E\left[Y^0_{i,t} - Y^0_{i,t'}\right]$ is constant across all units i that are observed at both t and t'.*

As is immediately apparent, the identifying assumption above, are equivalent to the one used in for example BJS and de Chaisemartin and D'Haultfœuille (2020). Other previous work by Callaway and Sant'Anna (2021) and Sun and Abraham (2021) have proposed different ways of weakening the parallel trends assumption, as well as corresponding adjustments to SGDD estimators. In the present framework, these alternative estimators and assumptiosn are covered by simply by appropriately restricting the data (and thus restricting the set of observations on which the parallel trends assumptions are imposed).[9] The efficiency results presented later thus apply to these SGDD estimators as well.

[9] 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 DID 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 DID estimators become equivalent to the never-treated approach of Sun and Abraham (2021) and Callaway and Sant'Anna (2021).

1.5 Error benchmarks

Define $\varepsilon_{i,t} = Y_{i,t} - E[Y_{i,t}]$ to be the error for unit i in time period t. Estimator efficiency will depend on the behavior of these errors. BJS derives the efficient estimator under the assumption that these errors are spherical, that is homeskedastic and serially uncorrelated. Spherical errors is a standard and widely-used benchmark. With regards to persistence, however, it is also an extreme benchmark since it imposes that there is no correlation in errors over time. If we interpret the errors as reflecting shocks to the units, serially uncorrelated implies that shocks are completely transitory.

In this paper, I instead consider the opposite benchmark that shocks are completely persistent. This corresponds to imposing a random walk assumption on the errors, e.g. that the differences in the errors, $\varepsilon_{i,t} - \varepsilon_{i,t-1}$, are homoskedastic and serially uncorrelated. To cover the case where data is missing for some periods and to help build intuition for the results presented later, it is convenient to formulate this assumption explicitly in terms of a set 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 8. *Random Walk Errors: For any $h \geq 1$, the errors satisfy*

$$\varepsilon_{i,t+h} - \varepsilon_{i,t} = \sum_{t'=t+1}^{t+h} \eta_{i,t'} \tag{4}$$

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

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

As equation 4 makes clear, this assumption implies that the difference in the error between two points in time can be written as the sum of the uncorrelated shocks that have occurred in-between. This insight will be useful for interpreting the efficiency results later.

Finally, it is worth emphasizing that I use the assumption of Random Walk Errors only as a benchmark to characterize *when* the different estimators have attractive efficiency properties. Neither the SGDD estimator, the RI estimator or the Stepwise Difference-in-Differences estimator introduced later require Random Walk Errors to be unbiased. The assumption will also not be necessary for inference. As usual, standard cluster-robust inference is likely to be the preferred

approach in almost all settings.

2 Theoretical efficiency results

I now present my results on efficient estimation of treatment effects. As usual when studying unbiased estimators, I use the term 'best estimator' to refer to the estimator with the lowest possible variance. I relegate all detailed proofs to Appendix A but provide sketch arguments in the main text when this is useful.

2.1 Efficiency under non-staggered adoption and a balanced panel

To set up the first efficiency results for SGDD, it is useful to recap the key insights underlining the existing efficiency results in BJS. The first insight is that in the present framework, the assumptions of *Parallel Trends* and *No Anticipation,* imply that the data satisfy the following regression model:

$$Y_{i,t} = \alpha_i + \beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{5}$$

Importantly, all the individual treatment effects of interest appear directly as coefficients in 5. If the errors term is spherical, a standard application of the Gauss-Markov theorem therefore implies that applying OLS to 5 will yield efficient estimators for any linear combination of treatment effects. Since the RI estimator proposed by BJS is equvialent to applying OLS in equation 5, it follows that the RI estimator is efficient in this case.

In standard texts on panel data, the efficiency result above is often phrased as saying that the 'within-estimator' is efficient under homoskedastic and serially uncorrelated errors. This is because the relevant OLS estimator can be conveniently computed by first applying within-unit demeaning to equation 5. Another standard result in this literature, however, is that if errors instead follow a random walk, the efficient estimator is instead the 'first difference' estimator which can be obtained by applying OLS to a first-differenced version of equation 5:

$$\Delta Y_{i,t} = \Delta \beta_t + \sum_{k=0}^{\bar{K}} \Delta H_{i,t}^k \gamma_{i,k} + \Delta \varepsilon_{i,t} \tag{6}$$

The first new efficiency result follow from showing that - in a balanced panel with non-staggered

adoption - SGDD estimators are equivalent to OLS estimation of the first-differenced regression equation 6:

Theorem 1. *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, $\widehat{\gamma^w}^{SGDD}$.*

Theorem 1 establishes that the SGDD estimator utilizes the data efficiently under the imposed assumptions thus formally establishing SGDD as an admissible estimator. In terms of the cross-sectional variation across units, this efficiency should be unsurprising. Assuming *Random Walk Errors* implies that shocks have the same variance across units and the individual-level DID estimators involves simple averages across units.

It might be more surprising that the SGDD estimator also exploits the data efficiently in the time dimension. When estimating the horizon h treatment effect for some unit, the SGDD estimator only uses data from h periods after treatment and from the *single* period immediately before the onset of treatment. As emphasized by BJS, however, there should generally be many more periods available prior to treatment onset which could also be used for estimation.

To build intution for why the SGDD estimator still turns out to be efficient, consider some specific unit i that gets treated at E_i and another (control) unit j which remains untreated at least until E_i. Now consider computing the difference in outcomes for these two units between the treatment period E_i and some arbitrary pretreatment period $b < E_i$. Under *Random Walk Errors*, these differences will equal the expected change in the outcome between b and E_i plus noise stemming from the sum of the uncorrelated shocks occuring between those periods:

$$(Y_{i,E_i} - Y_{i,b}) = E\left[Y_{i,E_i} - Y_{i,b}\right] + \sum_{t=b+1}^{E_i} \eta_{i,t}$$

$$(Y_{j,E_i} - Y_{j,b}) = E\left[Y_{j,E_i} - Y_{j,b}\right] + \sum_{t=b+1}^{E_i} \eta_{j,t}$$

The reason *No Anticipation* and *Parallel Trends* makes it possibly to identify treatment effects is that they imply $E\left[Y_{i,E_i} - Y_{i,b}\right] - E\left[Y_{j,E_i} - Y_{j,b}\right] = \gamma_{i0}$ so that the difference in the diffences above

equals the treatment effect for unit i plus a mean zero noise term reflecting the difference in the shocks:

$$(Y_{i,E_i} - Y_{i,b}) - (Y_{j,E_i} - Y_{j,b}) = \gamma_{i,0} + \sum_{t=b}^{E_i} (\eta_{i,t} - \eta_{j,t}) \tag{7}$$

However, the fact that the noise term is a sum over uncorrelated shocks occuring between b and E_i, now implies two things: First, for any difference-in-differences comparison of the form in 7, we will get the least noise (smallest variance) if the preperiod b is set to be the baseline period immediately before treatment, $b = E_i - 1$ because this implies that the noise terms is the difference between a single pair of shocks $\eta_{i,E_i} - \eta_{j,E_i}$. Second, making the comparison in 7 using any earlier preperiod $b < E_i - 1$ would simply mean adding additional noise by including more serially uncorrelated shocks. Both of these insights hold also if considering some later posttreatment period $E_i + h$ and/or if considerings means over more units. Accordingly, under Random Walk Errors, the most efficient estimator for a given treatment effect will only consider the single pre-period just before treatment onset. This is exactly what the SGDD estimator does.[10]

2.2 Efficiency with staggered adoption and/or missing data

Theorem 1 above show that the SGDD estimator has attractive efficiency properties under Random Walk errors. In addition to Random Walk Errors and the standard identifying assumptions, however, Theorem 1 also requires the data to be a balanced panel and treatment adoption to be non-staggered. The latter in particular is a restrictive condition. Staggered adoption is a common occurence in applications and allowing for it in estimation has been a major impetus for the development of modern difference-in-differences estimators. I thus now consider efficiency in the general case where treatment adoption may be staggered. I also weaken the balanced panel assumption to instead only require *No Holes in the Data*.

Analogously to above, the efficient estimator in this case can be derived as an application of OLS to the first-differenced equation 6. The efficient estimator turns out to be an adjusted difference-in-differences estimator that I term 'Stepwise Difference-in-Differences' (SWDD):

[10] Note that without the assumption of Random Walk Errors, the expression in 7 will of course still hold if we just define $\eta_{i,t} \equiv \varepsilon_{i,t} - \varepsilon_{i,t-1}$. In this case however the series of $\eta_{i,t}$s will typically be negatively correlated over time. This means that drawing on comparisons also from earlier preperiods can improve efficiency because the noise in the different comparisons partially cancels each other out.

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*

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

.

To understand why Stepwise DID is an appropriate name for this efficient estimator, consider the expression for the individual-level estimator, $\hat{\gamma}^{SWDD}_{i,h}$ and note that the expression inside brackets is simply a one-period difference-in-differences comparison involving a one-period change in the outcome for unit i and the corresponding average one-period change among all untreated units. Taking into account the outer sum, we thus see that the SWDD estimator is simply a sum over $h+1$ such one-period differences. In other words, where existing SGDD estimators estimate the horizon h treatment effect by comparing the total change in the outcome from $E_i - 1$ to $E_i + h$ across treated and untreated units, the SWDD estimator instead works step-by-step, by first constructing a series of one-period ahead comparions and then summing these up to arrive at the final estimate.

Next, to understand why the SWDD estimator is more efficient than SGDD with staggered adoption or unbalanced panels, first rewrite the individual-level SWDD estimator as:

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

Then note that the first sum over k telescopes so that we have:

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

The first term here is identical to the first term in the expression for the SGDD estimator and is simply the total change in the outcome of the treated unit i between E_i-1 to E_i+h. The differences between the DID and the SDD estimators thus come entirely from the second term, which relates to the untreated units.

Next, consider what happens *if* the set of observed untreated units is the same in all periods between period $E_i - 1$ and $E_i + h$. In this case the remaining sum over k also telescopes and will just equal the average total change in the outcome among these units:

$$\frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: K_{j,E_i-1}<0, \\ K_{j,E_i+h}<0}} (Y_{j,E_i+h} - Y_{j,E_i-1})$$

This is the same as the second term in the expression for the SGDD estimator so in this case the SWDD estimator will equal the SGDD estimator. It follows that the two estimators 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: It can occur because of missing data if an untreated observation enters or leaves the sample between $E_i - 1$ and $E_i + h$. More importantly, under staggered adoption, it will also occur whenever there is an untreated unit that becomes treated in between $E_i - 1$ and $E_i + h$. In both cases, the untreated unit in question will be excluded from the long-differences used in the SGDD estimator because it is not observed as untreated at *both* $E_i - 1$ and $E_i + h$. As long as it is observed for at least two periods, however, this untreated unit *will* be included in at least some of the one-period comparisons used in the SWDD estimator. This clarifies why the SWDD estimator is more efficient: The SWDD estimator is able to average over more untreated units in periods after E_i and under *Random Walk Errors* this is guaranteed to improve efficiency.

Figure 1 provides a simple illustration of the above point 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

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

t :	1	2	3	4	5	6	7
i :	Treatment status ($\cdot$ = missing)						
A	**0**	**1**	**1**	1	1	1	1
B	**0**	**0**	**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

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).

The discussion above also highlight that the efficiency gains of SWDD will be particularly large when the estimand corresponds to treatment effects at longer horizons: For estimating the treatment effect for a given individual i and horizon h, the efficiency gains will be large when there are many potential control units that get treated or drop out of the sample between the baseline period $E_i - 1$ and relevant posttreatment period $E_i + h$. If estimating treatment effects at longer time horizons, there is a longer gap between the baseline period and the relevant posttreatment period which mechanically means that more potential control units get treated or drop out. Conversely, in the extreme case where a researcher is interested only in the contemporaneous treatment effect at horizon 0, no potential control units can get treated or leave the sample, meaning that the SWDD and SGDD estimators coincide in this case. This immediately yields the following corollary:

Corollary 1. *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, $\widehat{\gamma_0^w}^{SGDD}$.

For researchers who care only about contemporanoues treatment effects immediately at the onset of treatment, SGDD estimators thus retain their efficiency under non-staggered adoption.

2.3 Additional results

Below I briefly discuss some additional theoretical results that are expanded on in the appendices.

First, the derivation of the SWDD estimator in Appendix A shows that it can be viewed as an efficient RI estimator of the type considered by BJS. One implication of this is that their approaches to both (cluster-robust) inference and computation also apply to the SWDD estimator. A Stata package implementing SWDD estimation and inference in this way is available on my website (`did_stepwise.ado`). The package also implements some extensions of the SWDD estimator to cover estimation with predetermined covariates (conditional parallel trends) and examine the identifying assumption by estimating pretrends. These extensions are discussed in Online Appendix C.

Second, in Online Appendix D, I describe two other special cases where SGDD remains efficient even with staggered adoption. The first is the case where the identifying assumption for eventually treated units are assumed to only hold from the baseline period and onwards, meaning that all earlier preperiods are excluded from the analysis (corresponding to the 'never treated' approach discussed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021)). In this case, SGDD is efficient for any treatment effect, under a suitable restriction on the extent of missing data. The second special case I consider is when treatment adoption is sufficiently spaced in time. If time periods where treatment adoption happens are spaced out by at least h periods and the data is a balanced panel, SGDD is efficient for treatment effects up to horizon h. In both cases, the maintained efficiency of SGDD reflects that in these special cases SWDD and SGDD estimators coincide because there are no additional untreated observations that SWDD leverages but SGDD does not.

Third, the analysis above compares estimators that require exactly the same assumptions for

unbiasedness. A natural question, however, is whether some estimators may be more robust to certain violations of the identifying assumptions. In Online Appendix E, I provide a result in this vein by showing that - also under staggered adoption - SGDD is the best unbiased estimator for treatment effects at any particular horizon h if the parallel trend assumption is weakened to hold *only* at this time horizon h.[11] In addition to establishing a formal robustness property of SGDD, the derivation of the result also implies that the SGDD estimator can be viewed as an efficient RI estimator under this weaker parallel trend assumption. In addition to clarifying the relationship between estimators, this result extends the tools and results provided by BJS to apply to SGDD estimators.

Finally, note that source of the efficiency gain of SWDD over SGDD applies also when comparing RI estimators to SGDD estimators. In the example of Figure 1 for example, an RI estimator would also leverage data on the additional control units that SWDD does but SGDD ignores. This source of efficiency gains seems to have received less attention in previous comparisons of RI vs SGDD, which have tended to emphasize the efficiency gains stemming from using additional preperiods. The additional source of efficiency gains help explain why RI estimators are often found to outperform SGDD estimators at longer horizons, even in situations where there is no additional preperiods available.

3 Numerical results

The theoretical analysis above shows that under Random Walk Errors, SGDD estimators provide efficiency gains relative to RI estimators if treatment is non-staggered, while SWDD estimators provide additional efficiency gains relative to SGDD under staggered adoption. To provide evidence on the extent and practical relevance of these efficiency gains, this section first discusses a numerical simulation and then results using data from Brenøe *et al.* (forthcoming).

[11] The practical elevance of this robustness property will depend on the application. The assumption that parallel trends hold only at certain horizons will occur in practice if the outcome is subject to unit-specific seasonality. If researchers do not know in advance at which horizons parallel trends hold, however, applying SGDD is likely to produce biased estimates of at least some treatment of interests because SGDD is biased for treatment effects at most horizons in this case. Conversely, if researchers know in advance at which horizons parallel trends hold, it is possible to construct alternative estimators which are unbiased and efficient for all treatment effects of interest.

3.1 Simulation results

To provide simulation evidence on the extent of the efficiency gains provided by SWDD and/or SGDD, I add persistent errors into a simulation setup originally introduced by BJS. Using the notation from Section 1, the data is a balanced panel with 250 units, observed over the periods $t = 1, 2, ..., 6$. Following BJS, I draw treatment assignment once under the assumption that E_i is *iid* uniform on the set $\{2, 3, ..., 6, \infty\}$. I then generate 500 simulations according to the following model, which satisfies the assumptions from Section 1:

$$\begin{aligned} Y_{i,t} &= \alpha_i + \beta_t + \sum_{k=0}^{4} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \\ \alpha_i &= -E_i \\ \beta_t &= 3t \\ \gamma_{i,h} &= 1 + h \\ \varepsilon_{i,t} &= \rho \varepsilon_{i,t-1} + \eta_{i,t} \end{aligned}$$

To impose Random Walk Errors, the baseline simulation uses $\rho = 1, \varepsilon_{i,1} = 0, \eta_{i,t} \overset{iid}{\sim} \mathcal{N}\left(0, \sqrt{\frac{2}{5}}\right)$.[12] For each simulation and each horizon $h = 0, 1, .., 4$, I produce estimates of ATT_h, using both the SGDD and SWDD estimators. As a benchmark, I also produce estimates using the Regression Imputation estimator of BJS. For each estimand, Table 1 reports the variance of each estimator relative to the most efficient of the three estimator. Results for theoretically efficient estimators are marked in italic. Columns correspond to different variations of the simulation setup as detailed below.

The first column shows that under the benchmark assumption of Random Walk Errors, both the SGDD and SWDD estimators performs very well at short horizons. At horizon 0, where SGDD/SWDD are best unbiased estimators, the variance of the RI estimator is 64 percent larger than that of the SGDD/SWDD. At horizon 1 the variance of RI is 25-33 percent larger. The efficiency gains from using the more efficient estimator here are thus equivalent to having between 25

[12] Note that setting the standard deviation of the shocks to $\sqrt{\frac{2}{5}}$ implies that the $Var(\varepsilon_{i,t,})$ increases linearly from 0 to 2 over the six periods.

Table 1: Simulation results

	Variance relative to best shown estimator:			
	Random Walk Errors	AR(1) $\rho = 0.8$	AR(1) $\rho = 0.5$	Non-staggered, RW Errors
Subgroup DID				
Horizon 0	*1.000*	1.000	1.000	*1.000*
Horizon 1	1.060	1.047	1.075	*1.000*
Horizon 2	1.161	1.061	1.040	*1.000*
Horizon 3	1.262	1.198	1.146	
Horizon 4	1.364	1.199	1.188	
Stepwise DID				
Horizon 0	*1.000*	1.000	1.000	*1.000*
Horizon 1	*1.000*	1.000	1.092	*1.000*
Horizon 2	*1.000*	1.000	1.073	*1.000*
Horizon 3	*1.000*	1.000	1.073	
Horizon 4	*1.000*	1.000	1.000	
Regression Imputation				
Horizon 0	1.644	1.527	1.075	1.555
Horizon 1	1.328	1.223	1.000	1.290
Horizon 2	1.195	1.055	1.000	1.193
Horizon 3	1.160	1.052	1.000	
Horizon 4	1.127	1.016	1.005	

The table shows simulation results for the three estimators when estimating average treatment effects at different horizons relative to treatment (across all units observed at the relevant horizon). Columns correspond to different variations of the simulation described in the text. The first column uses Random Walk errors. The second column uses AR(1) errors with an autocorrelation coefficient of 0.8. The third column uses AR(1) errors with an autocorrelation coefficient of 0.8. The fourth column uses Random Walk errors but modifies the simulation to have non-staggered adoption with all units treated in period 4 so that treatment effects can only be estimated up to horizon 2. For each simulation and estimand, the table reports the variance of the estimator relative to the best alternative among the three estimators. Italic denotes theoretically efficient estimators.

and 64 percent more data (assuming variance is inverse proportional to sample size).

Since the simulation has non-staggered adoption, only the SWDD estimator is theoretically efficient at longer horizons. The additional results in the first column shows that the SWDD estimator indeed provides considerable efficiency gains over SGDD at longer horizons. At horizon 3 and 4, the variance of the SGDD estimator is 26-36 percent larger than the SWDD estimator. Moreover, for these longer horizons, we in fact see that the RI estimator also outperforms SGDD by having an efficiency loss of only 13-16 percent relative to SWDD. As noted at the end of Section 2.3 this likely reflects that at longer horizons the Regression Imputation estimator leverages more untreated units in the same way that SWDD does.

In practice of course, few data sets are likely to exhibit Random Walk Errors exactly. Accordingly, the second column compares the estimators under the less extreme persistence assumption of AR(1) errors with parameter $\rho = 0.8$ (and $Var(\varepsilon_{i,t,}) = 1$). The relative performance of the estimators is quite similar in this case, although - as should be expected - the differences are less stark.

The third column consider AR(1) errors with parameter $\rho = 0.5$. In a naive sense, this is halfway in-between the ideal case for SWDD ($\rho = 1$) and the ideal case for RI ($\rho = 0$). We in fact see that SWDD and RI perform quite similarly here: At horizon 0 SWDD (and also SGDD) is more efficient, at horizon 1-3 Regression Imputation more efficient, while at horizon 4, SWDD is again slightly more efficient. Again however, SGDD, has a substantial efficiency loss at longer horizons relative to the other estimators.

Finally, the fourth column returns to Random Walk Errors but considers a non-staggered simulation where all eventually treated units have $E_i = 4$ (meaning that treatment effects are only identified up to horizon 2). In this case SGDD and SWDD are theoretically equivalent and efficient at all horizons. Accordingly, we see that both provide substantial efficiency gains relative to RI.

3.2 Practical relevance of the Random Walk and Spherical Error benchmarks

The simulation results in the previous section suggest that there may be large efficiency gains from choosing estimator based on the whether the outcome variables is subject to mostly impersistent shocks (closer to spherical errors) or mostly persistent shocks (closer to random walk errors). At the same time, economic theory often give clear predictions about which error benchmark is likely

to apply to a given variable.

A natural question of course is whether such predictions are in fact borne out in the types of data typically used in difference-in-differences designs. If most outcome variables in practice exhibit (approximately) spherical errors for example, the gain from considering SGDD/SWDD estimators may be small.

As a check on the practial relevance of the two error benchmarks, Appendix B analyzes data from Brenøe *et al.* (forthcoming) (BCHH from now on). Using Danish administrative data, BCHH applies a non-staggered difference-in-differences design to estimate the causal effect on firms when one of their female employees give birth and goes on parental leave. In addition to representing my own most recent application of a difference-in-differences design, BCHH is an interesting case study because of its diverse set outcomes variables. Closely related data and research designs have also appeared frequently in the literature (see for example Jäger and Heining (2022), Bertheau *et al.* (2022) and Schmutte and Skira (2023)).

For the firm outcomes in BCHH, the measured autocorrelation in the errors ($\varepsilon_{i,t}$) ranges from 0.085 to 0.995 depending on the outcome variable and method. This confirms 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 outcome variable closest to the random walk benchmark is the firm's total wage bill. This is in line with theories of frictional labor adjustment and sticky wages. The outcome variable closest to the spherical error benchmark is the total births among all firm employees. This is in line with individual fertility being driven mostly by idiosyncratic and transitory shocks.

Finally, applying RI and SGDD/SWDD estimators to the data from BCHH and comparing the estimated (cluster-robust) standard errors also suggests considerable efficiency gains from estimator choice. For a range of outcome variables, the estimated gain in precision from choosing the right estimator is equivalent to as much as 50 percent more data.

4 Conclusion

When analyzing difference-in-differences designs, applied researchers face a choice between several estimators that require the same assumptions for unbiasedness. This paper provides a set of new

efficiency results under persistent errors which - together with previous results in Borusyak *et al.* (forthcoming) - allow applied researchers to make a principled estimator choice aimed at improving precision:

If the outcome variable is likely to be characterized by mostly impersistent errors (e.g. transitory shocks or measurement errors), Regression Imputation estimators are likely to perform well for estimating any treatment effect. If the outcome variable is instead likely to be characterized by very persistent errors (persistent shocks), Subgroup Difference-in-Difference estimators in the vein of de Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2021) should perform well if estimating treatment effects immediately at treatment onset, or if treatment adoption is non-staggered. If researchers are also interested in later time horizons and treatment adoption is staggered however, the novel corrected Stepwise Difference-in-Differences (SWDD) estimator provides additional efficiency gains under persistent errors. A Stata package implementing SWDD is available on my website (`did_stepwise.ado`). Both simulation evidence and application to existing data suggest large efficiency gains from choosing the right estimator for the particular outcome under study.

In terms of future work, I note particularly that the stepwise adjustment underlying SWDD seem likely to both be manageable to implement and to offer potential efficiency gains also outside the canonical non-staggered difference-in-difference setting; Online Appendix C for example provides a simple extension to the case of conditioning on predetermined covariates. Exploring and incorporating the stepwise adjustment for existing extensions and estimator thus seem like a fruitful avenue for future work.

References

Bellégo, C., Benatia, D. and Dortet-Bernardet, V. (2023) The chained difference-in-differences, *Working Paper*.

Bertheau, A., Cahuc, P., Jäger, S. and Vejlin, R. (2022) Turnover costs: Evidence from unexpected worker separations, *Working Paper*.

Borusyak, K., Jaravel, X. and Spiess, J. (forthcoming) Revisiting event study designs: Robust and efficient estimation, *Review of Economic Studies*.

Brenøe, A. A., Canaan, S., Harmon, N. A. and Royer, H. (forthcoming) Is parental leave costly for firms and coworkers?, *Journal of Labor Economics*.

Callaway, B. and Sant'Anna, P. H. (2021) Difference-in-differences with multiple time periods, *Journal of Econometrics*, **225**, 200–230, themed Issue: Treatment Effect 1.

de Chaisemartin, C. and D'Haultfœuille, X. (2020) Two-way fixed effects estimators with heterogeneous treatment effects, *American Economic Review*, **110**, 2964–96.

de Chaisemartin, C. and D'Haultfœuille, X. (2022) Difference-in-differences estimators of intertemporal treatment effects, *Working Paper*.

Dube, A., Girardi, D., OscarJorda and Taylor, A. M. (2022) A local projections approach to difference-in-differences event studies, *Working Paper*.

Gardner, J. (2022) Two-stage differences in differences.

Jäger, S. and Heining, J. (2022) How substitutable are workers? evidence from worker deaths, *Working Paper*.

Marcus, M. and Sant'Anna, P. H. C. (2021) The role of parallel trends in event study settings: An application to environmental economics, *Journal of the Association of Environmental and Resource Economists*, **8**, 235–275.

Nickell, S. (1981) Biases in dynamic models with fixed effects, *Econometrica*, **49**, 1417–1426.

Roth, J. (2022) Pretest with caution: Event-study estimates after testing for parallel trends, *American Economic Review: Insights*, **4**, 305–22.

Roth, J., Sant'Anna, P. H., Bilinski, A. and Poe, J. (2023) What's trending in difference-in-differences? a synthesis of the recent econometrics literature, *Journal of Econometrics*, **235**, 2218–2244.

Schmutte, I. M. and Skira, M. M. (2023) The response of firms to maternity leave and sickness absence, *Journal of Human Resources*.

Sun, L. and Abraham, S. (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects, *Journal of Econometrics*, **225**, 175–199, themed Issue: Treatment Effect 1.

Wooldridge, J. M. (2021) Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators, *Working Paper*.

A Proof of Theorems 1 and 2

A.1 Proof of Theorem 1

Theorem 1 in fact follows as a direct consequence from Theorem 2 by noting that the SGDD and SWDD estimators are always equivalent when the data is a *Balanced Panel* and there is *Non-Staggered Adoption.*

A.2 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.* (forthcoming) 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.

A.2.1 The efficient estimator can be viewed as an OLS estimator from a first-differenced regression equation

As noted 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 Online Appendix E.4for proof in a more general case):

$$Y_{i,t} = \alpha_i + \beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{8}$$

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 8 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):

$$Y_{i,t} = \alpha_i + \beta_1 + \sum_{j=1}^{T} \mathbf{1}[t > j] \Delta\beta_j + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{9}$$

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 9, 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:

$$\Delta Y_{i,t} = \Delta\beta_t + \sum_{k=0}^{\bar{K}} \Delta H_{i,t}^k \gamma_{i,k} + \Delta\varepsilon_{i,t} \tag{10}$$

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 Online Appendix E.5 for a detailed argument in a more general case).

A.2.2 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 10. 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 10 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 is 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:

$$\Delta Y_{i,t} = \Delta\beta_t + H_{i,t}^0 \gamma_{i,0} + \sum_{k=1}^{\bar{K}} H_{i,t}^k \Delta\gamma_{i,k} + \Delta\varepsilon_{i,t} \tag{11}$$

Now consider applying OLS to 11. 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, note that we can of course recover the OLS estimators of interest simply by reversing the definition of $\Delta\gamma_{i,k}$:

$$\hat{\gamma}_{i,h}^{OLS} = \hat{\gamma}_{i,0}^{OLS} + \sum_{k=1}^{h} \widehat{\Delta^1\gamma_{i,k}}^{OLS} \tag{12}$$

Additionally, the regression equation 11 is of the same form as the ones considered by BJS.[13]

[13]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 11 using only untreated observations. That is, estimate the following equation using only observations $(i,t) \in \Omega^{D,1}$ such that $D_{i,t} = 0$:

$$\Delta Y_{i,t} = \Delta\beta_t + \Delta\varepsilon_{i,t} \tag{13}$$

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:

$$\hat{\gamma}_{i,0}^{OLS} = \Delta Y_{i,E_i} - \widehat{\Delta Y_{i,E_i}} \tag{14}$$

$$\widehat{\Delta\gamma_{i,k}}^{OLS} = \Delta Y_{i,E_i+k} - \widehat{\Delta Y_{i,E_i+k}} \tag{15}$$

4. Obtain the OLS estimator $\hat{\gamma}_{i,h}^{OLS}$ for $h > 0$ by applying 12.

A.2.3 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 13 is simple to characterize here as it will simply equal the OLS estimate of $\Delta\beta_t$ from 13 (note that this coefficient is guaranteed to not have been affected by the normalization applied earlier). This estimate is simply:

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

Plugging into 14 and 15 we then get:

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

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

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

Plugging into 12 then completes the proof by showing that $\hat{\gamma}_{i,h}^{OLS} = \hat{\gamma}_{i,h}^{SWDD}$.

B Evidence on practical relevance using Brenøe *et al.* (forthcoming)

In this section, I provide evidence on the practical relevance of the two error benchmarks and the extent of possible efficiency gains from estimator choice. To do this I analyze the Danish administrative data from Brenøe *et al.* (forthcoming) (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)

After modifying it to fit the theoretical framework above,[14] 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

[14]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 results presented 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 in Online Appendix F 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.

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.[15]

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 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 errors persistence in the different outcome variable. Using all 13 years of available data for untreated units, the second column fits a twoway fixed effecs model and computes the empirical autocorrelation of the residuals. With a limited number of time periods, this residual autocorrelation is known to understate the true autocorrelation of the errors so the third column applies the Nickell (1981)-correction which recovers the true autocorrelation under the assumption of AR(1) errors.[16] As expected, we see a large spread in the degree of error persistence across the difference outcomes. At the lower end total births appears quite close to the serially uncorrelated benchmark with autocorrelations of 0.087 and 0.187 in the two columns. At the higher end the total firm wage bill appears close to the random walk benchmark with estimated autocorrelations of 0.766 and 0.995. This underscores the empirical relevance of both the spherical errors and random walk benchmarks.

Finally 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

[15] I implement the RI estimator via the `did_imputation` Stata package and implement SGDD/SWDD via my own `did_stepwise` package which relies `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)

[16] More precisely, the empirical autocorrelation of the residuals is consistent only when the number of time periods goes to infinity. Under the assumption that errors are AR(1), the Nickell (1981)-correcton is consistent when the number of units goes to infinity.

and numerical results provided earlier. For outcomes with impersistent errors, RI estimators have a 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 .

Table 2: Comparing performance on data from Brenøe *et al.* (forthcoming)

	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.* (forthcoming), 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).

ONLINE APPENDICES

C 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.

C.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:

$$Y_{i,t} = \alpha_i + X_i'\beta_t + \sum_{k=0}^{\bar{K}} H^k_{i,t}\gamma_{i,k} + \varepsilon_{i,t} \tag{16}$$

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 assumption as in the main text (see Section C.2 for the proof):

Theorem 3. *Assume that the data satisfy 16 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*

$$\hat{\gamma}_{i,h}^{SWDD} = \sum_{k=0}^{h} \left[(Y_{i,E_i+k} - Y_{i,E_i+k-1}) - \frac{\mathbf{1}}{\mathbf{N}} \sum_{\substack{j: K_{j,E_i+k-1}<0, \\ K_{j,E_i+k}<0}} \kappa_{i,t,j} \left(Y_{j,E_i+k} - Y_{j,E_i+k-1}\right) \right] \tag{17}$$

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

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

Comparing 17 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 17 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 17 above, the SWDD 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).[17]

C.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

$$\hat{\gamma}_{i,h}^{OLS} = \sum_{k=0}^{h} \left[\Delta Y_{i,t} - \underset{\substack{j: K_{j,E_i+k-1}<0, \\ K_{j,E_i+k}<0}}{\frac{\mathbf{1}}{\mathbf{N}} \sum} \widehat{\Delta Y_{i,t}} \right] \tag{18}$$

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:

[17] For example, let $\omega_{i,t}$ be a set of weight satisfying $\omega_{i,t} = 1$ whenever $D_{i,t} = 1$ and assume that parallel trends hold only after applying these weights (in the sense that $E\left[\omega_{i,t}\left(Y_{i,t}^0 - Y_{i,t-1}^0\right)\right]$ is constant across units). If there is also *No Anticipation* then applying 17 with $\kappa_{i,t,j} = \omega_{i,t}$ yields an unbiased estimator.

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

$$= \frac{1}{N} \sum_{\substack{j:\, K_{j,t}<0, \\ K_{j,t-1}<0}} \left(X_i' \left(\frac{1}{N} \sum_{\substack{j:\, K_{j,t}<0, \\ K_{j,t-1}<0}} X_j X_j' \right)^{-1} X_j \right) \Delta Y_{j,t}$$

Plugging for $\widehat{\Delta Y_{i,t}}$ in 18 then completes the proof.

C.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:

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

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

D Efficiency of SGDD with staggered adoption, additional special cases

In addition to the case of non-staggered adoption, there are two additional cases of interest in which the SGDD estimator is equivalent to the SWDD estimator and is thus efficient. The first case isif the data includes no additional preperiods for eventually treated units but that there is otherwise no missing data. This occurs for example if the identifying assumption for eventually treated units are only assumed to hold from the baseline period and onwards, so that earlier preperiods are excluded (corresponding to the 'never treated' approach discussed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021)). :

Corollary 2. *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 second case 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 3. *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, $\widehat{\gamma_h^w}^{SGDD}$.*

E 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 parallelt 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.

E.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 looking h time periods ahead:

Assumption 9. *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 *Full 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.

E.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_h^w, 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 E.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 E.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 treatmet effect, γ_h^w, is the SGDD estimator, $\widehat{\gamma_h^w}^{DID}$.*

E.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 E.4 and E.5 for clarity.

E.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 E.4 for a full derivation):

$$Y_{i,t} = \alpha_{i,c(t)} + \beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \tag{19}$$

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 19 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):

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

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 19, 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 19 means that for all $(i,t) \in \Omega^{D,h+1}$ the following regression equation holds:

$$\Delta^{h+1} Y_{i,t} = \Delta^{h+1}\beta_t + \sum_{k=0}^{\bar{K}} \Delta^{h+1} H^k_{i,t} \gamma_{i,k} + \Delta^{h+1}\varepsilon_{i,t} \tag{20}$$

Under the normalization chosen above, the non-normalized coefficient vector θ is identified in this differenced regression and applying OLS will yield 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 $\hat{\gamma}^{OLS}_{i,h}$ denote the corresponding estimator of $\gamma_{i,h}$.

We next show that this OLS estimator is efficient. Under *Random Walk Errors*, the regression

equation 20 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 19 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 E.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.

E.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 20. 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 20 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:

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

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

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:

$$\Delta^{h+1}Y_{i,t} = \Delta^{h+1}\beta_t + \sum_{k=0}^{h} H_{i,t}^k \gamma_{i,k} + \sum_{k=h+1}^{\bar{K}} H_{i,t}^k \Delta^{h+1}\gamma_{i,k} + \Delta^{h+1}\varepsilon_{i,t} \tag{21}$$

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 steps:

1. Estimate equation 21 using only untreated observations. That is, estimate the following equation using only observations $(i,t) \in \Omega^{D,h+1}$ such that $D_{i,t} = 0$:

$$\Delta^{h+1}Y_{i,t} = \Delta^{h+1}\beta_t + \Delta^{h+1}\varepsilon_{i,t} \tag{22}$$

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:

$$\hat{\gamma}_{i,h}^{OLS} = \Delta^{h+1}Y_{i,E_i+h} - \widehat{\Delta^{h+1}Y_{i,E_i+h}} \tag{23}$$

[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

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

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

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

E.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 22 (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:

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

.

Plugging into 23 we get:

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

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

E.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 19 from Appendix E.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 19:

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 19.

E.5 Proof that unbiased estimators depend only on differences

Consider the version of equation 19 from Appendix E.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 19 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 $E[\varepsilon] = 0$, the expecation of the estimator can be written:

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

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

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

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

F Alternative application to Brenøe *et al.* (forthcoming)

The main specification in Brenøe *et al.* (forthcoming) 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.* (forthcoming) address by using standard errors clustered on firm. In the analysis in the main Appendix, I modified the data to fit the theoretical setup of the paper. Here I instead show results mainting the same sample and weigthing and using standard errors clustered on firm. 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 shows the result. Comparing to results in the main text we see that the relative performance of the estimators is virtually identical.

Table 3: Comparing performance on data from Brenøe *et al.* (forthcoming)

	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.* (forthcoming), 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).