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

Nikolaj A. Harmon

University of Copenhagen

First version: November 2022

This version: December 2024

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 an adjusted 'Stepwise Difference-in-Differences' estimator is the best unbiased estimator for all later effects. The results provide a simple guide to estimator choice in practice.

*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 Econometric Society European Winter Meetings 2023. 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 can lead to problems if the preferred estimator is not specified in advance.[1] In this paper, I derive new theoretical results that help applied researchers choose between the most common estimators in a simple but principled way.

For the canonical difference-in-differences design with an absorbing treatment and no covariates, two broad groups of estimators have become popular and have been implemented in commonly used statistical software:[2] 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.* (2023)). 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.* (2024); Gardner (2022); see also Wooldridge (2021)).

In choosing between SGDD and RI estimators, efficiency considerations should play an important role. In a given analysis sample, both groups of estimators are unbiased under the same assumptions so picking the efficient alternative means getting more precise estimates using the same data and assumptions. Based on existing results, however, efficiency comparisons between SGDD and RI are lopsided and incomplete. While Borusyak *et al.* (2024) (BJS from now on) has established that RI is the best unbiased estimator under spherical errors, the efficiency properties of SGDD estimators are largely unknown. A particular concern is that SGDD estimators may have generally poor efficiency

[1] A pertinent example is the case where a researcher produces estimates and confidence intervals using two or more different estimators and then ex post has to decide whether a given parameter value is rejected by the data. Rejecting only if the value falls outside both confidence intervals will lead to a lower than expected rate of false positives and imply an unnecesary loss of power. Rejecting when the parameter falls outside just one of the confidence intervals will inflate the rate of false positives.

[2] Examples of STATA/R packages that implement SGDD estimators are `did_multiplegt`, `csdid`, `eventstudyplot` and `lpdid`. Examples of packages that implement RI estimators are `did_imputation` and `did2s`.

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 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. Many microeconomic processes are in fact modelled as being subject to permanent or strongly persistent shocks so efficiency properties under this benchmark should be empirically relevant.

I first consider the case where treatment adoption is non-staggered, e.g. all eventually treated units get treated in the same period. In this case, 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 difference-in-differences estimators. A simple simulation study adapted from BJS show that the efficiency gains of SGDD can be substantial when errors are persistent: Relative to RI estimators, SGDD provides efficiency gains that are equivalent to as much as 60 percent more data (21 percent smaller standard errors) when errors follow a random walk .

Next I consider the general case of staggered treatment adoption. 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 an adjusted 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 estimates treatment effects step-by-step in a series of one-period-ahead comparisons. I clarify how this stepwise estimation leads to efficiency gains: under staggered treatment adoption, the SWDD estimator is able to leverage data on more untreated units at longer horizons. A simple simulation shows that the efficiency gains can be substantial:

Relative to SGDD, using SWDD under staggered adoption provides efficiency gains equivalent to as much as 35 percent more data (14 percent smaller standard errors) when estimating effects 4 periods after treatment onset.

Combined with previous work, these results provide 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 - RI 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. Reanalyzing data from the difference-in-differences design in Brenøe *et al.* (2024) I confirm the practical relevance of this approach. The behavior of the outcome variables in Brenøe *et al.* (2024) range from nearly no serial correlation in errors to near-random walk errors. Choosing the right estimator for each outcome delivers reductions in the width of confidence intervals equivalent to as much as 50 percent more data (18 percent smaller standard errors).

In addition to the direct implications for estimator choice in practice, this paper also makes some technical contributions, particularly clarifying the relationship between estimators and underscoring the versatilify of the methodology developed in BJS. In addition to relying on the BJS framework, the proof techniques in this paper imply that both SGDD estimators and the adjusted SWDD estimator can be viewed as an efficient RI estimator based on a particular linear regression model. This extends the tools BJS provide for RI estimators to also apply for 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).

The efficient SWDD estimator derived here also relates to other previous work. In their seminal discussion of efficiency in difference-in-differences designs Marcus and Sant'Anna (2021) proposes an alternative '$ny+$'-estimator which they conjecture to have attractive efficiency properties. For the same data and estimand, this estimator turns out to be equivalent to the SWDD estimator. This paper thus establishes formal efficiency properties for the '$ny+$'-estimator. Since the original circulation of the present paper, Bellégo *et al.* (2024) has also proposed a class of 'Chained Difference-in-Differences' estimators to address complications arising when balanced panels are not available.

These estimators are also closely related to SWDD and the efficiency results in this paper apply to them as well.

The focus of this paper is to help researchers choose between simple, popular difference-in-differences estimators based on finite sample efficiency results under two error benchmarks. In data sets where neither error benchmark is a good approximation however, more complex (multi-step) estimators may provide additional efficiency gains, at least asymptotically in large samples. Earlier work by Marcus and Sant'Anna (2021) provide a framework for asympotically efficient GMM. Since the original circulation of this paper, BJS has described a natural Feasible GLS estimator which is asymptotically efficient and bridges their RI efficiency results with the ones presented here. Additionally, Arkhangelsky *et al.* (2021) and Clarke *et al.* (2023) have derived and implemented a Synthetic Difference-in-Differences estimator which may also provide efficiency gains under some error structures.

Finally, an alternative approach to estimator choice is to consider the extent of bias in 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 BJS). The online appendix for this paper provides additional results in this direction by establishing robustness of the SGDD estimator to a certain violation of parallel trends.

1 Framework and assumptions

Except for notation, 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, $i = 1, 2, ..., N$, observed over several periods, $t = 1, 2, ..., T$. 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. For ease of exposition, I assume that the data is a balanced panel throughout the main text. The online appendix provides a simple extension of the main theorem to unbalanced panels.

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. Units that

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

are never treated correspond to $E_i = \infty$. This nests both staggered treatment adoption and the non-staggered case where treatment happens at single point in time. As noted, the analysis will treat the observed data as containing a fixed set of units with a given treatment timing so $D_{i,t}$ and E_i are non-stochastic.

For expositional convenience, I also define some additional variables and notation. I let $\bar{K}_i = T - E_i$ denote the number of additional post-treatment periods where i is observed and let $\bar{K} = \max_i \bar{K}_i$ be the maximum number of such post-treatment periods observed for any unit in the data.

Averages over units that satisfy certain conditions will play a prominent role throughout the text. 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] As an example, the expression below corresponds to the the average period t outcome for units who are untreated at time $t + k$:

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

1.1 Potential outcomes, treatment effects and estimands

Treatment effects are defined relative to a situation where units never experience the treatment. Accordingly, $Y^0_{i,t}$ 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 at different time horizons relative to the onset of treatment. I let $\gamma_{i,h} = E\left[Y_{i,E_i+h} - Y^0_{i,E_i+h}\right]$ denote unit i's treatment effect, at the time when they have experienced the treatment for h previous periods. I refer to this as the horizon h treatment effect for i. Treatment effects at horizon $h = 0$ corresponds to the contemporaneous effect at the onset of treatment.

With these building blocks, I will consider the case where the estimand of interest is some weighted sum of treatment effects at a specific horizon h: $\gamma^w_h = \sum_{i:\bar{K}_i \geq h} w_i \gamma_{i,h}$ 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^w_h is the conditional average horizon h treatment effect for units first treated at time t, written as $CATT_{t,h} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{i:K_{i,t+h}=h} \gamma_{i,h}$.

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

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.

For most theoretical results however, I allow for the more general case where a researchers may also be interested in averaging treatment effects across different horizons. This corresponds to the general 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

$$\hat{\gamma}_{i,h}^{SGDD} = (Y_{i,E_i+h} - Y_{i,E_i-1}) - \frac{\mathbf{1}}{\mathbf{N}} \sum_{j:\, D_{j,E_{i+h}=0}} (Y_{j,E_i+h} - Y_{j,E_i-1}) \tag{3}$$

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

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

off the average corresponding change among an appropriate subgroup of 'relevant controls'. These are units that remained untreated in the data over the entire period from $E_i - 1$ to $E_i + h$. With an absorbing treatment and a balanced panel such 'relevant controls' are characterized succintly by having $D_{j,E_i+h} = 0$. Finally, to arrive at an estimator for the weighted sum of treatment effects, the relevant weights are simply applied to the individual-level estimates as shown in (1) and (2).

For the same estimand, the SGDD estimator defined above is identical to the estimator of de Chaisemartin and D'Haultfœuille (2020), the 'not-yet-treated' estimator of Callaway and Sant'Anna (2021) and the LP-DID estimator of Dube *et al.* (2023). In particular, note that if we set the estimand γ_h^w to be $CATT_{t,h}$, we arrive at the familiar expression:

$$\widehat{CATT_{t,h}}^{DID} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{i:\, E_i = t} (Y_{i,t+h} - Y_{i,t-1}) - \frac{\mathbf{1}}{\mathbf{N}} \sum_{j:\, D_{j,t+h}=0} (Y_{j,t+h} - Y_{j,t-1})$$

Other modern difference-in-differences estimators fall within the class of SGDD estimators if one extends the definition to allow for unbalanced panels. This includes the 'never-treated' estimators of Callaway and Sant'Anna (2021) and Sun and Abraham (2021). Formally, these estimators can be viewed as dropping certain observations and then imposing identifying assumptions and forming SGDD estimators *only* on the resulting unbalanced panel. As I return to in Section 2.3 and the online appendix, the main theorem thus extends to these 'never-treated' estimators as well.

1.3 Restrictions on treatment timing

Throughout the analysis, I will require that the data contains sufficient control units so that SGDD estimators are well-defined. In a balanced panel, necessary and sufficient conditions for this are that there is some untreated unit in the last period and that no units start out already treated:

Assumption 1. *Sufficient Control Units: There exists some unit i such that $D_{i,T} = 0$. Moreover, for all units i we have $D_{i,1} = 0$.*

Conditional on considering SGDD estimators this assumption is innocous. If it fails there is some treated unit and post-period where no relevant untreated comparison units exist. Such units/periods would always have to be dropped to apply SGDD.

For some of the results later, I will additionally consider the more restricted case where treatment

adoption is non-staggered so that all units that are eventually treated receive the treatment at the same time:

Assumption 2. *Non-Staggered Adoption: For any pair of units (i, j) such that $E_i, E_j \neq \infty$, we have $E_i = E_j$.*

1.4 Identifying assumptions

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

Assumption 3. *No Anticipation: $Y_{i,t} = Y_{i,t}^0$ whenever $D_{i,t} = 0$.*

As written this assumption imposes that eventual treatment does not affect outcomes in periods before treatment occurs. As is well known however, a simple relabeling of the treatment variable covers cases where outcomes are affected some known number of periods before treatment occurs.

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

Assumption 4. *Parallel Trends: For any two periods t and t', $E\left[Y_{i,t}^0 - Y_{i,t'}^0\right]$ is constant across i.*

This version of the parallel trends assumption is identical to the one in BJS using the same framework, and equivalent to the assumption in de Chaisemartin and D'Haultfœuille (2020) and Dube *et al.* (2023) when translated to the present (finite sample) framework. The assumption also corresponds to the 'not-yet-treated' approach of Callaway and Sant'Anna (2021) modulo a small restriction on what constitues the first time period.[6] As noted previously, the main results later also apply also to a range of alternative approaches which restrict estimation to a subset of observations and impose parallel trends *only* on this subset (see Section 2.3).

1.5 Error benchmark

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 shows that RI is the best unbiased estimator when these errors

[6]The 'not-yet' treated approach of Callaway and Sant'Anna (2021) only uses observations - and only imposes parallel trends - starting the period before the first unit gets treated. This corresponds restricting the data and relabeling the time variable so that $\min_i E_i = 2$.

are spherical, that is homeskedastic and serially uncorrelated. Spherical errors is a 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 errors corresponds to the assumption 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, i.e. that the *differences* in the errors are homoskedastic and serially uncorrelated. Formally, I define $\eta_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$ to be the shock to unit i at time t and let η denote the correpsonding NT-dimensional vector of all shocks. I then consider the following assumption:

Assumption 5. *Random Walk Errors: The shocks η are mean zero, homoskedastic and uncorrelated over time and units: $E(\eta) = 0$, $Var(\eta) = \mathbb{I}_{NT}\sigma^2$.*

It is worth emphasizing that I use this assumption only as a benchmark to characterize *when* the different estimators have attractive efficiency properties. Neither the SGDD estimator, the RI estimator or the SWDD 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 most settings.

2 Theoretical efficiency results

I now present my results on efficient estimation along with simple proofs and discussion.[7] As usual when studying unbiased estimators, I use the term 'best estimator' to refer to the estimator with the lowest possible variance.

2.1 Efficiency under non-staggered adoption

The first result establishes that the commonly-used SGDD estimators are efficient when treatment adoption is non-staggered and errors are strongly persistent:

[7] The online appendix provides expanded proofs also covering the case of unbalanced panels.

Proposition 1. *Assume that there is Sufficent Control Units, No Anticipation, Parallel Trends, Random Walk Errors, and Non-Staggered Adoption. Then the best unbiased estimator of any treatment effect γ^w is the Subgroup Difference-in-Differences Estimator, $\widehat{\gamma^w}^{SGDD}$.*

Proof. The proposition follows as a special case of Theorem 1 below.

Proposition 1 establishes that the SGDD estimator utilizes the data efficiently (formally it establishes 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 SGDD involves simple averages across units.

It might be more surprising that the SGDD estimator also exploits 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 preperiods available which could also be used for estimation.

Formally, the efficiency of SGDD turns out to reflect its analogy with first-difference estimators for panel data (see the proof of Theorem 1). To build direct intution for this efficiency property, however, consider estimating the horizon h treatment effect for some specific unit i that gets treated at E_i. For ease of exposition, additionally assume that at $E_i + h$ there is only a *single* unit j that remains untreated. Using the units i and j, there are now many different difference-in-differences comparison we could consider in estimation: For *some arbitrary* baseline period $b < E_i$, let $\hat{\gamma}^{ALT}_{i,h}$ be the comparison that goes from period b to period $E_i + h$:

$$\hat{\gamma}^{ALT}_{i,h} = (Y_{i,E_i+h} - Y_{i,b}) - (Y_{j,E_i+h} - Y_{j,b})$$

Under *Parallel Trends* and *No Anticipation* this can be rewritten as:

$$\hat{\gamma}^{ALT}_{i,h} = \gamma_{i,h} + \sum_{k=b+1}^{E_i+h} (\eta_{i,t} - \eta_{j,t}) \tag{4}$$

Equation (4) shows that the difference-in-difference comparison $\hat{\gamma}^{ALT}_{i,h}$ is an unbiased estimator; it equals the treatment effect of interest, $\gamma_{i,h}$, plus a (mean zero) noise term. This noise term is a sum over differences in the idiosyncratic shocks $\eta_{i,t}$ and $\eta_{j,t}$ that the units experience each period

from the baseline period b and until the postperiod $E_i + h$. Under *Random Walk Errors*, however, these shocks are serially uncorrelated. This yields two conclusions:

First, to get the smallest variance in our difference-in-difference comparision, $\hat{\gamma}_{i,h}^{ALT}$, we should choose the latest possible preperiod as the baseline ($b = E_i - 1$) because this implies that the noise term includes as few shocks as possible. This choice of baseline period, however, means that $\hat{\gamma}_{i,h}^{ALT}$ is equal to the SGDD estimator, $\hat{\gamma}_{i,h}^{SGDD}$.

Second, if we consider difference-in-difference comparisons that use some earlier preperiod as the baseline, $b < E_i - 1$, equation (4) shows that the difference between $\hat{\gamma}_{i,h}^{ALT}$ and $\hat{\gamma}_{i,h}^{SGDD}$ is simply that the noise term for $\hat{\gamma}_{i,h}^{ALT}$ will include *more* serially uncorrelated shocks. These additional shocks will only add additional variance and lower precision.

Summing up, the example illustrates that the SGDD estimator is efficient under strongly persistent errors exactly *because* it only uses the last period before treatment onset. Considering earlier preperiods only adds additional noise from earlier shocks. The arguments and intuition goes through unchanged if we consider the general case of having more untreated units or estimating weighted averages of individual treatment effects. Note that the arguments above also provide a rigorous justification for the general practice of using the last period prior to treatment as the baseline in difference-in-differences.

2.2 Efficiency with staggered adoption

Next, I turn to the more general case of staggered adoption where units may enter treatment at different times. The efficient estimator in this case turns out to be an adjusted difference-in-differences estimator that I term 'Stepwise Difference-in-Differences' (SWDD):

Theorem 1. *Assume that there is Sufficent Control Units, No Anticipation, Parallel Trends and 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:*

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

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

$$\hat{\gamma}_{i,h}^{SWDD} = \sum_{k=0}^{h} \left[(Y_{i,E_i+k} - Y_{i,E_i+k-1}) - \underset{j:\, D_{j,E_i+k}=0}{\frac{\mathbf{1}}{\mathbf{N}} \sum} (Y_{j,E_i+k} - Y_{j,E_i+k-1}) \right] \tag{5}$$

.

Proof. For $k = 0, 1, ..., \bar{K}$, let $H_{i,t}^k$ be a dummy for whether at time t, unit i is treated and has experienced the treatment for exactly k previous periods:

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

The assumptions of *No Anticipation* and *Parallel Trends* then implies that the data satisfies the following model (see for example BJS):

$$Y_{i,t} = \alpha_i + \beta_t + \sum_{k=0}^{\bar{K}} H_{i,t}^k \gamma_{i,k} + \varepsilon_{i,t} \qquad , \qquad E\left[\varepsilon_{i,t}\right] = 0 \tag{6}$$

Mirroring a well-known result regarding efficient estimation in panels with random walk errors, the proof proceeds by considering the first-differenced version:

$$\Delta Y_{i,t} = \Delta\beta_t + \sum_{k=0}^{\bar{K}} \Delta H_{i,t}^k \gamma_{i,k} + \Delta\varepsilon_{i,t} \qquad , \qquad E\left[\Delta\varepsilon_{i,t}\right] = 0 \tag{7}$$

Under the assumption of *Sufficient Control Units*, any individual treatment effect of interest, $\gamma_{i,h}$, is identified in (7) and applying OLS will yield the unbiased estimator $\hat{\gamma}_{i,h}^{OLS}$. Under *Random Walk Errors* a standard application of the Gauss-Markov Theorem to panel data implies that these these OLS estimators are the best unbiased estimators of the individual treatment effects $\{\gamma_{i,h}\}_{i,h:\bar{K_i} \geq h}$. Moreover, taking their linear combination $\widehat{\gamma^w}^{OLS} = \sum_{i,h:\bar{K_i} \geq h} w_{i,h} \hat{\gamma}_{i,h}^{OLS}$ will yield the best unbiased estimator of γ^w.

Finally, simple algebra shows that the Regression Imputation Theorem of BLS can be applied in (7) to characterize the closed form of the efficient estimator $\widehat{\gamma^w}^{OLS}$. This turns out to equal the SWDD estimator.

□

To understand why Stepwise Difference-in-Differences is an appropriate name for this efficient estimator, consider the expression for the individual-level estimator, $\hat{\gamma}_{i,h}^{SWDD}$ in (5). The expression inside brackets is simply a one-period-ahead difference-in-differences comparison involving a one period change in the outcome for unit i and the corresponding average change among untreated units. Taking into account the outer sum shows that the SWDD estimator is simply a sum over $h+1$ such one period difference-in-differences. In other words, where the 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 can be seen as working step-by-step: It first constructs a series of one-period-ahead comparions and then sums these up to arrive at the final estimate.

To better understand how and why the SWDD estimator differs from SGDD, first rewrite the individual-level SWDD estimator as:

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

Then note that the first sum over k immediately telescopes:

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

The first term in this expression is identical to the first term in the expression for the individual-level SGDD estimator and is simply the total change in the outcome of the treated unit i between $E_i - 1$ and $E_i + h$. The difference between the SWDD and the SGDD estimators thus come entirely from the second term, which relates to the untreated units.

Now consider what happens *if* no unit switches from treated to untreated over the period E_i to $E_i + h$. In this case the set of untreated units is unchanged over these periods so the second sum over k also telescopes and will just equal the average total change in the outcome among these units:

$$\sum_{k=0}^{h} \left(\frac{\mathbf{1}}{\mathbf{N}} \sum_{j:\, D_{j,E_i+k}=0} (Y_{j,E_i+k} - Y_{j,E_i+k-1}) \right) = \frac{\mathbf{1}}{\mathbf{N}} \sum_{j:\, D_{j,E_i+h}=0} (Y_{j,E_i+h} - Y_{j,E_i-1})$$

This is the second term in the expression for the SGDD estimator. When no unit switches from treated to untreated over the period E_i to E_i+h the SWDD estimator and SGDD estimator are thus equivalent. This shows the connection between Proposition 1 and Theorem 1: under non-staggered adoption if some unit is treated at E_i, all units that are untreated at E_i remain untreated forever. SGDD is therefore always equivalent to the efficient SWDD estimator in this case.

In general with staggered adoption however there may be some units that are untreated at E_i but have switched into treatment by period $E_i + h$. Such units give rise to the efficiency gains of SWDD over SGDD: Because these units are not observed as untreated at *both* $E_i - 1$ and $E_i + h$, these units are never included as controls in the SGDD estimator. In constrast, they *will* be included in some of the one-period comparisons used in the SWDD estimator. Under *Random Walk Errors,* leveraging such additional untreated units is guaranteed to improve efficiency. See Section D of the online appendix for a simple visual illustration of the above.

An implication of the discussion above is that the efficiency gains of SWDD over SGDD will be larger when estimating effects at longer horizons. When looking at a longer horizon h, there will mechanically be more units who switch from treated to untreated over the periods from E_i to E_i+h and therefore drop out the SGDD estimator. Conversely, in the extreme case where a researcher is interested only in the contemporaneous treatment effect at horizon $h = 0$, no such units drop out of the SGDD estimator and the SWDD and SGDD estimators always coincide. This immediately yields the following corollary:

Corollary 1. *Assume that there are Sufficent Control Units, No Anticipation, Parallel Trends and 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 staggered adoption.[8]

2.3 Additional results and discussion

I close this section with two additional remarks regarding the theoretical results. First, the online appendix extends Theorem 1 to cover various forms of unbalanced panels. In addition to addressing

[8] A related case where SGDD can also remains efficient is if the periods where treatment adoption happens are sufficiently spaced out over time. See Section C of the online appendix for details.

possible missing data, this also extends the theorem to cases where researchers exclude certain observations in order to use a weaker version of parallel trends: since the identifying assumptions only need to hold across observations included in estimation, dropping observations means that a weaker version of parallel trends is required for unbiasedness. The general version of Theorem 1 imply that SWDD is the best unbiased estimator in these cases as well. A particularly notable case is the 'never-treated' approach of Callaway and Sant'Anna (2021)and Sun and Abraham (2021) which effectively restricts the data by dropping all but the last pretreatment observation for each treated unit. Under this data restriction however, SGDD turns out to always be equivalent to the efficient SWDD estimator. In the 'never-treated' approach, SGDD is thus efficient with persistent errors even when adoption is staggered.

Second, in addition to their practical implications, the results above help clarify the relationship between estimators. The last step in the proof of Theorem 1 shows that the SWDD estimator can be viewed as an efficient RI estimator for a first-differenced regression model. Moreover, the online appendix extends this to show that any SGDD estimator of treatment effects at a particular horizon can be seen as an efficient RI estimator from a particular regression model. As a result, the tools provided by BJS for RI estimators also apply to SGDD and SWDD estimators, including their approaches to computation and (cluster-robust) inference. A Stata package implementing SWDD estimation and inference in this way is available on my website (`did_stepwise.ado`). The package also implements extensions of the SWDD estimator to cover estimation with predetermined covariates (as in BJS), and to examine the identifying assumption by estimating pretrends.[9]

3 Numerical results

To provide simulation evidence on efficiency, I add persistent errors into a simulation originally introduced by BJS. Using the notation from Section 1, the data contains 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 $\{2, 3, ..., 6, \infty\}$ and then generate 500 simulations according to the following model:

[9] See Section E of the online appendix.

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

To impose *Random Walk Errors*, the baseline simulation uses $\rho = 1, \varepsilon_{i,1} = 0, \eta_{i,t} \overset{iid}{\sim} \mathcal{N}\left(0, \sqrt{\frac{2}{5}}\right)$.[10] 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 RI estimator of BJS. Panel A of Table 1 shows results. For each estimand, the table reports the simulated variance of each estimator relative to the most efficient of the three estimators. Results for theoretically efficient estimators are in italic. Columns correspond to different variations of the simulation setup as detailed below.

The first column shows that under *Random Walk Errors*, both the SGDD and SWDD estimators perform very well at short horizons. At horizon 0, where SGDD/SWDD is efficient, 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 of SWDD/SGDD are thus equivalent to having 25-64 percent more data (variance inverse proportional to sample size) or to an 11-22 percent reduction in standard errors.

Since the simulation has staggered adoption, only the SWDD estimator is theoretically efficient at longer horizons. Additional results in the first column show 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 for SWDD. Moreover, for these longer horizons, we in fact see that the RI estimator also outperforms SGDD with an efficiency loss of only 13-16 percent relative to SWDD. This reflects that at longer horizons the RI estimator leverages more untreated units in the same way that SWDD does.

[10] The standard deviation of $\sqrt{\frac{2}{5}}$ ensures that $Var(\varepsilon_{i,t,})$ increases linearly from 0 to 2 over the sample periods.

Table 1: Comparing estimator variance in simulated and real data

Panel A: Simulations with varying error benchmarks

	Random Walk Errors	AR(1) $\rho = 0.8$	AR(1) $\rho = 0.5$	Non-stag., RW Errors
	Simulated variance relative to best shown estimator:			
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	

Panel B: Outcome variables from Brenøe *et al.* (2024)

	Total births at firm	Total leave days at firm	Number of employees	Total firm sales
Raw residual autocorrelation:	0.087	0.218	0.645	0.712
Nickell-corrected AR(1) coef.:	0.187	0.334	0.833	0.920
	Estimated variance relative to best shown estimator:			
Subgroup DID/Stepwise DID				
Horizon 0	1.530	1.354	1.000	1.000
Horizon 1	1.352	1.308	1.000	1.000
Horizon 2	1.396	1.342	1.000	1.000
Horizon 3	1.363	1.379	1.000	1.000
Regression Imputation				
Horizon 0	1.000	1.000	1.383	1.510
Horizon 1	1.000	1.000	1.226	1.201
Horizon 2	1.000	1.000	1.179	1.119
Horizon 3	1.000	1.000	1.161	1.080

The table compares estimator variance when estimating ATT_h at different horizons. Panel A shows simulation results. Columns correspond to different variations of the simulation. For each simulation and estimand, the table reports the simulated variance of the estimator relative to the best alternative. Italic denotes theoretically efficient estimators. Panel B shows results based on data in Brenøe *et al.* (2024). Columns correspond to different outcome variables. The first two rows shows measures of the error autocorrelation for each outcome variable using untreated firms across 13 years. The first row shows the raw autocorrelation in residuals from a TWFE model, the second row shows Nickell-corrected estimates assuming AR(1) errors. The remaining rows compares the different estimators. For each outcome variable and estimand, the table reports the estimated variance of the estimator relative to the best alternative. Treatment is non-staggered so SWDD and SGDD are equivalent in these data.

In practice of course, few data sets are likely to exhibit *Random Walk Errors* exactly. The second column compares 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 $\rho = 0.5$. Mechanically, this is halfway 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 similarly here. Again however, SGDD shows a substantial efficiency loss at longer horizons.

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

3.1 Practical relevance of estimator choice based on the error benchmarks

Results above suggest that there may be substaintal 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). As a check on the practical relevance of this, I examine data from Brenøe *et al.* (2024) (BCHH from now on). Using yearly Danish administrative data from 2001-2013, BCHH applies 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. BCHH is an interesting case study because of its diverse set of 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)). Panel B of Table 1 adapts the BCHH data to match the setup of the current paper and then compares estimator performance across four different firm outcome variables: Total births among employees, total leave days, number of employees and total sales.[11]

Before comparing estimator performace, Table 1 first provides evidence on the practical relevance of the random walk errors benchmark. The first rows of the table show two measures of the

[11]See the online appendix for additional details of the BCHH data and corresponding results for all outcome variables.

error persisence in each outcome variable, computed using untreated firms over all 13 years in the data. The first row shows the raw autocorrelation in residuals from a TWFE effect model (a consistent estimator of error autocorrelation as $T \longrightarrow \infty$) , while the second row shows the Nickell (1981)-corrected autocorrelation (a consistent estimator as $N \longrightarrow \infty$ if errors are AR(1)). The measured autocorrelations range from 0.087 to 0.920. This confirms the empirical relevance of both the spherical error benchmark (true autocorrelation of 0) and the random walk benchmark (true autocorrelation of 1).

The remaining rows of Panel B compares RI and SGDD/SWDD on the BCHH data. For each estimator, the usual cluster-robust standard errors are computed and the estimated variances (squared standard errors) are compared.[12] Conclusions closely mirror those from the simulation. SGDD/SWDD performs better for outcome variables with persistent errors, while RI performs better for outcome variables with impersistent errors. The estimated precision gains from choosing the right estimator also appear substantial, equivalent to as much as 50 percent more data or 18 percent smaller standard errors.

4 Conclusion: Estimator choice in practice

When analyzing difference-in-differences designs, researchers face a choice between several popular estimators that require the same assumptions for unbiasedness. This paper provides a set of new efficiency results under persistent errors. Together with previous results, these enable applied researchers to make a simple, 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 in the vein of Borusyak *et al.* (2024) 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-Differences 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

[12] Standard errors are clustered on the unit (firm). 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 numerically equivalent if using `csdid` with analytical, pointwise standard errors.

time horizons and treatment adoption is staggered, the adjusted Stepwise Difference-in-Differences estimator provides additional efficiency gains under persistent errors. A Stata package implementing Stepwise Difference-in-Differences is available on my website (`did_stepwise.ado`).

To help applied researchers distinguish outcomes with persistent and impersistent errors, two remarks are in order. First, it bears clarifification that what matters for estimator efficiency is *not* the persistence of the outcome variable itself but only the persistence of its errors. A survey measure of a person's weight, for example, will typically be very persistent over time but may have completely impersistent errors if reported weight only fluctuates because of measurement error. Second, the following question may serve as a useful heuristic for determining error persistence from theory or institutional knowledge: If a given unit at some point experiences an increase in the outcome variable relative to other units, what is the natural expectation for this unit next period? Under impersistent (spherical) errors, the outcome should tend to drop back down relative to other units in data. Under persistent (random walk) errors, the outcome should instead tend to stay high.

The results also suggest avenues for future work. First, the stepwise adjustment underlying Stepwise Difference-in-Differences can likely be adapted to offer efficiency gains also outside the canonical difference-in-differences setting. Second, as noted in the introduction, there may be cases where researchers can achieve additional efficiency gains by using more complex GMM, Feasible GLS or Synthetic DID approaches. To help harvest such additional efficiency gains, future work can develop good practical implementations of the former two methods and provide additional evidence on the finite sample performance of all relevant methods.

References

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W. and Wager, S. (2021) Synthetic difference-in-differences, *American Economic Review*, **111**, 4088–4118.

Bellégo, C., Benatia, D. and Dortet-Bernardet, V. (2024) The chained difference-in-differences, *Journal of Econometrics*.

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. (2024) Revisiting event study designs: Robust and efficient estimation, *Review of Economic Studies*.

Brenøe, A. A., Canaan, S., Harmon, N. A. and Royer, H. (2024) 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.

Clarke, D., Pailanir, D., Athey, S. and Imbens, G. W. (2023) Synthetic difference-in-differences estimation, *IZA Discussion Paper*.

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., Jorda, O. and Taylor, A. M. (2023) A local projections approach to difference-in-differences event studies, *NBER Working Paper*.

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

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*.