Ashoka University
Economics Discussion Paper 163

Efficient Estimation of Treatment Effects under Staggered Adoption: GMM Approach

May 2026

Parush Arora, Ashoka University
Rishabh Bijani, Ashoka University

https://ashoka.edu.in/economics-discussionpapers

Efficient Estimation of Treatment Effects under Staggered Adoption: GMM Approach

Parush Arora* and Rishabh Bijani*

‘

Abstract

The two-way fixed effects (TWFE) estimator is biased for the average treatment effect on the treated (ATT) when treatment is adopted in a staggered manner and effects are heterogeneous across cohorts and event times. A growing literature has responded with heterogeneity-robust estimators that restore unbiasedness and attain efficiency under the assumption of spherical errors. However, the spherical errors assumption is rarely true as panel data is typical modeled under the assumption of serial correlation. In light of this, we develop a generalized method of moments (GMM) framework that delivers an unbiased and efficient estimator of the ATT under arbitrary within-unit serial correlation. The framework is built on a modified incidence matrix that maps each 2×2 DiD comparison to its corresponding cohort-and-time-specific ATT (CATT) and explicitly debiases "forbidden" comparisons. Optimal weighting based on the inverse moment-covariance matrix yields the minimum-variance estimator within the class of unbiased estimators and consistency and asymptotic normality follow from standard GMM arguments. Monte Carlo evidence confirms substantial efficiency gains over Callaway and Sant'Anna (2021) and Sun and Abraham (2021), and modest but systematic gains over Gardner et al. (2024), Wooldridge (2025) and Borusyak et al. (2024), with the largest gains concentrated at short post-treatment horizons. Replicating Beck et al. (2010) and Cheng and Hoekstra (2013), we show that the choice of estimator is not only a technical decision but can potentially change the conclusion of the paper.

Keywords: Difference-in-differences; Two-way fixed effects; Average treatment effect on the treated; Heterogeneous treatment effects; Staggered timing; Generalized method of moments.
JEL classification: C10, C21, C22, C23.

*Department of Economics, Ashoka University.

1 Introduction

Difference-in-differences (DiD) is a workhorse design for estimating average treatment effects in settings where units are observed over time and treatment is adopted at different dates by different groups. The canonical implementation is the two-way fixed effects (TWFE) regression

$$Y_{igt} = \alpha_i + \lambda_t + \theta\, D_{it} + \varepsilon_{it}, \tag{1}$$

where Y_{igt} is the outcome of unit i in cohort g at time t, α_i and λ_t are unit and time fixed effects, D_{it} is a binary treatment indicator, and θ is interpreted as the ATT. Under staggered adoption, $\hat{\theta}_{\text{TWFE}}$ identifies the ATT only when the cohort-and-time-specific ATT (CATT) is homogeneous across (g, t) cells.

A recent literature has shown that, when CATTs are heterogeneous, $\hat{\theta}_{\text{TWFE}}$ is generally biased (Borusyak and Jaravel, 2017; De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; Athey and Imbens, 2022). The bias originates from 2×2 DiD comparisons in which an already-treated cohort serves as the control group; Goodman-Bacon (2021) term these *forbidden comparisons*. Several heterogeneity-robust estimators have been proposed in response (De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2024; Gardner et al., 2024; Wooldridge, 2025). Perhaps the most direct fix is the flexible TWFE specification of Wooldridge (2025), which fully interacts the treatment indicator with cohort-by-time dummies:

$$Y_{igt} = \alpha_i + \lambda_t + \sum_{g} \sum_{t \geq g} \theta_{gt}\, D_{it} + \varepsilon_{it}. \tag{2}$$

Under the standard assumption that $\{Y_{ig1}, \ldots, Y_{igT}, D_{i1}, \ldots, D_{iT}\}_{i=1}^{N}$ are independent across i (allowing serial correlation but not cross-sectional dependence), (2) can be estimated by ordinary least squares. Under spherical errors, the Gauss–Markov theorem implies that this is the best linear unbiased estimator. Equivalent estimators are proposed by Borusyak et al. (2024) and Gardner et al. (2024), who differ from Wooldridge (2025) in their first-stage specification, computational implementation, and inference but coincide numerically in many empirical settings.

The Gauss–Markov property breaks down when errors are non-spherical. In panel data, serial correlation is the norm rather than the exception. De Chaisemartin and d'Haultfoeuille (2023) report that serial dependence can erode the efficiency advantage of imputation-based estimators relative to alternatives such as Callaway and Sant'Anna (2021) and Sun and Abraham (2021). To our knowledge, no existing heterogeneity-robust estimator achieves efficiency under arbitrary within-unit serial correlation. Existing estimators deliver consistency and valid (cluster-robust) inference, but their weighting of 2×2 comparisons does not exploit the serial covariance structure of moments.

We try to close this gap by proposing a unified GMM framework where it rests on three ideas. First, every 2×2 DiD comparison—including forbidden ones—is treated as a moment condition for a particular CATT, with the bias of forbidden comparisons absorbed into a modified incidence matrix Q_H that maps moments to the parameter vector β of CATTs. Second,

the GMM weighting matrix A is set to the (estimated) inverse of the asymptotic covariance matrix of the moments. Under serial correlation, this delivers the minimum-variance estimator within the class of estimators based on 2×2 comparisons. Third, by stacking debiased forbidden comparisons alongside clean comparisons, the framework retains the information in already-treated controls rather than discarding it. We establish consistency and asymptotic normality of the resulting GMM estimator and provide closed-form analytical expressions for the optimal weighting matrix under serial correlation. As a complementary contribution, we extend the flexible TWFE estimator of Wooldridge (2025) to the case of serially correlated errors via iterated generalized least squares (GLS).

Monte Carlo experiments under heterogeneous CATTs and serially correlated errors confirm that optimal weighting delivers tangible efficiency gains. Compared with Callaway and Sant'Anna (2021) and Sun and Abraham (2021), the GMM estimator achieves substantial variance reductions of 30%–70% across persistence levels. Compared with Gardner et al. (2024) and Wooldridge (2025), gains are smaller—roughly 12%–19%—but systematic, increasing with the persistence parameter ρ and concentrating at the impact horizon. Intuitively, the optimal weighting matrix downweights short-lag, highly correlated DiD moments and upweights longer-lag moments that carry more orthogonal information about a given CATT. Gains shrink but do not vanish when errors are also cross-sectionally correlated through a common-factor structure—a setting in which no estimator in this literature attains efficiency. The iterated GLS extension of the flexible TWFE estimator turns out to be inferior to GMM, the OLS-based flexible TWFE, and Gardner et al. (2024), reflecting the noise introduced by estimating a high-dimensional unit-level covariance matrix.

We illustrate the framework with two empirical applications. Replicating Beck et al. (2010), we show that the estimated effect of staggered intrastate bank-branch deregulation on income inequality across U.S. states ranges from significantly negative to significantly positive depending on the estimator employed. The choice of estimator is therefore not a technical refinement: it can determine whether the central conclusion of the paper is positive, negative, or null. By contrast, replicating Cheng and Hoekstra (2013), we find that the estimators broadly agree on the homicide effect of castle-doctrine laws, because the identifying variation is dominated by clean comparisons.

The remainder of the paper is organized as follows. Section 2 describes the setup, the problem of forbidden comparisons, and the way recent estimators address it. Section 3 develops the GMM estimator, derives its asymptotic properties, characterizes the optimal weighting matrix, and discusses estimation and covariate adjustment. Section 4 reports simulation evidence. Section 5 applies the framework to Beck et al. (2010) and Cheng and Hoekstra (2013). Section 6 concludes.

2 Framework

2.1 Setup and assumptions

We observe the panel $\{Y_{ig1}, \ldots, Y_{igT}, D_{i1}, \ldots, D_{iT}\}_{i=1}^{N}$, where $Y_{igt} \in \mathbb{R}$ is the outcome of unit i in cohort g at time t, and $D_{it} \in \{0, 1\}$ is a binary treatment indicator. Units are partitioned into G cohorts indexed by their first treatment date; we let $g = \infty$ index the never-treated cohort and $g \in \{2, \ldots, T\}$ index the treated cohorts, with N_g units in cohort g and $\sum_g N_g = N$. Staggered timing means that at least two cohorts have distinct treatment dates. The framework permits the absence of a never-treated cohort.

We maintain four assumptions. **Assumption 1 (cross-sectional independence)** states that $\{Y_{ig1}, \ldots, Y_{igT}, D_{i1}, \ldots, D_{iT}\}_{i=1}^{N}$ are independent and identically distributed across i; serial dependence within units is left unrestricted, and panels need not be balanced. **Assumption 2 (irreversibility of treatment)** states that $D_{i,t-1} = 1$ implies $D_{it} = 1$. **Assumption 3 (parallel trends)** states that the expected counterfactual outcome under no treatment evolves identically across cohorts: for all g and all $t > t'$, the difference $E[Y_{igt} \mid D_{it} = 0] - E[Y_{ig,t'} \mid D_{i,t'} = 0]$ does not depend on g. We impose this assumption at the unit level, although our estimator is also valid under cohort-level parallel trends. The assumption can be conditioned on covariates to accommodate (i) covariate-specific trends in Y_{igt} and (ii) systematic differences in the distribution of covariates across cohorts; we discuss this extension in Section 3.5. **Assumption 4 (no anticipation)** states that there is no treatment effect in pre-treatment periods, so that $E[Y_{igt} \mid D_{it} = 0]$ equals the untreated potential outcome for any $t < g$. Assumptions 3 and 4 are standard in the program-evaluation literature; see, among many others, Abbring and Van den Berg (2003), Sianesi (2004), Abadie (2005), and Athey and Imbens (2022).

Let $\beta_{g,g+k}$ denote the CATT for cohort g at horizon $k \geq 0$ relative to first treatment. The aggregate ATT under heterogeneous effects is defined as a weighted average

$$\theta = \sum_{k} \sum_{g} w_{g,g+k} \, \beta_{g,g+k}, \tag{3}$$

where the weights $\{w_{g,g+k}\}$, with $\sum_g \sum_k w_{g,g+k} = 1$, are chosen by the researcher to define the parameter of interest; Callaway and Sant'Anna (2021) discuss several common aggregation schemes.

2.2 Two-by-two DiD comparisons

We index the 2×2 DiD comparisons used to identify $\beta_{g,g+k}$ by their control cohort and pre-period. Let $\hat{\beta}^{cm}_{g,g+k}$ denote a 2×2 DiD comparing cohort g in periods $g + k$ (post) and $g - m$

(pre) against control cohort c:

$$\begin{aligned}
\text{Never-treated:} \quad & \hat{\beta}^{\infty m}_{g,g+k} = (Y_{g,g+k} - Y_{g,g-m}) - (Y_{\infty,g+k} - Y_{\infty,g-m}), \\
\text{Not-yet-treated:} \quad & \hat{\beta}^{lm}_{g,g+k} = (Y_{g,g+k} - Y_{g,g-m}) - (Y_{g+l,g+k} - Y_{g+l,g-m}), \\
\text{Already-treated:} \quad & \hat{\beta}^{jm}_{g,g+k} = (Y_{g,g+k} - Y_{g,g-m}) - (Y_{g-j,g+k} - Y_{g-j,g-m}),
\end{aligned} \tag{4}$$

where $Y_{g,g+k} = \frac{1}{N_g} \sum_{i \in g} Y_{i,g,g+k}$ is the within-cohort average outcome. The superscript m denotes the number of periods between the pre-period and treatment, l denotes the periods until the not-yet-treated control's own treatment date, and j denotes the periods since the already-treated control's treatment date. The relevant index sets are $m \in \{1, \ldots, g-1\}$, $j \in \{m+1, \ldots, g-1\}$, $k \in \{0, \ldots, T-g\}$, and $l \in \{k+1, \ldots, T-g\}$. The restriction $j > m$ ensures that the already-treated control's treatment date strictly precedes the pre-period reference date, i.e. $g-j < g-m$. The restriction $l > k$ ensures that the not-yet-treated control's treatment date strictly succeeds the post-period reference date, i.e. $g+l > g+k$.

Under homogeneous effects, $E[\hat{\beta}^{\infty m}_{g,g+k} \mid D] = E[\hat{\beta}^{lm}_{g,g+k} \mid D] = E[\hat{\beta}^{jm}_{g,g+k} \mid D] = \beta_{g,g+k} = \theta$, and any weighted average

$$\hat{\theta} = \sum_m \sum_c \sum_k \sum_g w^{cm}_{g,g+k} \hat{\beta}^{cm}_{g,g+k}, \qquad \sum_m \sum_c \sum_k \sum_g w^{cm}_{g,g+k} = 1, \tag{5}$$

identifies θ. Goodman-Bacon (2021) show that $\hat{\theta}_{\text{TWFE}}$ takes precisely this form, with the weights determined by the relative variance of treatment timing across cohorts and by relative cohort sizes. The decomposition makes the implicit weighting in TWFE explicit and clarifies the conditions under which it identifies θ.

2.3 Forbidden comparisons under heterogeneous effects

When CATTs are heterogeneous, $\hat{\beta}^{jm}_{g,g+k}$ is no longer unbiased for $\beta_{g,g+k}$. In particular,

$$\begin{aligned}
E\left[\hat{\beta}^{jm}_{g,g+k}\right] &= E\left[(Y_{g,g+k} - Y_{g,g-m}) - (Y_{g-j,g+k} - Y_{g-j,g-m})\right] \\
&= \beta_{g,g+k} - \left(\beta_{g-j,g+k} - \beta_{g-j,g-m}\right),
\end{aligned} \tag{6}$$

so the bias is the change in the control cohort's CATT between the pre- and post-period. Under homogeneous effects this difference is zero and $\hat{\beta}^{jm}_{g,g+k}$ is unbiased. Under heterogeneity, however, plugging $\hat{\beta}^{jm}_{g,g+k}$ into (5) yields a biased estimator of θ. A worked example illustrating the source and magnitude of the bias is given in Appendix A.

The literature has responded with several heterogeneity-robust estimators. De Chaisemartin and d'Haultfoeuille (2020) propose an estimator applicable in settings where treatment can be switched on and off; De Chaisemartin et al. (2022) extend the approach to continuous treatments.

Callaway and Sant'Anna (2021) construct ATT estimates using never-treated or not-yet-treated cohorts as controls, eliminating forbidden comparisons by construction. Sun and Abraham (2021) propose interaction-weighted estimators for event-study designs with heterogeneous effects. Borusyak et al. (2024) propose an imputation estimator that is efficient under spherical errors. Gardner et al. (2024) estimate the ATT in two stages: first by regressing the outcome on cohort and time fixed effects in the never-treated subsample, and second by predicting counterfactuals for treated observations and averaging. Wooldridge (2025) propose the fully flexible TWFE specification in (2). The estimators of Borusyak et al. (2024), Gardner et al. (2024), and Wooldridge (2025) are numerically equivalent in many settings but differ in inference, first-stage specification, and computational implementation.

2.4 Efficiency under serial correlation

While the heterogeneity-robust estimators above achieve consistency and provide valid inference under serial correlation, none attains efficiency in this environment. Callaway and Sant'Anna (2021) construct CATT estimates as weighted averages of clean 2×2 DiD comparisons whose weights do not reflect the serial covariance structure of the moments. They derive an influence-function-based variance estimator in which each unit contributes a single influence score that sums over t, so cross-period covariances are absorbed before averaging across units. Inference is therefore valid and standard errors are cluster-robust at the unit level without requiring an explicit model of serial dependence. The estimator of Sun and Abraham (2021) is valid under arbitrary within-unit serial correlation by the same argument; because it is a linear combination of regression coefficients, standard clustered sandwich formulas apply directly. Borusyak et al. (2024) establish efficiency under spherical errors via the Gauss–Markov theorem and provide a conservative cluster-robust variance estimator under arbitrary within-unit dependence. Gardner et al. (2024) accommodate serial correlation through the GMM sandwich variance, securing valid inference, but the weighting scheme does not exploit serial dependence, leaving efficiency gains unrealized. Wooldridge (2025) recommends cluster-robust standard errors by unit, which are valid under arbitrary within-unit dependence, but the estimator itself is no longer BLUE under non-spherical errors. Borusyak et al. (2024) sketch in their appendix an efficient estimator under serial dependence that requires the moment covariance matrix to be known and recommend an auxiliary low-dimensional parametric model in practice; they do not implement or evaluate this proposal.

A natural alternative is to estimate the flexible TWFE specification (2) by iterated GLS, exploiting the block-diagonal structure of the unit-level covariance matrix. We evaluate this approach in our simulation study. As we show, in finite samples the noise introduced by estimating a high-dimensional unit-level covariance matrix offsets the efficiency benefits of correctly specifying the error structure.

3 The GMM Estimator

This section develops the GMM estimator. We begin with the homogeneous-effects case to introduce the notation and the role of the incidence matrix, then extend the framework to heterogeneous effects via a debiased incidence matrix that absorbs forbidden comparisons.

3.1 Homogeneous effects

The defining identity of CATT in a 2×2 DiD provides the moment conditions of the GMM problem. Let $N_{2\times2}$ denote the total number of 2×2 DiD comparisons that can be formed from the panel, and define

$$\phi^{cm}_{g,g+k} = (Y_{i,g,g+k} - Y_{i,g,g-m}) - (Y_{j,c,g+k} - Y_{j,c,g-m}) - \beta_{g,g+k}, \tag{7}$$

for $i \neq j$. Under Assumptions 1–4, $E[\phi^{cm}_{g,g+k}] = 0$ for every valid (g, k, c, m). Stacking all $N_{2\times2}$ moments yields

$$\phi(\beta) = \left[\phi^{\infty 1}_{i,2,2} \quad \phi^{31}_{i,2,2} \quad \cdots \quad \phi^{\infty 1}_{i,T,T}\right]', \tag{8}$$

an $N_{2\times2} \times 1$ vector with $E[\phi(\beta)] = 0$, where β is the $N_\beta \times 1$ vector of CATTs. The sample analog of the $(c, m, g, g + k)$-th moment is the difference between the corresponding 2×2 DiD estimate and the CATT,

$$\begin{aligned} \hat{\phi}^{cm}_{g,g+k} &= (Y_{g,g+k} - Y_{g,g-m}) - (Y_{c,g+k} - Y_{c,g-m}) - \beta_{g,g+k} \\ &= \hat{\beta}^{cm}_{g,g+k} - \beta_{g,g+k}, \end{aligned} \tag{9}$$

so that

$$\hat{\phi}(\beta) = \left[\hat{\phi}^{\infty 1}_{2,2} \quad \hat{\phi}^{31}_{2,2} \quad \cdots \quad \hat{\phi}^{\infty 1}_{T,T}\right]'. \tag{10}$$

The vector $\hat{\phi}(\beta)$ accommodates unbalanced panels and does not require the existence of a never-treated cohort: when no never-treated units are available, the GMM estimator can rely entirely on moment conditions formed from not-yet-treated or already-treated controls.

To obtain a compact matrix representation, let $\Delta^s_{g,g+k}$ index the s-th 2×2 DiD comparison whose expectation equals $\beta_{g,g+k}$, with $s \in \{1, \ldots, N_{g,g+k}\}$ and $N_{g,g+k}$ the total number of such comparisons. Let Δ be the $N_{2\times2} \times 1$ column vector of all 2×2 DiD estimates and Q the $N_{2\times2} \times N_\beta$ incidence matrix mapping each comparison to its corresponding CATT. Then

$$\hat{\phi}(\beta) = \Delta - Q\beta, \tag{11}$$

where

$$\Delta = \begin{bmatrix} \Delta_{2,2}^{1} \\ \vdots \\ \Delta_{2,2}^{N_{2,2}} \\ \vdots \\ \Delta_{T,T}^{1} \\ \vdots \\ \Delta_{T,T}^{N_{T,T}} \end{bmatrix}, \qquad Q = \begin{bmatrix} \boldsymbol{\ell}_{2,2} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\ell}_{2,3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\ell}_{T,T} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_{2,2} \\ \beta_{2,3} \\ \vdots \\ \beta_{T,T} \end{bmatrix},$$

and $\boldsymbol{\ell}_{g,g+k}$ is an $N_{g,g+k} \times 1$ column vector of ones. The GMM estimator minimizes the quadratic form of the sample moments,

$$\hat{\beta} = \arg\min_{\beta} \hat{\phi}(\beta)' A \, \hat{\phi}(\beta) = (Q'AQ)^{-1} Q'A\Delta, \tag{12}$$

where A is an $N_{2\times2} \times N_{2\times2}$ positive semi-definite weighting matrix. The aggregate ATT is then

$$\hat{\theta} = \boldsymbol{w}'\hat{\beta}, \tag{13}$$

with $\boldsymbol{w}$ the column vector of aggregation weights chosen by the researcher.

3.2 Heterogeneous effects and a debiased incidence matrix

Under heterogeneous effects, $\hat{\beta}^{jm}_{g,g+k}$ is contaminated by (6), so the moment conditions split into three categories:

$$\begin{aligned} \hat{\phi}^{\infty m}_{g,g+k} &= \hat{\beta}^{\infty m}_{g,g+k} - \beta_{g,g+k}, \\ \hat{\phi}^{lm}_{g,g+k} &= \hat{\beta}^{lm}_{g,g+k} - \beta_{g,g+k}, \\ \hat{\phi}^{jm}_{g,g+k} &= \hat{\beta}^{jm}_{g,g+k} - \beta_{g,g+k} + \beta_{g-j,g+k} - \beta_{g-j,g-m}. \end{aligned} \tag{14}$$

The first two categories take the same form as in the homogeneous case. The third absorbs the bias of forbidden comparisons by adding the contaminating CATTs of the already-treated control cohort to the moment condition. Stacking all three categories and rearranging, we obtain

$$\hat{\phi}(\beta) = \Delta - Q_H \beta, \tag{15}$$

where the modified incidence matrix Q_H takes the form

$$Q_H = \begin{bmatrix} \boldsymbol{\ell}_{2,2} & \ell^{2}_{2,3} & \cdots & \ell^{N_\Delta}_{T,T} \\ \ell^{1}_{2,2} & \boldsymbol{\ell}_{2,3} & \cdots & \ell^{N_\Delta}_{T,T} \\ \vdots & \vdots & \ddots & \vdots \\ \ell^{1}_{2,2} & \ell^{2}_{2,3} & \cdots & \boldsymbol{\ell}_{T,T} \end{bmatrix}.$$

Each off-diagonal block $\ell^p_{g,g+k}$ is an $N_{g,g+k} \times 1$ column vector that records, for the rows corresponding to comparisons of $\beta_{g,g+k}$, how a CATT in column p enters as a contaminant. Specifically, the s-th entry of $\ell^p_{g,g+k}$ equals

$$\ell^p_{g,g+k}[s] = \begin{cases} -1 & \text{if comparison } s \text{ uses already-treated control } g-j \text{ and column } p \text{ is } \beta_{g-j,g-m}, \\ 1 & \text{if comparison } s \text{ uses already-treated control } g-j \text{ and column } p \text{ is } \beta_{g-j,g+k}, \\ 0 & \text{otherwise.} \end{cases}$$

The GMM estimator under heterogeneous effects becomes

$$\hat{\beta} = (Q_H' A Q_H)^{-1} Q_H' A \Delta. \tag{16}$$

By construction, Q_H debiases forbidden comparisons through the incidence structure rather than excluding them, allowing the GMM objective to absorb the information they contain. A worked example illustrating the construction of Q_H in a three-cohort, three-period design is provided in Appendix B.

3.3 Efficiency

The weighting matrix A governs how the moment conditions interact in the GMM objective. Several familiar choices arise as special cases. If A is diagonal with weights matching those of the Goodman–Bacon decomposition, the resulting estimator is numerically equivalent to TWFE, which is BLUE under the Gauss–Markov theorem when errors are spherical. Setting $A = I$ weights all 2×2 DiDs equally; raising any diagonal element places more weight on the corresponding comparison. A diagonal A implicitly treats the moment conditions as independent. This is never literally true in our setting: a single unit-by-time observation enters multiple moment conditions, and serial correlation in the errors propagates dependence across moments through their leads and lags. Allowing A to be non-diagonal, and in particular to reflect the moment covariance structure, is therefore essential for efficiency.

By standard GMM theory (Hansen, 1982), the optimal weighting matrix satisfies $A \xrightarrow{p} \Omega_\phi^{-1}$, where

$$\Omega_\phi = \lim_{N\to\infty} \mathrm{Var}\left[\sqrt{N}\,\hat{\phi}(\beta)\right] = \lim_{N\to\infty} E\left[N\,\hat{\phi}(\beta)\hat{\phi}(\beta)'\right]. \tag{17}$$

Setting $A = \hat{\Omega}_\phi^{-1}$ for a consistent estimator $\hat{\Omega}_\phi$ minimizes the asymptotic variance of $\hat{\beta}$. Using the weak law of large numbers and the continuous mapping theorem, we have

$$\mathrm{Var}\left[\sqrt{N}\,\hat{\phi}(\beta)\right] = \mathrm{Var}\left[\sqrt{N}\,\Delta\right] = N\,\Omega_\Delta, \tag{18}$$

where $\Omega_\Delta = \mathrm{Var}[\Delta]$ has the block structure

$$\Omega_\Delta = \begin{bmatrix} \mathrm{Var}[\Delta^1_{2,2}] & \mathrm{Cov}[\Delta^1_{2,2}, \Delta^2_{2,2}] & \cdots & \mathrm{Cov}[\Delta^1_{2,2}, \Delta^{N_{T,T}}_{T,T}] \\ \mathrm{Cov}[\Delta^2_{2,2}, \Delta^1_{2,2}] & \mathrm{Var}[\Delta^2_{2,2}] & \cdots & \mathrm{Cov}[\Delta^2_{2,2}, \Delta^{N_{T,T}}_{T,T}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[\Delta^{N_{T,T}}_{T,T}, \Delta^1_{2,2}] & \mathrm{Cov}[\Delta^{N_{T,T}}_{T,T}, \Delta^2_{2,2}] & \cdots & \mathrm{Var}[\Delta^{N_{T,T}}_{T,T}] \end{bmatrix}.$$

3.3.1 Analytical structure under serial correlation

To exemplify, we derive $\text{Var}[\hat{\phi}(\beta)]$ under the simplifying assumption of stationary outcomes with $\text{Var}[Y_{igt}] = \sigma_0$ and $\text{Cov}[Y_{igt}, Y_{igs}] = \sigma_{|t-s|}$ for $t \neq s$. The assumption can be relaxed to allow for time-varying variances and covariances at the cost of additional notation.[1] The diagonal elements of Ω_Δ take the form

$$\text{Var}[\Delta^s_{g,g+k}] = \left(\frac{2}{N_g} + \frac{2}{N_c}\right)(\sigma_0 - \sigma_d),$$

where N_c is the size of the control cohort and d is the calendar distance between the post- and pre-period of the comparison. The off-diagonal elements depend on whether the focal and control cohorts coincide across comparisons:

$$\begin{cases} \text{if } g \neq g' \text{ and } c \neq c' \text{ and } \{g,c\} \cap \{g',c'\} = \emptyset: \\ \quad \text{Cov}\left[\hat{\beta}^{cm}_{g,g+k}, \hat{\beta}^{c'm'}_{g',g'+k'}\right] = 0, \\ \text{if } g = g' \text{ and } c \neq c': \\ \quad \text{Cov}\left[\hat{\beta}^{cm}_{g,g+k}, \hat{\beta}^{c'm'}_{g',g'+k'}\right] = \frac{1}{N_g}\left(\sigma_{|k-k'|} - \sigma_{|k-m'|} - \sigma_{|m-k'|} + \sigma_{|m-m'|}\right), \\ \text{if } g \neq g' \text{ and } c = c': \\ \quad \text{Cov}\left[\hat{\beta}^{cm}_{g,g+k}, \hat{\beta}^{c'm'}_{g',g'+k'}\right] = \frac{1}{N_c}\left(\sigma_{|k-k'|} - \sigma_{|k-m'|} - \sigma_{|m-k'|} + \sigma_{|m-m'|}\right), \\ \text{if } g' = c \text{ and } g \neq c' \text{ (or symmetric):} \\ \quad \text{Cov}\left[\hat{\beta}^{cm}_{g,g+k}, \hat{\beta}^{c'm'}_{g',g'+k'}\right] = -\frac{1}{N_g}\left(\sigma_{|k-k'|} - \sigma_{|k-m'|} - \sigma_{|m-k'|} + \sigma_{|m-m'|}\right), \\ \text{if } g = g' \text{ and } c = c': \\ \quad \text{Cov}\left[\hat{\beta}^{cm}_{g,g+k}, \hat{\beta}^{c'm'}_{g',g'+k'}\right] = \left(\frac{1}{N_g} + \frac{1}{N_c}\right)\left(\sigma_{|k-k'|} - \sigma_{|k-m'|} - \sigma_{|m-k'|} + \sigma_{|m-m'|}\right), \end{cases}$$

with $\sigma_{|r-r'|} = \sigma_0$ when $r = r'$. To obtain Ω_ϕ, we take the asymptotic limit element-wise. The diagonal entries scale as $\lim_{N\to\infty} \text{Var}[\sqrt{N}\,\Delta^s_{g,g+k}] = 4(\sigma_0 - \sigma_d)$, and the cohort-size factors drop out of the covariance terms in the limit, yielding

$$\begin{aligned} \text{if } g \neq g',\ c \neq c',\ \{g,c\} \cap \{g',c'\} = \emptyset: &\quad \lim_{N\to\infty} N\,\text{Cov}[\cdot] = 0, \\ \text{if } g = g',\ c \neq c' \text{ or } g \neq g',\ c = c': &\quad \lim_{N\to\infty} N\,\text{Cov}[\cdot] = \sigma_{|k-k'|} - \sigma_{|k-m'|} - \sigma_{|m-k'|} + \sigma_{|m-m'|}, \\ \text{if } g' = c,\ g \neq c' \text{ (or symmetric):} &\quad \lim_{N\to\infty} N\,\text{Cov}[\cdot] = -\left(\sigma_{|k-k'|} - \sigma_{|k-m'|} - \sigma_{|m-k'|} + \sigma_{|m-m'|}\right), \\ \text{if } g = g',\ c = c': &\quad \lim_{N\to\infty} N\,\text{Cov}[\cdot] = 2\left(\sigma_{|k-k'|} - \sigma_{|k-m'|} - \sigma_{|m-k'|} + \sigma_{|m-m'|}\right). \end{aligned}$$

Because σ_0 and the σ_d are unknown in practice, Ω_ϕ must be estimated. One option is to run iterated GLS on the TWFE specification, recovering estimates of σ_0 and σ_d from the residuals.

[1] We flag, however, that imposing time-invariance is a substantive simplification: panel data often exhibit heteroskedasticity over the calendar dimension, and a fully nonstationary covariance structure requires additional indexing.

3.3.2 Iterated GMM

A computationally simpler alternative is iterated GMM, an extension of two-step GMM (Hansen et al., 1996). In the first step, $\hat{\beta}$ is computed using an arbitrary positive-definite weighting matrix (we use $A = I$ in practice). The first-step estimator is then used to construct a consistent estimate $\hat{\Omega}_\phi$ of the moment covariance matrix, and the GMM objective is re-minimized with $A = \hat{\Omega}_\phi^{-1}$. The procedure is iterated until both $\hat{\beta}$ and A stabilize.

A subtlety arises because Ω_ϕ is generically singular. Each DiD moment is a linear combination of four group-time means Y_{gt}. As a consequence, $\text{rank}(\Omega_\phi) = |G|(T-1)$, where $G \subseteq \{2, \ldots, T\}$ is the set of treated cohort onset periods, so Ω_ϕ is singular whenever $N_{2\times 2} > |G|(T-1)$, which is typically the case. This singularity is benign for the GMM estimator: Ω_ϕ enters only through the sandwich $Q'\Omega_\phi^{-1}Q$, an $N_\beta \times N_\beta$ object. The matrix Ω_ϕ is effectively projected down to the parameter space through Q, and directions in the null space of Ω_ϕ that lie outside the range of Q' contribute negligibly. In our implementation we add a small ridge term νI to Ω_ϕ before inversion. This shifts every eigenvalue by ν, ensuring numerical positive-definiteness. Within the range of Q', the addition is asymptotically negligible; in the null space, the inverse eigenvalue $1/\nu$ is large but is annihilated by Q in the sandwich, leaving $Q'\Omega_\phi^{-1}Q$ well-conditioned.

Algorithm: Iterated GMM

1. Set $A^{(0)} = I$ and compute $\hat{\beta}^{(0)} = (Q_H^\top Q_H)^{-1} Q_H^\top \Delta$.
2. Iterate for $j = 1, \ldots, J_{\max}$.
   (a) Form treatment-adjusted outcomes $\tilde{Y}_{it} = Y_{it} - \hat{\tau}_{it}^{(j-1)}$, where $\hat{\tau}_{it}^{(j-1)}$ plugs in $\hat{\beta}^{(j-1)}$ for treated cells.
   (b) Two-way demean $\tilde{Y}_{it}$ on unit and time fixed effects and store residuals $\hat{v}_{it}$.
   (c) Estimate the auto-covariances $\hat{\sigma}_d = \frac{1}{N(T-d)} \sum_{i,t} \hat{v}_{it}\hat{v}_{i,t+d}$ for $d = 0, \ldots, T-1$ and update $\hat{\Omega}_\phi$.
   (d) Update $\hat{\beta}^{(j)} = (Q_H^\top \hat{\Omega}_\phi^{-1} Q_H)^{-1} Q_H^\top \hat{\Omega}_\phi^{-1} \Delta$

To ensure that $\hat{\Omega}_\phi$ is positive definite in each iteration, add νI where ν is the minimum possible value. Alternatively, one can run GLS on (15) and the resulting estimator has the same closed form solution as GMM.

3.4 Inference

The variance of the GMM CATT estimator is

$$\begin{aligned} \text{Var}[\hat{\beta}] &= \text{Var}\left[(Q_H' A Q_H)^{-1} Q_H' A \Delta\right] \\ &= (Q_H' A Q_H)^{-1} Q_H' A \, \Omega_\Delta \, A Q_H (Q_H' A Q_H)^{-1}, \end{aligned} \tag{19}$$

and the variance of the aggregate ATT estimator is

$$\mathrm{Var}[\hat{\theta}] = \boldsymbol{w}' \,\mathrm{Var}[\hat{\beta}]\boldsymbol{w} = \boldsymbol{w}'(Q_H' A Q_H)^{-1} Q_H' A\, \Omega_\Delta\, A Q_H (Q_H' A Q_H)^{-1}\boldsymbol{w}. \tag{20}$$

A consistent estimator is obtained by replacing Ω_Δ with $\hat{\Omega}_\Delta$, constructed from $\hat{\sigma}_0$ and the $\hat{\sigma}_d$ recovered through iterated GMM estimation:

$$\widehat{\mathrm{Var}}[\hat{\beta}] = (Q_H' A Q_H)^{-1} Q_H' A\, \hat{\Omega}_\Delta\, A Q_H (Q_H' A Q_H)^{-1}. \tag{21}$$

This formula yields standard errors and pointwise confidence intervals for both CATTs and the aggregate ATT.

The consistency and asymptotic normality of $\hat{\beta}$ follow from standard results for GMM (Hansen, 1982; Newey and McFadden, 1994). Under Assumptions 1–4 and standard regularity conditions namely, identification ($E[\phi_i(\beta)] = 0$ if and only if β is the true parameter), compactness of the parameter space, continuity of $\phi_i(\beta)$, an integrable envelope function, and any positive definite matrix $\hat{A} \xrightarrow{p} A$ (A is positive definite), we have $\hat{\beta} \xrightarrow{p} \beta$ keeping T as fixed, and by the continuous mapping theorem $\hat{\theta} = \boldsymbol{w}'\hat{\beta} \xrightarrow{p} \boldsymbol{w}'\beta = \theta$. Under additional regularity (interior β, continuous differentiability of ϕ_i in a neighborhood of β, integrable derivative envelope, non-singular $Q_H' A Q_H$, existence of Ω_ϕ),

$$\sqrt{N}\,(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\big(0,\, (Q_H' A Q_H)^{-1}(Q_H' A\, \Omega_\phi\, A Q_H)(Q_H' A Q_H)^{-1}\big)\,,$$

and by the delta method

$$\sqrt{N}\,(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\big(0,\, \boldsymbol{w}'(Q_H' A Q_H)^{-1}(Q_H' A\, \Omega_\phi\, A Q_H)(Q_H' A Q_H)^{-1}\boldsymbol{w}\big)\,.$$

At the optimal weighting matrix $A = \hat{\Omega}_\phi^{-1}$, the asymptotic variance simplifies to $\boldsymbol{w}'(Q_H' \Omega_\phi^{-1} Q_H)^{-1}\boldsymbol{w}$.

3.5 Conditioning on observed covariates

In many empirical applications the unconditional parallel-trends assumption is implausible, and it is more credible to assume that parallel trends hold only after conditioning on a vector of observed covariates X_i. This arises naturally when (i) covariates generate cohort-specific trends in Y_{igt} or (ii) the distribution of covariates differs across cohorts, so unconditional comparisons confound treatment effects with compositional differences. We replace Assumption 3 with its conditional counterpart, **Assumption** 3′ **(conditional parallel trends)**: for all g and all $t > t'$, $E[Y_{igt} \mid D_{it} = 0, X_{it}] - E[Y_{ig,t'} \mid D_{i,t'} = 0, X_{it}]$ does not depend on g. That is, conditional on X_{it}, the evolution of untreated potential outcomes is the same across cohorts.

To incorporate covariates we modify the moment conditions at the unit level and aggregate. The key idea is to construct covariate-adjusted 2×2 DiD estimates that partial out the effect of X_i on outcome trends before forming moments. Following Callaway and Sant'Anna (2021),

$$\tilde{\beta}^{cm}_{g,g+k} = \frac{1}{N_g} \sum_{i \in g} \Big[\big(M_{g,g+k} - M_{g,g-m}\big) - \big(M_{c,g+k} - M_{c,g-m}\big) \Big], \tag{22}$$

where $M_{g,t}$ is an estimate of the conditional mean function $E[Y_{igt} \mid D_{it} = 0, X_{it}]$. If the conditional mean is linear, $M_{g,t} = X'_{it}\beta_X$, then β_X is estimated by regressing $Y_{i,g,g+k}$ on $X_{i,g+k}$ within cohort g at time $g+k$, after which the fitted values are averaged. The covariate-adjusted moment conditions

$$\tilde{\phi}^{cm}_{g,g+k} = \tilde{\beta}^{cm}_{g,g+k} - \beta_{g,g+k} \tag{23}$$

satisfy $E[\tilde{\phi}^{cm}_{g,g+k}] = 0$ under Assumption 3′. Stacking all covariate-adjusted DiD estimates into $\tilde{\Delta}$, the GMM system retains the same structure,

$$\tilde{\phi}(\beta) = \tilde{\Delta} - Q_H\beta, \tag{24}$$

and the GMM estimator becomes

$$\tilde{\beta} = (Q'_H A Q_H)^{-1} Q'_H A\, \tilde{\Delta}. \tag{25}$$

Only the input vector changes; everything else—the incidence matrix, the optimal weighting, the inference formulas—is unchanged. The outcome-regression approach extends naturally to inverse probability weighting (IPW) or doubly robust (DR) estimation (Callaway and Sant'Anna, 2021), in which the outcome adjustment is combined with propensity-score reweighting; the resulting estimator is consistent if either model is correctly specified, and the corresponding 2×2 estimates can be substituted directly into $\tilde{\Delta}$. The over-identification test developed in Section **??** remains valid under covariate adjustment and continues to test the conditional parallel-trends assumption.

4 Simulation Study

We evaluate the finite-sample performance of the GMM estimator against existing heterogeneity-robust alternatives in 500 simulated panels with $T = 33$ time periods. We consider both small-sample ($N_g = 10$) and large-sample ($N_g = 50$) designs to assess whether the noise from estimating the moment covariance matrix under GMM, or the unit-level covariance matrix under iterated GLS, offsets the benefits of correctly specifying the error structure—a concern that is most acute when the sample size is small. The outcome is generated as

$$Y_{it} = \alpha_i + \lambda_t + \tau_{it} + \varepsilon_{it}, \tag{26}$$

with unit fixed effect $\alpha_i \sim \mathcal{N}(0,1)$, time fixed effect $\lambda_t \sim \mathcal{N}(0,1)$, and treatment effect

$$\tau_{it} = \beta_g \,(1 + r_g)^{\,t-g_i}\, D_{it},$$

where g_i is the first treatment date for unit i ($g_i = 0$ if never-treated), $D_{it} = \mathbf{1}\{t \geq g_i\}$, β_g is the cohort-specific impact effect, and r_g is the cohort-specific growth rate of the treatment effect. We consider six cohorts: five treated cohorts with $T_g \in \{10, 13, 16, 19, 22\}$ and one never-treated cohort. The cohort-specific impact effects and growth rates are

$$\beta_g \in \{-16, -12, -10, -9, -2\}, \qquad r_g \in \{0.01, 0.04, 0.08, 0.10, 0.07\}.$$

Aggregation uses unit weights, in line with Callaway and Sant'Anna (2021).

We compare the GMM estimator with five alternatives: Callaway and Sant'Anna (2021), hereafter **CS**; Sun and Abraham (2021), hereafter **SA**; Gardner et al. (2024), hereafter **Gardner**; Wooldridge (2025), hereafter **Flex TWFE**; and an iterated GLS implementation of Wooldridge (2025), hereafter **GLS Flex TWFE**, which estimates the unit-level covariance matrix under serial correlation. Two error-generating processes are considered. The first is an AR(1) process,

$$\varepsilon_{it} = \rho\, \varepsilon_{i,t-1} + u_{it}, \qquad u_{it} \sim \mathcal{N}(0,1),$$

with $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ to span low to high persistence. The second introduces cross-sectional dependence through a factor structure,

$$\varepsilon_{it} = \Phi_{i,\cdot}\, f_{t,\cdot}^{\top} + \sqrt{D_i}\, \eta_{it},$$

where $\Phi_{i,k} \sim \mathcal{N}(0, 0.4^2)$ are unit loadings on k_f factors, f_t is a vector of common AR(1) factors with persistence $\rho = 0.5$, $D_i \sim \text{Unif}(0.5, 2.0)$ is unit-specific idiosyncratic variance, and $\eta_{it} \sim \mathcal{N}(0,1)$. This design induces both serial and cross-sectional error dependence.

4.1 Results under serially correlated errors

Table 1 reports bias and variance under AR(1) errors. CS and SA produce numerically equivalent estimates, as do Flex TWFE and Gardner; we report each pair as a single row (CS/SA and Flex TWFE/Gardner). All four estimators—CS, Gardner, GLS Flex TWFE, and GMM—are numerically unbiased at every value of ρ, but they differ markedly in efficiency.

The efficiency advantage of GMM and Gardner over CS is large, ranging from 30% to 70% variance reduction across ρ. The gap shrinks at higher persistence, where CS becomes more competitive. The GLS Flex TWFE estimator delivers higher variance than both Gardner and GMM, plausibly because its covariance matrix has dimension $T \times T$ per unit and is poorly estimated when the sample is small relative to the parameter dimension. The GMM estimator performs best across all configurations and delivers a modest but systematic gain over Gardner, with variance reduction of roughly 12%–19% in both small and large samples. The gain is largest when ρ is high and marginal when ρ is small. Interestingly, the relative gain over Gardner is somewhat larger in the smaller sample, contrary to a naive expectation. The optimal weighting matrix incorporates the error dependence structure with a substantially lower-dimensional object than the unit-level covariance matrix used by GLS Flex TWFE, which mitigates overfitting.[2] The benefits of correctly specifying the moment covariance therefore outweigh the noise introduced by estimating it—a balance that the GLS Flex TWFE estimator does not strike.

[2] We re-estimate Table 1 under homogeneous effects ($\beta_g = -5$ and $r_g = 0$ for all g). The results are similar, with TWFE recovering the most efficient estimator, as expected under the Gauss–Markov theorem.

Table 1: Performance of ATT estimators when Errors are Serially Correlated

ρ	Estimator	$N_g = 10$		$N_g = 50$	
		Bias	Variance	Bias	Variance
0.1	CS/SA	0.0016	0.0464	-0.0005	0.0104
	Flex TWFE/Gardner	-0.0199	**0.0136**	0.0027	0.0027
	GLS Flex TWFE	-0.0206	0.0151	0.0014	0.0027
	GMM	-0.0200	**0.0136**	0.0025	**0.0026**
0.3	CS/SA	0.0200	0.0566	-0.0040	0.0095
	Flex TWFE/Gardner	0.0321	0.0207	0.0031	**0.0033**
	GLS Flex TWFE	0.0336	0.0220	0.0095	0.0035
	GMM	0.0324	**0.0199**	0.0031	0.0035
0.5	CS/SA	0.0122	0.0462	-0.0110	0.0076
	Flex TWFE/Gardner	0.0298	0.0310	-0.0054	0.0043
	GLS Flex TWFE	0.0337	0.0334	-0.0029	0.0048
	GMM	0.0273	**0.0299**	0.0055	**0.0039**
0.7	CS/SA	-0.0349	0.0674	0.0016	0.0105
	Flex TWFE/Gardner	-0.0209	0.0465	0.0111	0.0078
	GLS Flex TWFE	-0.0112	0.0529	0.0078	0.0096
	GMM	-0.0109	**0.0361**	0.0103	**0.0069**
0.9	CS/SA	0.0351	0.0662	0.0000	0.0132
	Flex TWFE/Gardner	0.0568	0.0694	0.0004	0.0112
	GLS Flex TWFE	0.0465	0.0962	-0.0017	0.0146
	GMM	0.0401	**0.0561**	0.0099	**0.0092**

Notes: The lowest variance is highlighted.

4.2 Results under cross-sectionally correlated errors

Table 2 reports performance under the factor-model error process, which adds cross-sectional dependence to within-unit serial correlation. The performance gap between GMM and Gardner narrows: GMM delivers a 7% variance reduction at $N_g = 50$ and only marginal gains at $N_g = 10$. No estimator is efficient in this environment, since none of the candidates models cross-sectional dependence; nevertheless, GMM remains the best performer in the pool of inefficient estimators.

Table 2: Performance of ATT estimators when errors are serially and cross-sectionally correlated

	$N_g = 10$		$N_g = 50$	
Estimator	Bias	Variance	Bias	Variance
CS/SA	-0.0192	0.1027	0.0036	0.0251
Flex TWFE/Gardner	-0.0020	0.0338	0.0001	0.0103
GLS Flex TWFE	0.0040	0.0424	0.0027	0.0113
GMM	-0.0021	0.0335	0.0007	0.0096

Notes: Bias and variance of the aggregate ATT estimator across 500 simulations under the factor-model error process described in the text.

4.3 Where do the gains come from?

Table 3 disaggregates the variance comparison between GMM and Gardner by event time k for $\rho = 0.7$. The efficiency advantage of GMM is sharply concentrated at short post-treatment horizons. At impact ($k = 0$), GMM delivers lower variance for all five CATTs in both sample sizes, with average variance reduction of roughly 47%–49%. The advantage remains substantial for $k = 1$–3 at about 17%–19% and decays slowly thereafter. Beyond $k \geq 8$, GMM and Gardner perform comparably.

Table 3: Average percentage variance reduction in GMM relative to Gardner by relative time, AR(1) errors with $\rho = 0.7$

		Avg. % variance reduction	
Rel. time	# of CATTs	Sim I ($N_g = 10$)	Sim II ($N_g = 50$)
$k = 0$	5	47.1%	49.3%
$k = 1$–3	15	16.9%	18.6%
$k = 4$–7	20	4.5%	5.6%
$k = 8$–11	20	2.3%	3.4%
$k = 12$–15	15	3.1%	4.6%
$k \geq 16$	15	0.9%	2.1%
Overall	90	8.3%	7.4%

This pattern is consistent with the structure of the optimal weighting matrix. At short post-treatment horizons, the cross-moment covariances induced by serial correlation are substantial, and the optimal weighting matrix downweights these correlated short-lag DiD comparisons in favor of longer-lag comparisons that carry more orthogonal information about a given CATT. As k grows, the marginal contribution of the AR(1) structure to efficiency diminishes, and GMM converges to Gardner.

5 Empirical Applications

5.1 Bank-branch deregulation and inequality (Beck et al., 2010)

Beck et al. (2010) examined how the staggered intrastate deregulation of bank branching affected income inequality across 49 U.S. states between 1976 and 2006. Baker et al. (2022) revisited the analysis and found that the estimated ATT is sensitive to the choice of estimator. One source of sensitivity is the inclusion of 12 states that were already deregulated at the start of the sample period (the always-treated cohort). Modern staggered-DiD estimators require support at every event-time, that is, both treated and untreated (or not-yet-treated) units must exist at each comparison horizon. We therefore examine two samples: the full panel of 49 states corresponding to the original Beck et al. (2010), and a reduced panel of 36 states obtained by dropping the 12 always-treated states.

The baseline specification is the TWFE DiD

$$Y_{st} = \alpha_s + \lambda_t + \beta \, \text{Dereg}_{st} + \varepsilon_{st}, \tag{27}$$

where Y_{st} is the log Gini coefficient for state s at year t, α_s and λ_t are state and year fixed effects, and standard errors are clustered by state. Following Table II, Panel A, Column (2) of Beck et al. (2010), the baseline result is $\hat{\beta} \approx -0.022$ with standard error 0.008, indicating that deregulation reduced inequality. Our replication recovers $\hat{\beta} = -0.0220$ with standard error 0.0075 on 1,519 state–year observations.

Table 4 reports the TWFE estimate together with the Goodman–Bacon decomposition. In the full sample (Panel A), the negative TWFE estimate of Beck et al. (2010) is driven almost entirely by 2×2 DiD comparisons in which already-treated or always-treated states serve as controls: only 12.5% of the weight is assigned to comparisons against never-treated or not-yet-treated controls, whereas the forbidden and always-treated comparisons account for 87.5% of the weight and yield large negative averages. Strikingly, the average 2×2 DiD using clean (not-yet-treated) comparisons is positive, 0.0061. In the reduced sample (Panel B), the TWFE estimate falls in magnitude to -0.0134 but remains significant; it is again driven by forbidden comparisons, which now carry 71.9% of the weight.

We re-estimate (27) using ATT estimators that are robust to heterogeneous effects. The estimators fall into three groups: the standard TWFE regression; four heterogeneity-robust estimators (CS, SA, Gardner, and Flexible TWFE); and our GMM estimator under three alternative

Table 4: TWFE estimates and Goodman–Bacon decomposition, Beck et al. (2010)

	Panel A Full sample	**Panel B** Reduced sample
TWFE estimate	−0.0213	−0.0134
	(0.0036)	(0.0045)
Later vs. always-treated		
Weight share	0.5559	—
Avg. 2×2 estimate	−0.0276	—
Later vs. already-treated		
Weight share	0.3195	0.7194
Avg. 2×2 estimate	−0.0210	−0.0210
Earlier vs. later-treated		
Weight share	0.1246	0.2806
Avg. 2×2 estimate	0.0061	0.0061
Observations	1,519	1,116
States	49	36
Years	31	31

Notes: TWFE coefficient on the deregulation indicator and Goodman-Bacon (2021) decomposition of the underlying 2×2 DiD components. Standard errors clustered by state in parentheses.

weighting matrices, namely identity, diagonal inverse-covariance, and full inverse-covariance. The results are reported in Table 5.

The wide dispersion of estimates in Table 5 reflects three distinct sources of divergence: the bias from forbidden comparisons embedded in TWFE; differences in control-group definition and aggregation among the heterogeneity-robust estimators; and differences in how efficiently each estimator exploits the available moment conditions.

The CS, Gardner, and Flexible TWFE estimators all eliminate forbidden comparisons by construction, yet they yield markedly different results. CS finds a negative but insignificant ATT in Panel A (−0.0101) and a small positive ATT in Panel B (+0.0062), consistent with a near-zero true effect. Gardner yields positive and significant estimates of approximately +0.020 in both panels. The Flexible TWFE estimator diverges sharply across panels: it yields −0.0165 (insignificant) in Panel A but +0.0195 (significant) in Panel B. The divergence reflects a conceptual difference: the Flexible TWFE, unlike Gardner's two-stage approach, includes always-treated states in the common time fixed-effects estimation. Since these 12 states contribute only post-treatment observations, they shift the estimated time trends and thereby alter the counterfactual baseline for all other cohorts. Gardner avoids this contamination by estimating unit and time fixed effects exclusively from untreated observations in the first stage, so its estimate is invariant to whether always-treated states are retained.

The SA estimator stands apart, yielding large, negative, and highly significant estimates (−0.0354 in Panel A, −0.0536 in Panel B). The divergence is attributable to its aggregation

Table 5: Effect of branch-banking deregulation on ln(Gini), Beck et al. (2010)

	Panel A (49 states)		*Panel B* (36 states)	
Estimator	ATT	SE	ATT	SE
Pooled TWFE	-0.0213^{***}	(0.0076)	-0.0134^{*}	(0.0071)
CS	-0.0101	(0.0078)	0.0062	(0.0053)
SA	-0.0354^{***}	(0.0114)	-0.0536^{***}	(0.0063)
Gardner	0.0195^{***}	(0.0067)	0.0195^{***}	(0.0067)
Flexible TWFE	-0.0165	(0.0141)	0.0195^{***}	(0.0069)
GMM-Identity	0.0266	(0.0190)	0.0266	(0.0187)
GMM-Diagonal	0.0247	(0.0187)	0.0247	(0.0180)
GMM-Full	0.0002	(0.0141)	0.0007	(0.0140)

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ based on two-sided z-tests. TWFE, SA, CS, and Gardner standard errors are clustered by state. GMM-Identity standard errors use the sandwich formula $(Q'Q)^{-1}Q'\hat{\Omega}Q(Q'Q)^{-1}$; GMM-Diagonal uses $(Q'AQ)^{-1}Q'A\hat{\Omega}AQ(Q'AQ)^{-1}$ with $A = \text{diag}(\hat{\Omega})^{-1}$; GMM-Full uses the delta-method formula $(Q'\hat{\Omega}^{-1}Q)^{-1}$. $\hat{\Omega}$ is estimated from auto-covariances of treatment-adjusted, two-way-demeaned residuals. CS uses not-yet-treated states as controls with doubly robust estimation and group-level aggregation. SA uses interaction-weighted aggregation; always-treated states are coded as never-treated ($g = \infty$) in Panel A. Red denotes a significant negative estimate, blue denotes a significant positive estimate.

scheme: SA produces interaction-weighted ATT estimates that average CATTs over event time rather than over cohorts. If early-treatment cohorts have systematically different underlying inequality trends—which is plausible for states that deregulated before others—the event-time weighting can amplify cohort-composition effects. In Panel A, the inclusion of always-treated states as an implicit never-treated reference group further distorts the baseline comparison. The SA estimate should therefore be interpreted with caution in this application.

The divergence between Gardner ($\approx +0.020$) and CS (near-zero) reflects a subtler difference. Both eliminate forbidden comparisons, but they exploit the data differently. Gardner relies on parallel trends holding globally—it uses pre-treatment variation from all untreated units to estimate counterfactual trends for all treated cohorts. CS estimates cohort-specific counterfactuals using only the cohort's own not-yet-treated comparisons. When treatment-effect heterogeneity is strong and cohort compositions differ, the two approaches diverge. Panel B is the cleaner comparison: in the absence of always-treated contamination, Flexible TWFE agrees with Gardner (+0.0195), suggesting that the imputation approach consistently delivers a positive estimate once the data support it. The Panel A discrepancy for Flexible TWFE (-0.0165) should be attributed to distorted time fixed effects caused by always-treated units rather than to a genuine treatment effect.

The three GMM estimators operate on 5,829 not-yet-treated 2×2 DiD comparisons that span 236 cohort-time CATT cells, differing only in the weighting matrix applied to those moments. GMM-Identity assigns equal weight to all 5,829 DiD moments and yields $\hat{\theta} \approx +0.027$. GMM-Diagonal upweights moments with low marginal variance and downweights high-variance moments, yielding $\hat{\theta} \approx +0.025$—a modest shift toward zero that reflects the partial information in the diagonal of Ω. GMM-Full applies the full inverse-covariance matrix Ω^{-1} as the weighting matrix, and the estimate collapses to nearly zero ($\approx +0.0002$ in Panel A and $\approx +0.0007$ in Panel B). The 5,829 DiD moments are clearly not independent, since moments sharing a focal cohort, a control cohort, or an overlapping time window are positively correlated; properly accounting for that correlation in the weighting matrix neutralizes the apparent positive effect that survives in the diagonal-weighted GMM.

Taken together, the unbiased estimators in Table 5 deliver a sobering message: the choice of estimator is not a technical detail in staggered DiD designs but can determine whether the central conclusion of a study is positive, negative, or null. Our results raise more questions than they answer, and we recommend that future studies systematically check the sensitivity of their estimates to the choice of estimator.

5.2 Castle-doctrine laws and homicide (Cheng and Hoekstra, 2013)

For our second application, we replicate Cheng and Hoekstra (2013), who study the state-level staggered rollout of 'castle doctrine' laws—which expand the legal justification for the use of lethal force in self-defense—on homicide rates in the United States between 2000 and 2010. The baseline specification is

$$Y_{st} = \alpha_s + \lambda_t + \beta \, \text{Post}_{st} + \varepsilon_{st}, \tag{28}$$

where Y_{st} is the log homicide rate per 100,000 population for state s at year t, α_s and λ_t are state and year fixed effects, and standard errors are clustered at the state level. Following Table 5, Panel B, Column 1 of Cheng and Hoekstra (2013), the original unweighted OLS estimate is $\hat{\beta} \approx 0.088$ (0.064), implying that the law reform increased homicides by approximately 8%. Our replication using the binary indicator recovers $\hat{\beta} = 0.0694$ (0.0334) on 550 state–year observations.

Table 6: TWFE estimate and Goodman–Bacon decomposition, Cheng and Hoekstra (2013)

TWFE estimate	
D_{post}	0.0694
	(0.0334)
Goodman–Bacon decomposition	
Earlier vs. later-treated	
Weight share	0.0771
Avg. 2×2 estimate	−0.0286
Later vs. earlier-treated	
Weight share	0.0241
Avg. 2×2 estimate	0.0456
Treated vs. untreated	
Weight share	0.8988
Avg. 2×2 estimate	0.0784
Observations	550
States	50
Years	11

Notes: TWFE coefficient on the post-reform indicator and Goodman-Bacon (2021) decomposition of the underlying 2×2 DiD components. Standard errors clustered by state in parentheses.

Table 6 reports the TWFE estimate and its Goodman–Bacon decomposition. The setting features 29 never-treated states and clean treated-vs.-untreated comparisons account for 89.9% of the TWFE coefficient, indicating that forbidden comparisons play only a minor role—in stark contrast to Beck et al. (2010).

Table 7: Estimated effects of castle-doctrine laws on log homicide rate, Cheng and Hoekstra (2013)

Estimator	ATT	SE
TWFE	0.0694*	(0.0334)
CS	0.1104*	(0.0387)
Sun–Abraham	0.1104*	(0.0534)
Gardner	0.0669	(0.0570)
Flexible TWFE	0.1007*	(0.0366)
GMM	0.0856	(0.0780)

Notes: * denotes significance at the 5% level. Standard errors in parentheses. All estimators include state and year fixed effects.

Table 7 re-estimates (28) using the GMM estimator and the heterogeneity-robust alternatives. Consistent with the Goodman–Bacon diagnostic of low forbidden-comparison content, all estimators agree in sign and are broadly consistent in magnitude, pointing to a positive effect of the reform on homicide rates. CS, SA, and Flexible TWFE yield significant estimates of approximately 0.10, while Gardner and GMM are insignificant but positive and closer to the original pooled TWFE. The broad consensus is consistent with the decomposition: when identifying variation is dominated by clean comparisons, the heterogeneity-robust estimators converge—a result that does not hold for Beck et al. (2010).

6 Conclusion

We have addressed the identification challenge in staggered difference-in-differences designs: contamination of the ATT by forbidden comparisons that arise when an already-treated cohort serves as a control under treatment-effect heterogeneity. The GMM framework developed in this paper accommodates this bias through a modified incidence matrix that incorporates all 2×2 DiD comparisons, and the resulting estimator is unbiased, consistent, asymptotically normal, and—under the optimal weighting matrix—efficient under arbitrary within-unit serial correlation. By absorbing rather than discarding the information in already-treated controls, the framework recovers efficiency gains that the existing literature leaves unrealized. The framework accommodates unbalanced panels, does not require a never-treated cohort, and extends naturally to covariate-adjusted settings via outcome regression, inverse probability weighting, or doubly robust estimation.

Several extensions are natural avenues for future work. The current framework assumes irreversible treatment; allowing for reversible treatment, in the spirit of De Chaisemartin and d'Haultfoeuille (2020), would broaden applicability. Implementation is computationally more involved than for several alternatives, and we are developing an R package for general use; ports to Python and Stata would similarly broaden adoption. The empirical applications underscore that, in canonical datasets, the choice of method is not a refinement but can be the difference

between opposite conclusions, and we recommend systematic sensitivity analysis as standard practice in staggered DiD studies.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies 72*(1), 1–19.

Abbring, J. H. and G. J. Van den Berg (2003). The nonparametric identification of treatment effects in duration models. *Econometrica 71*(5), 1491–1517.

Athey, S. and G. W. Imbens (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics 226*(1), 62–79.

Baker, A. C., D. F. Larcker, and C. C. Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics 144*(2), 370–395.

Beck, T., R. Levine, and A. Levkov (2010). Big bad banks? the winners and losers from bank deregulation in the united states. *The Journal of Finance 65*(5), 1637–1667.

Borusyak, K. and X. Jaravel (2017). Consistency and inference in bartik research designs. Technical report, Technical Report, Working paper 2017. 187 and,"Revisiting event study designs.

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies 91*(6), 3253–3285.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of econometrics 225*(2), 200–230.

Cheng, C. and M. Hoekstra (2013). Does strengthening self-defense law deter crime or escalate violence? evidence from expansions to castle doctrine. *Journal of Human Resources 48*(3), 821–854.

De Chaisemartin, C., X. d'Haultfoeuille, F. Pasquier, D. Sow, and G. Vazquez-Bare (2022). Difference-in-differences estimators for treatments continuously distributed at every period. *arXiv preprint arXiv:2201.06898*.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review 110*(9), 2964–2996.

De Chaisemartin, C. and X. d'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The econometrics journal 26*(3), C1–C30.

Gardner, J., N. Thakral, L. T. To, and Y. Luther (2024). Two-stage differences in differences. *https://jrgcmu.github.io/2sdd$_g$tty.pdf*.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of econometrics 225*(2), 254–277.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics 14*(3), 262–280.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Sianesi, B. (2004). An evaluation of the swedish system of active labor market programs in the 1990s. *Review of Economics and statistics 86*(1), 133–155.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics 225*(2), 175–199.

Wooldridge, J. M. (2025). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Empirical Economics 69*(5), 2545–2587.

Appendix

A Forbidden Comparisons: A Worked Example

To fix ideas, consider a setting with three time periods and three cohorts. The first cohort ($g = \infty$) consists of never-treated units, the second ($g = 2$) is treated in period 2, and the third ($g = 3$) is treated in period 3. The cohort-time mean outcomes Y_{gt} take the values

$$\{Y_{\infty 1}, Y_{\infty 2}, Y_{\infty 3}\} = \{100, 100, 100\},$$
$$\{Y_{21}, Y_{22}, Y_{23}\} = \{110, 130, 125\},$$
$$\{Y_{31}, Y_{32}, Y_{33}\} = \{140, 140, 165\}.$$

Under parallel trends, the true CATTs are $\beta_{2,2} = 20$, $\beta_{2,3} = 15$, and $\beta_{3,3} = 25$. There are six possible 2×2 DiD estimates that can be used to estimate these three CATTs. Figure 1 provides a graphical illustration.

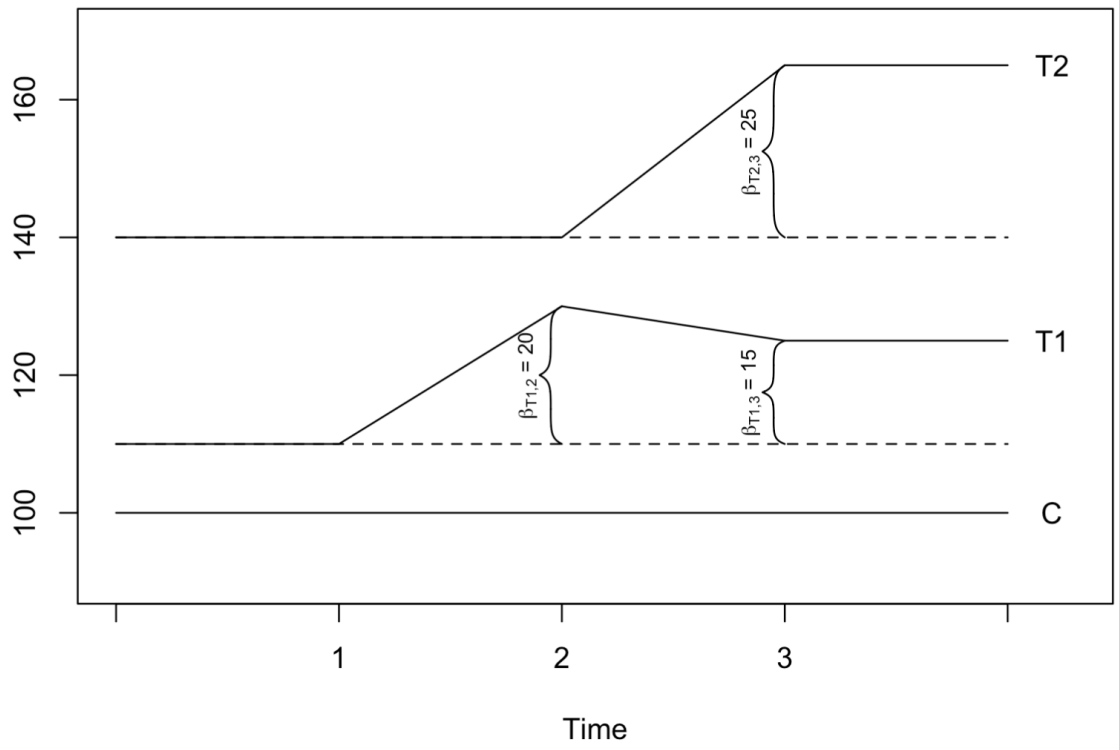

Figure 1: Example of staggered treatment adoption with heterogeneous effects. The horizontal axis represents time and the vertical axis the average outcome by cohort. The solid line "C" represents the never-treated cohort, "T1" the earlier-treated cohort, and "T2" the later-treated cohort. The CATTs of interest—$\beta_{T1,2}$, $\beta_{T1,3}$, and $\beta_{T2,3}$—are labeled with curly brackets. Dashed lines represent the counterfactual paths under no treatment for the treated groups.

The four DiDs that use the never-treated cohort ($g = \infty$) as the control are

$$\begin{aligned}
\hat{\beta}^{\infty 1}_{2,2} &: (Y_{22} - Y_{21}) - (Y_{\infty 2} - Y_{\infty 1}) = 20 = \beta_{2,2}, \\
\hat{\beta}^{\infty 2}_{2,3} &: (Y_{23} - Y_{21}) - (Y_{\infty 3} - Y_{\infty 1}) = 15 = \beta_{2,3}, \\
\hat{\beta}^{\infty 2}_{3,3} &: (Y_{33} - Y_{31}) - (Y_{\infty 3} - Y_{\infty 1}) = 25 = \beta_{3,3}, \\
\hat{\beta}^{\infty 1}_{3,3} &: (Y_{33} - Y_{32}) - (Y_{\infty 3} - Y_{\infty 2}) = 25 = \beta_{3,3}.
\end{aligned} \tag{29}$$

All four are unbiased. The DiD using the not-yet-treated cohort ($c = 3$) as the control,

$$\hat{\beta}^{31}_{2,2} : (Y_{22} - Y_{21}) - (Y_{32} - Y_{31}) = 20 = \beta_{2,2}, \tag{30}$$

is also unbiased. By contrast, the DiD using the already-treated cohort as the control,

$$\hat{\beta}^{21}_{3,3} : (Y_{33} - Y_{32}) - (Y_{23} - Y_{22}) = 30 \neq \beta_{3,3}, \tag{31}$$

is the forbidden comparison. It recovers $\beta_{3,3} - (\beta_{2,3} - \beta_{2,2})$ rather than $\beta_{3,3}$. The expected value is contaminated by the difference between the control cohort's CATTs at the post- and pre-period. In general,

$$\begin{aligned}
E[\hat{\beta}^{jm}_{g,g+k}] &= E\left[(Y_{g,g+k} - Y_{g,g-m}) - (Y_{g-j,g+k} - Y_{g-j,g-m})\right] \\
&= \beta_{g,g+k} - (\beta_{g-j,g+k} - \beta_{g-j,g-m}),
\end{aligned} \tag{32}$$

with the bias term $(\beta_{g-j,g+k} - \beta_{g-j,g-m})$ vanishing under homogeneous effects.[3]

To see the consequences for the aggregate ATT, define

$$\theta_E = w_{2,2}\beta_{2,2} + w_{2,3}\beta_{2,3} + w_{3,3}\beta_{3,3}, \tag{33}$$

with $w_{2,2} + w_{2,3} + w_{3,3} = 1$. The CATT estimates take the form

$$\begin{aligned} \hat{\beta}_{2,2} &= w_{2,2}^{\infty 1}\hat{\beta}_{2,2}^{\infty 1} + w_{2,2}^{31}\hat{\beta}_{2,2}^{31}, \\ \hat{\beta}_{2,3} &= \hat{\beta}_{2,3}^{\infty 2}, \\ \hat{\beta}_{3,3} &= w_{3,3}^{\infty 2}\hat{\beta}_{3,3}^{\infty 2} + w_{3,3}^{\infty 1}\hat{\beta}_{3,3}^{\infty 1} + w_{3,3}^{21}\hat{\beta}_{3,3}^{21}, \end{aligned} \tag{34}$$

with $w_{2,2}^{\infty 1} + w_{2,2}^{31} = 1$ and $w_{3,3}^{\infty 2} + w_{3,3}^{\infty 1} + w_{3,3}^{21} = 1$. Substituting (34) into (33) yields

$$\begin{aligned} \hat{\theta}_E &= w_{2,2}\hat{\beta}_{2,2} + w_{2,3}\hat{\beta}_{2,3} + w_{3,3}\hat{\beta}_{3,3} \\ &= w_{2,2}(w_{2,2}^{\infty 1}\hat{\beta}_{2,2}^{\infty 1} + w_{2,2}^{31}\hat{\beta}_{2,2}^{31}) + w_{2,3}\hat{\beta}_{2,3}^{\infty 2} + w_{3,3}(w_{3,3}^{\infty 2}\hat{\beta}_{3,3}^{\infty 2} + w_{3,3}^{\infty 1}\hat{\beta}_{3,3}^{\infty 1} + w_{3,3}^{21}\hat{\beta}_{3,3}^{21}). \end{aligned}$$

Taking expectations and using (32),

$$E[\hat{\theta}_E] = (w_{2,2} + w_{3,3}w_{3,3}^{21})\beta_{2,2} + (w_{2,3} - w_{3,3}w_{3,3}^{21})\beta_{2,3} + w_{3,3}\beta_{3,3}.$$

The implicit weight on $\beta_{2,3}$ is shifted to $\beta_{2,2}$ by the amount $w_{3,3}w_{3,3}^{21}$, so the bias is

$$\theta_E - E[\hat{\theta}_E] = -w_{3,3}w_{3,3}^{21}(\beta_{2,2} - \beta_{2,3}).$$

The bias is a function of the aggregation weights and the CATTs, and vanishes under homogeneous effects. In severe cases the implicit weight on $\beta_{2,3}$ can become negative.

B Constructing Q_H: A Worked Example

Consider the three-cohort, three-period design of Appendix A. The six 2×2 DiD estimates are summarized in Table 8. We collect them into the vector $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_6)^\top$, ordered by control-group type and then by pre-period m.

[3] The original draft writes $\beta_{g-j,g+k-m}$ for the second term in the bias, which is a typo: the correct expression is $\beta_{g-j,g-m}$, the CATT of the already-treated control evaluated at the pre-period reference date.

Table 8: Summary of the six 2×2 DiD estimates from the example.

	2×2 DiD	g	k	Control	m	Formula
Δ_1	$\hat{\beta}^{\infty 1}_{2,2}$	$g=2$	$k=0$	never-treated ($c=\infty$)	1	$(Y_{22}-Y_{21})-(Y_{\infty 2}-Y_{\infty 1})=20$
Δ_2	$\hat{\beta}^{\infty 1}_{2,3}$	$g=2$	$k=1$	never-treated ($c=\infty$)	1	$(Y_{23}-Y_{21})-(Y_{\infty 3}-Y_{\infty 1})=15$
Δ_3	$\hat{\beta}^{\infty 2}_{3,3}$	$g=3$	$k=0$	never-treated ($c=\infty$)	2	$(Y_{33}-Y_{31})-(Y_{\infty 3}-Y_{\infty 1})=25$
Δ_4	$\hat{\beta}^{\infty 1}_{3,3}$	$g=3$	$k=0$	never-treated ($c=\infty$)	1	$(Y_{33}-Y_{32})-(Y_{\infty 3}-Y_{\infty 2})=25$
Δ_5	$\hat{\beta}^{31}_{2,2}$	$g=2$	$k=0$	not-yet-treated ($c=3$)	1	$(Y_{22}-Y_{21})-(Y_{32}-Y_{31})=20$
Δ_6	$\hat{\beta}^{21}_{3,3}$	$g=3$	$k=0$	already-treated ($c=2$)	1	$(Y_{33}-Y_{32})-(Y_{23}-Y_{22})=30 \neq \beta_{3,3}$

Notes: The original draft uses superscript 0 for the never-treated control. We use ∞ throughout for consistency with the main text.

Under heterogeneous effects, the forbidden comparison Δ_6 satisfies $E[\hat{\beta}^{21}_{3,3}] = \beta_{3,3} - (\beta_{2,3} - \beta_{2,2})$. All other rows are unbiased, so rows 1–5 of Q_H coincide with Q. The incidence matrix is

$$Q_H = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix}}_{6\times 3} \begin{matrix} \leftarrow \Delta_1 \\ \leftarrow \Delta_2 \\ \leftarrow \Delta_3 \\ \leftarrow \Delta_4 \\ \leftarrow \Delta_5 \\ \leftarrow \Delta_6 \text{ (corrected)} \end{matrix}$$

The entry $+1$ in the sixth row, $\beta_{2,2}$ column absorbs the contaminating pre-period CATT of the already-treated control ($\beta_{g-j,g-m} = \beta_{2,2}$). The entry -1 in the sixth row, $\beta_{2,3}$ column removes the contaminating post-period CATT of the same control ($\beta_{g-j,g+k} = \beta_{2,3}$). Together, the two entries debias Δ_6 and restore $E[\Delta_6 - (Q_H\boldsymbol{\beta})_6] = 0$.

Verification.

$$Q_H\boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 20 \\ 15 \\ 25 \end{pmatrix} = \begin{pmatrix} 20 \\ 15 \\ 25 \\ 25 \\ 20 \\ 20 - 15 + 25 \end{pmatrix} = \begin{pmatrix} 20 \\ 15 \\ 25 \\ 25 \\ 20 \\ 30 \end{pmatrix} = E[\boldsymbol{\Delta}].$$

C An Outcome-Level Alternative for the Weighting Matrix

Let Y be the $NT \times 1$ column vector of outcomes at the observation level. We can write the moment vector Δ as

$$\Delta = RY, \tag{35}$$

where R is a $\sum_g \sum_k N_{g,g+k} \times NT$ incidence matrix whose entries take the value $1/N_g$ or $-1/N_g$ depending on how $Y_{g,g+k}$ enters the formula for $\Delta^s_{g,g+k}$, and zero otherwise. As an illustration,

$$\begin{aligned}\Delta^1_{3,3} = \hat{\beta}^{\infty 1}_{3,3} &= (Y_{33} - Y_{32}) - (Y_{\infty 3} - Y_{\infty 2}) \\ &= \left(\tfrac{1}{N_3}\sum_{i\in 3} Y_{i33} - \tfrac{1}{N_3}\sum_{i\in 3} Y_{i32}\right) - \left(\tfrac{1}{N_\infty}\sum_{i\in\infty} Y_{i\infty 3} - \tfrac{1}{N_\infty}\sum_{i\in\infty} Y_{i\infty 2}\right),\end{aligned}$$

which we can write as $\Delta^1_{3,3} = RY$ with the appropriate row of R.[4] Modeling Y via the standard TWFE specification,

$$Y = \alpha_i + \lambda_t + \theta D + \nu, \tag{36}$$

where ν has $E[\nu] = 0$ and $\mathrm{Var}[\nu] = \Omega_\nu$, we obtain

$$\mathrm{Var}[\Delta] = \mathrm{Var}[RY] = R\,\Omega_\nu\,R'.$$

Under cross-sectional independence, Ω_ν is block-diagonal with $T \times T$ blocks Ω_i describing within-unit serial dependence:

$$\Omega_\nu = \begin{bmatrix} \Omega_1 & & \\ & \ddots & \\ & & \Omega_N \end{bmatrix}.$$

Two strategies for estimating Ω_ν are useful in practice. Under an AR(1) model $\nu_{it} = \rho\nu_{i,t-1} + u_{it}$ with $\mathrm{Var}[\nu_{it}] = \sigma^2$ and $E[u_{it}] = 0$, we have $\mathrm{Cov}(\nu_t, \nu_{t-p}) = \rho^p\sigma^2$ and

$$\Omega_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots \\ \rho & 1 & \rho & \cdots \\ \rho^2 & \rho & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Estimating Ω_ν thus reduces to estimating the single parameter ρ: estimate (36) by OLS, recover $\hat{\nu}_{it}$, regress $\hat{\nu}_{it}$ on $\hat{\nu}_{i,t-1}$, and obtain $\hat{\rho}$. Alternatively, one can estimate Ω_ν without imposing structure via the following iterative procedure. Estimate (36) by OLS to obtain $\hat{\nu}$. Form the unit-level outer-product estimator

$$\hat{\Omega}_i = \begin{bmatrix} \hat{\nu}_1^2 & \hat{\nu}_1\hat{\nu}_2 & \cdots & \hat{\nu}_1\hat{\nu}_T \\ \hat{\nu}_2\hat{\nu}_1 & \hat{\nu}_2^2 & \cdots & \hat{\nu}_2\hat{\nu}_T \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\nu}_T\hat{\nu}_1 & \hat{\nu}_T\hat{\nu}_2 & \cdots & \hat{\nu}_T^2 \end{bmatrix},$$

[4] The original draft writes $\Delta^1_{3,3} = \beta^{\infty 1}_{3,3}$, equating a sample object with a population parameter. We interpret this as a notational shortcut: the displayed object is $\hat{\beta}^{\infty 1}_{3,3}$, a particular linear combination of Y.

estimate the GLS analog

$$\hat{\theta} = (Q_H' R \hat{\Omega}_\nu^{-1} R' Q_H)^{-1} Q_H' R \hat{\Omega}_\nu^{-1} R' \Delta,$$

update the residuals $\hat{\nu} = Y - \hat{\alpha}_i - \hat{\lambda}_t - \hat{\theta} D$, and iterate until $\hat{\beta}$ stabilizes. The eventual $\hat{\Omega}_\nu$ is the estimated unit-level covariance matrix, and $\widehat{\text{Var}}[\Delta] = R \hat{\Omega}_\nu R'$.