Ashoka University
Economics Discussion Paper 166

A Bayesian Approach to Partial Homogeneity in Staggered Difference-in-Difference

July 2026

Parush Arora, Ashoka University
Rohan Wagle, Ashoka University

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

A Bayesian Approach to Partial Homogeneity in Staggered Difference-in-Difference

Parush Arora* and Rohan Wagle*

Abstract

In staggered difference-in-differences (DiD) designs, units enter treatment at different calendar times, so the treatment effect is not a single number but a set of Cohort-Average Treatment effects on the Treated (CATTs), one per cohort-time cell. Estimating every CATT as its own parameter, as the standard fully flexible estimator does, is unbiased but inefficient when some of these effects are in fact equal, whereas pooling them all into a single two-way fixed effects (TWFE) coefficient is efficient but biased whenever the heterogeneity is genuine. We frame the choice between these extremes as a partition-selection problem on the cohort-time cells and address it with a Dirichlet Process (DP) mixture prior on the CATTs. The model favors parsimonious groupings without fixing their number, and a collapsed Gibbs sampler delivers point estimates, credible intervals that marginalize the unknown partition, and co-clustering probabilities for every pair of CATTs. With the error variance held fixed, the model's maximum a posteriori (MAP) partition reduces to an ℓ_0-penalized regression, connecting the Bayesian formulation to the homogeneity-pursuit literature. In a calibrated simulation, the model cuts the sampling variance of the cohort-time effects by 26–52% relative to the fully flexible estimator, without the pooled estimator's bias, provided the distinct effects are separated enough to be recovered, and the posterior delivers near-nominal confidence-interval coverage by averaging over the unknown partition. In two applications the method recovers a precision-improving partial-homogeneity structure in one, where the cohort-time effects are genuinely heterogeneous, and reports that full pooling is adequate in the other, where they are not.

Keywords: difference-in-differences, staggered treatment, TWFE, model specification, Bayesian estimation, Gibbs sampler, treatment effect heterogeneity, causal inference.

JEL Codes: C11, C14, C23, C52.

*Department of Economics, Ashoka University.

1 Introduction

Difference-in-differences (DiD) has become one of the workhorses of empirical economics. In its canonical form, a treatment cohort and a control cohort are observed before and after a single policy change, and the researcher estimates a single average treatment effect (ATT or average treatment effect on treated). In a simple setting with two time periods, ATT is recovered by taking the difference between the change in the average outcome over time for the treatment and control cohorts, commonly referred to as a simple 2×2 DiD estimate. In a general setting with N observations and T time periods, the most popular way to estimate the ATT with a binary treatment indicator is the two-way fixed effects regression (TWFE):

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

where Y_{it} is the observed outcome for i^{th} unit in time t, α_i is the unit fixed effect, λ_t is the time fixed effect, D_{it} is the treatment dummy and τ is the ATT. We will refer to Eq. 1 as pooled TWFE. The pooled TWFE estimator ($\hat{\tau}_{\text{pool}}$) is directly related to the 2×2 DiD estimator, as it can be represented as a weighted average of all possible 2×2 DiD estimates (Goodman-Bacon, 2021).

Modern policy evaluations, however, rarely fit this clean template. Minimum wage increases, Medicaid expansions, trade liberalizations, and school reforms typically roll out across states, cities, or firms in a staggered fashion, with different units adopting treatment at different calendar times. We update the notation so that Y_{igt} and D_{igt} denote the outcome and treatment indicator for unit i in cohort g at time t. The estimator $\hat{\tau}_{\text{pool}}$ is unbiased in staggered timing if the treatment effect for each $g \times t$ case (cohort-time specific ATT or CATT) is homogeneous.

The econometrics of staggered DiD has seen remarkable progress in recent years. de Chaisemartin and D'Haultfœuille (2020), Callaway and Sant'Anna (2021), Sun and Abraham (2021), Goodman-Bacon (2021), Gardner (2022), and Borusyak et al. (2024), among others, have shown that the TWFE regression can produce a biased estimate of the ATT when there is heterogeneity among CATTs. de Chaisemartin and D'Haultfœuille (2023) and Roth et al. (2023) survey this literature. The bias arises due to the inclusion of an already-treated cohort as a control, which violates the parallel trends assumption. Goodman-Bacon (2021) showed that the decomposition of TWFE can contain negative weights when the treatment effects are heterogeneous. The recommended solution is to either drop the already-treated cohorts from the control or estimate the CATTs one coefficient per (cohort-time) cell and then aggregate them as desired. For the latter, Wooldridge (2025) discussed the flexible TWFE that allows interaction between the treatment dummy and the cohort-time dummy and recovers unbiased CATTs. Their estimate is identical to Gardner (2022) and Borusyak et al. (2024) in certain settings and more efficient than Callaway and Sant'Anna (2021) and Sun and Abraham (2021), which drop the already-treated cohorts from the control. These heterogeneity-robust estimators are efficient only under spherical errors. Arora and Bijani (2026) show that efficiency under arbitrary within-unit serial correlation is attainable through an optimally weighted GMM in the space of clean comparisons.

This fully flexible approach solves the bias problem but introduces a new challenge, namely model specification. With G treated cohorts and T time periods, there can be $K = \sum_g (T - g + 1)$ distinct CATT cells, where $g = 0$ represents the never-treated cohort and $g = 2, 3, \ldots, T$ indexes the period in which the cohort is first treated. Estimating all K separately is unbiased, but may be inefficient if some CATTs share the same true effect. Conversely, pooling all CATTs into a single parameter (classical TWFE) is efficient but biased when effects genuinely differ across cohorts or periods.

The specification question is which CATTs can be pooled, and which must be kept separate? This is a model selection problem on the space of partitions of the K CATT cells. Getting the partition right matters, because under the true partition, the estimator is both unbiased and more efficient than the fully flexible approach, exploiting the additional identifying restriction that some effects are equal without imposing false restrictions on others.

Although the recent staggered-DiD literature has focused overwhelmingly on removing the bias of pooled TWFE, the complementary question of how to efficiently pool the resulting cohort-time effects has received comparatively little attention. The problem of grouping otherwise distinct coefficients that share a common value is, however, well studied in the broader panel-data and statistics literatures, and our approach draws on established machinery from both traditions. On the penalized-regression side, penalizing pairwise differences to recover latent groups underlies the classifier-Lasso of Su et al. (2016) and the homogeneity-pursuit estimators of Ke et al. (2015) and Wang et al. (2018), as well as the grouping-pursuit surface of Shen and Huang (2010). Related ideas appear in the grouped fixed-effects approach of Bonhomme and Manresa (2015) and in work on determining the number of latent groups (Lu and Su, 2017). Relative to these ℓ_1-type fusion penalties, we use an ℓ_0 penalty because it produces exact equality of coefficients rather than continuous shrinkage. On the Bayesian side, clustering coefficients through a Dirichlet Process mixture or an explicit partition prior is standard in Bayesian nonparametrics (Ferguson, 1973; Antoniak, 1974; Hartigan, 1990; Barry and Hartigan, 1992), and Bayesian nonparametric methods have also been applied directly to causal-effect estimation (Hill, 2011). Our contribution is to bring these tools to bear on the specification problem in staggered DiD, to characterize the resulting bias-variance and identification trade-offs in that setting, and to connect the two traditions, since the ℓ_0 fusion penalty is precisely the MAP of the DP model when the error variance is held fixed.

Our restriction is also related to, but distinct from, recent work that bounds the magnitude of treatment-effect heterogeneity rather than its structure. Kwon and Sun (2026) restrict the variance of conditional average treatment effects and conduct bias-aware sensitivity analysis over that bound, shrinking heterogeneous effects smoothly toward a common value; Armstrong et al. (2025) adapt to misspecification of a restricted model, and Armstrong and Kolesár (2018) give the underlying optimal-inference theory; in a similar honest-inference spirit but for a different DiD assumption, Rambachan and Roth (2023) allow bounded violations of parallel trends. Where these papers impose a continuous bound and never assert that any two effects are exactly equal, we impose a discrete partial-homogeneity restriction and attempt to recover the grouping from the data. The

two views are complementary, in that the continuous bound is agnostic about which effects coincide, whereas the discrete view is informative about the grouping structure itself, at the cost of a stronger assumption whose selection uncertainty must be confronted (a point we return to below and in the simulations).

This paper makes three contributions. First, we formalize the specification problem in staggered DiD as a partition-selection problem and characterize the bias and variance of the fully flexible, partially homogeneous, and fully pooled estimators under a partial-homogeneity data-generating process. Second, we propose a Bayesian solution, a Dirichlet Process (DP) mixture prior over the CATTs that assigns positive probability to every partition while favoring parsimonious groupings. A collapsed Gibbs sampler estimates the model, and the posterior delivers point estimates, credible intervals that marginalize the unknown partition, and co-clustering probabilities for every pair of CATTs. Third, we record a connection to the penalized-regression literature. Holding the error variance fixed, the model's maximum a posteriori (MAP) partition is an ℓ_0-penalized least-squares estimator whose penalty equals the model's log prior, and a BIC-type estimator when the variance is free, which links the Bayesian formulation to ℓ_0 homogeneity pursuit.

The DP prior is the standard nonparametric choice for models in which the number of components is unknown (Ferguson, 1973; Antoniak, 1974). Escobar and West (1995) show how the DP mixture models can be used for the density estimation and regression with an unknown number of mixture components. Neal (2000) develops efficient MCMC algorithms for posterior inference, including the collapsed Gibbs sampler that allows for efficient drawing from the posterior (which we employ). Müller and Quintana (2004) provide a survey of nonparametric Bayesian methods for applied researchers. Miller and Harrison (2018) study the finite mixture model with a prior on the number of components, which shares the spirit of our approach.

The remainder of the paper is organized as follows. Section 2 formalizes the specification problem, establishes the bias-variance trade-off under partial homogeneity. Section 3 develops the Dirichlet Process model, covering its specification, estimation, and inference, and records the ℓ_0-penalized estimator that coincides with its MAP partition. Section 4 reports the simulation results, Section 5 presents two empirical applications, and Section 6 concludes.

2 The Framework

2.1 Setup and Assumptions

Consider a balanced panel with N units observed over T time periods. We are interested in estimating the effect of a binary treatment, $D_{it} \in \{0, 1\}$, on the outcome variable, $Y_{igt} \in \mathbb{R}$. Assume we observe the sample $\{Y_{ig1}, \ldots, Y_{igT}, D_{ig1}, \ldots, D_{igT}\}$ where $i = 1, \ldots, N$ indexes units and $t = 1, \ldots, T$ indexes time. Units belong to one of g treated cohorts (indexed by

treatment entry date $g \in \{2, 3, \ldots, T\} = \mathcal{G}$) or a never-treated group ($g = 0$). Period 1 is assumed to be pre-treatment for all cohorts, which means $g \neq 1$. Let N_g be the number of units in the cohort g and thus $\sum_g N_g = N$. We make the following assumptions about the treatment process:

Assumption 1 (Random Sample): $\{Y_{ig1}, \ldots, Y_{igT}, D_{i1}, \ldots, D_{iT}\}_{i=1}^{N}$ are independent and identically distributed. This assumption also implies that we have a balanced panel.

Assumption 2 (Irreversibility of Treatment): It implies that once a unit is treated, that unit remains treated in the next period. Units do not "forget" about the treatment experience. If $D_{i,t-1} = 1$ then $D_{i,t} = 1$.

Assumption 3 (Parallel Trends): It implies that the expectation of Y_{igt} without the treatment follows the same evolution over time in each g. This is equivalent to assuming that $E[Y_{igt}|D_{it} = 0] - E[Y_{ig,t-1}|D_{i,t-1} = 0]$ does not vary with g. We are making this assumption at the unit level (but our estimator is valid under the assumption of parallel trends at the cohort level as well). The assumption may hold either unconditionally or only after conditioning on covariates. We allow the conditional version, which accommodates (1) covariate-specific trends in Y_{igt} and (2) different distributions of covariates across cohorts, without requiring it.

Assumption 4 (No Treatment Anticipation): There is no treatment effect in pre-treatment periods. The outcome Y_{igt} in any period before g is, on average, equal to the untreated outcome.

2.2 The Specification Problem

Under homogeneous effects the CATTs all equal the ATT and $\hat{\tau}_{\text{pool}}$ is unbiased. Under heterogeneity they differ, and the target ATT is a weighted average of the CATTs,

$$\tau = \sum_g \sum_t w_{gt}\, \tau_{gt}, \tag{2}$$

where τ_{gt} is the CATT for cohort g at calendar time t and Callaway and Sant'Anna (2021) discusses the choice of weights. As reviewed in Section 1, $\hat{\tau}_{\text{pool}}$ is then biased for τ, because already-treated units enter as controls (the forbidden comparison) and some weights in the Goodman-Bacon (2021) decomposition turn negative. The literature offers several unbiased alternatives, restricting the controls to never- or not-yet-treated units (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021), allowing treatment to switch on and off (de Chaisemartin and D'Haultfœuille, 2020), or estimating one coefficient per cohort-time cell through the flexible and imputation estimators of Wooldridge (2025), Gardner (2022), Borusyak et al. (2024), and Liu et al. (2024), which are unbiased and efficient under spherical errors, with Arora and Bijani (2026) extending efficiency to arbitrary within-unit

serial correlation. We build on the flexible TWFE of Wooldridge (2025), which interacts the treatment dummy with the cohort-time dummies,

$$Y_{igt} = \alpha_i + \lambda_t + \sum_{g \in \mathcal{G}} \sum_{t \geq g} \tau_{gt} D_{igt} + \varepsilon_{it}, \tag{3}$$

and recovers unbiased CATTs $\hat{\tau}_{gt}$, from which the ATT is estimated as $\hat{\tau}_{\text{flex}} = \sum_g \sum_t w_{gt} \hat{\tau}_{gt}$.

Let Y be the $NT \times 1$ vector of outcomes and $D = [D_1, \ldots, D_K]$ be the $NT \times K$ matrix of cohort-time dummies. Let $\tilde{Y} = Y - \bar{Y}_i - \bar{Y}_t + \bar{Y}$ be the vector of within-transformed outcomes and $\tilde{D} = [\tilde{D}_1, \ldots, \tilde{D}_K]$ be the matrix of within-transformed cohort-time dummies where $\tilde{D}_k = D_k - \bar{D}_{ki} - \bar{D}_{kt} + \bar{D}_k$. Here, $\bar{Y}_i$ and $\bar{D}_{ki}$ are the unit-specific averages over time, $\bar{Y}_t$ and $\bar{D}_{kt}$ are the time-specific averages over units and $\bar{Y}$ and $\bar{D}_k$ are the overall averages. Recall the within-transformed model obtained after double-demeaning to partial out unit and time fixed effects:

$$\tilde{Y} = \tilde{D}\,\tau + \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \sim \big(\mathbf{0},\, \sigma^2 I_{NT}\big), \tag{4}$$

where $\tau = (\tau_1, \ldots, \tau_K)'$ collects all K CATTs. Ordinary least squares on (4) yields the fully flexible estimator

$$\hat{\tau}_{\text{flex}} = \big(\tilde{D}'\tilde{D}\big)^{-1}\tilde{D}'\tilde{Y} = [\hat{\tau}_{1,\text{flex}}, \ldots, \hat{\tau}_{K,\text{flex}}]' \qquad \text{Var}\big(\hat{\boldsymbol{\tau}}_{\text{flex}}\big) = \sigma^2\big(\tilde{D}'\tilde{D}\big)^{-1}. \tag{5}$$

The k^{th} diagonal element of $\sigma^2[(\tilde{D}'\tilde{D})^{-1}]$ is the sampling variance of the k^{th} CATT estimate. Because each CATT is estimated from the variation in its own within-transformed indicator, this variance is governed by how much identifying variation CATT k possesses individually. In large panels with many cohorts and long post-treatment windows, K is large and the variation attributable to any single CATT cell is correspondingly modest, leading to imprecise estimates even when the overall sample size NT is large.

The opposite extreme imposes a single common coefficient τ for every cohort-time cell. Writing $\tilde{D}_{\text{pool}} = \sum_{k=1}^{K} \tilde{D}_k$ for the aggregate within-transformed treatment indicator, the fully pooled estimator is

$$\hat{\tau}_{\text{pool}} = \big(\tilde{D}'_{\text{pool}}\tilde{D}_{\text{pool}}\big)^{-1}\tilde{D}'_{\text{pool}}\tilde{Y}, \qquad \text{Var}\big(\hat{\tau}_{\text{pool}}\big) = \frac{\sigma^2}{\big\|\tilde{D}_{\text{pool}}\big\|^2}. \tag{6}$$

Since $\tilde{D}_{\text{pool}}$ aggregates the variation of all K cells, $\|\tilde{D}_{\text{pool}}\|^2 > \|\tilde{D}_k\|^2$ for any individual k, and the pooled estimator is correspondingly precise. The cost is bias. The pooled estimator is consistent for a variance-weighted average of the true CATTs, $\sum_{k=1}^{K} w_k \tau_k^*$, not for any individual τ_k^*. Formally, the bias of $\hat{\tau}_{\text{pool}}$ as an estimator of τ_k^* is

$$\mathbb{E}\big[\hat{\tau}_{\text{pool}}\big] - \tau_k^* = \sum_{j=1}^{K} w_j\big(\tau_j^* - \tau_k^*\big), \qquad w_j = \frac{\|\tilde{D}_j\|^2}{\sum_{\ell=1}^{K}\|\tilde{D}_\ell\|^2}, \tag{7}$$

where we use the approximation that the within-transformed dummies are nearly orthogonal across distinct cohort-time cells (an approximation in balanced panels, since

double-demeaning leaves small nonzero cross-cell products). The bias in (7) is zero only when all true CATTs are equal, precisely the case in which the pooled model is correctly specified. In any other case, the pooled estimator distorts every cohort-specific treatment effect toward a global average, rendering it useless for heterogeneity-aware policy analysis. This aggregation bias is distinct from the Goodman–Bacon negative-weight bias of pooled TWFE, in which already-treated units serve as controls (the forbidden comparison). The flexible starting point adopted here removes that negative-weight bias, so the specification problem we study concerns only the aggregation bias.

The flexible specification (3) is an ordinary linear regression, so covariate adjustment needs no new machinery. To accommodate the conditional version of Assumption 3, we augment it with covariates X_{igt}, time-varying controls or covariate-cohort interactions that permit covariate-specific trends,

$$Y_{igt} = \alpha_i + \lambda_t + X'_{igt}\beta + \sum_{g\in\mathcal{G}} \sum_{t\geq g} \tau_{gt}\, D_{igt} + \varepsilon_{it}. \tag{8}$$

By the Frisch–Waugh–Lovell theorem the CATT estimates are identical whether X enters the regression directly or is partialled out beforehand. Writing M for the residual-maker that projects off the unit and time fixed effects together with X, and redefining the within-transformed quantities as

$$\tilde{Y} = MY, \qquad \tilde{D}_k = MD_k, \tag{9}$$

which generalizes the double-demeaning of (4) to residualize on the covariates as well, the flexible estimator (5), the Bayesian model of Section 3, and the ℓ_0-PH estimator all apply verbatim with $\tilde{Y}$ and $\tilde{D}$ read as these covariate-residualized quantities. Covariate adjustment thus enters entirely through the within transformation and leaves the partition-selection problem unchanged.

2.3 Partial Homogeneity

Let the true parameter vector $\tau = (\tau_1, \ldots, \tau_K)'$ exhibit *partial homogeneity*: there is a partition $\mathcal{P} = \{C_1, \ldots, C_m\}$ of $\{1, \ldots, K\}$ into $m \leq K$ groups with

$$\tau_k = \phi_p \quad \text{for all } k \in C_p, \qquad p = 1, \ldots, m, \tag{10}$$

where the group effects $\phi_1, \ldots, \phi_m$ are distinct. The groups may have arbitrary sizes $|C_1|, \ldots, |C_m|$, and we call $\mathcal{P}$ the true partition. Partial homogeneity is the regime $m < K$, where the true model has only m distinct parameters and the fully flexible model is over-parameterized by $K - m$. Nothing below restricts the shape of $\mathcal{P}$; the efficiency gain accrues to every group C_p with $|C_p| \geq 2$, and a singleton group contributes none. (The special case of one large group together with singletons is the sharpest illustration but is not assumed.)

To estimate the correctly specified model under $\mathcal{P}$, define the group regressors $\tilde{D}_p = \sum_{k\in C_p} \tilde{D}_k$ and collect them into the $NT \times m$ matrix $\tilde{D}^* = [\tilde{D}_1, \ldots, \tilde{D}_m]$, where for a singleton

$C_p = \{k_p\}$ we have $\tilde{D}_p = \tilde{D}_{k_p}$. We assume a homoskedastic error and proceed with the restricted model

$$\tilde{Y} = \tilde{D}^* \phi + \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \sim \big(0,\, \sigma^2 I_{NT}\big), \tag{11}$$

whose OLS estimator is

$$\hat{\phi}_{\mathrm{PH}} = \big(\tilde{D}^{*\prime}\tilde{D}^*\big)^{-1}\tilde{D}^{*\prime}\tilde{Y}, \qquad \mathrm{Var}\big(\hat{\phi}\big) = \sigma^2\big(\tilde{D}^{*\prime}\tilde{D}^*\big)^{-1}. \tag{12}$$

The p-th diagonal element of $\sigma^2(\tilde{D}^{*\prime}\tilde{D}^*)^{-1}$ governs the precision of $\hat{\phi}_p$, the estimate of the common effect of group C_p. For any $k \in C_p$ the estimate of $\tau_k = \phi_p$ is $\hat{\phi}_p$, which pools the identifying variation of every cell in C_p, whereas the flexible estimator uses only the variation in $\tilde{D}_k$. This difference in the effective information set is the source of the variance gap. We call $\hat{\phi}_{\mathrm{PH}}$ the partial homogeneity (PH) estimator for the rest of the paper.

We now establish that, group by group, the PH estimator dominates the fully flexible estimator. Because both are unbiased for ϕ_p under the true model, the Gauss–Markov theorem determines which is more efficient.

Proposition 1 (PH efficiency dominance)**.** *Under the true partition $\mathcal{P}$, for every group C_p and every $k \in C_p$,*

$$\mathrm{Var}\big(\hat{\phi}_p\big) \;\le\; \mathrm{Var}\big(\hat{\tau}_{k,flex}\big), \tag{13}$$

with strict inequality whenever $|C_p| \ge 2$, and with equality for a singleton group.

Proof. Under $\mathcal{P}$ we have $\tau_k = \phi_p$ for $k \in C_p$, so $\hat{\tau}_{k,\mathrm{flex}}$ is a linear, unbiased estimator of ϕ_p. By the Gauss–Markov theorem, $\hat{\phi}_p$ is the best linear unbiased estimator (BLUE) of ϕ_p among all linear functions of $\tilde{Y}$, hence $\mathrm{Var}(\hat{\tau}_{k,\mathrm{flex}}) \ge \mathrm{Var}(\hat{\phi}_p)$. The inequality is strict whenever $\hat{\phi}_p \neq \hat{\tau}_{k,\mathrm{flex}}$ with positive probability, which holds precisely when $|C_p| \ge 2$, since $\hat{\phi}_p$ then pools the variation of every cell in C_p while $\hat{\tau}_{k,\mathrm{flex}}$ uses only that of $\tilde{D}_k$. When $|C_p| = 1$ the two estimators coincide. □

The result holds for every group at once, so the total efficiency gain is the sum of the within-group gains, one for each group of size at least two.

To make the variance reduction explicit, we work under the additional assumption that the within-transformed cohort-time dummies are mutually orthogonal, $\tilde{D}_j'\tilde{D}_k = 0$ for $j \neq k$, which holds approximately for balanced panels in which each unit belongs to exactly one cohort and which we impose exactly to obtain closed-form expressions. Under orthogonality, $\tilde{D}'\tilde{D} = \mathrm{diag}(n_1, \ldots, n_K)$ with $n_k = \|\tilde{D}_k\|^2$ the effective sample size for CATT k. The flexible variance is then

$$\mathrm{Var}\big(\hat{\tau}_{k,\mathrm{flex}}\big) = \frac{\sigma^2}{n_k}, \qquad k = 1, \ldots, K, \tag{14}$$

while $\|\tilde{D}_p\|^2 = \sum_{k \in C_p} n_k$, so the PH variance for group C_p is

$$\mathrm{Var}\big(\hat{\phi}_p\big) = \frac{\sigma^2}{\displaystyle\sum_{k \in C_p} n_k}. \tag{15}$$

Taking the ratio for any $k \in C_p$,

$$\frac{\mathrm{Var}(\hat{\phi}_p)}{\mathrm{Var}(\hat{\tau}_{k,\mathrm{flex}})} = \frac{n_k}{\sum_{j \in C_p} n_j} \;\leq\; 1, \tag{16}$$

strictly less than one whenever $|C_p| \geq 2$ and decreasing in the effective sample sizes of the other cells in the group. In the balanced case $n_k = n$ the ratio is $1/|C_p|$, so the PH estimator is exactly $|C_p|$ times more efficient than the flexible estimator for every cell in group C_p. The gain grows with group size, as pooling $|C_p|$ cells concentrates $|C_p|$ times as much effective sample size onto a single coefficient.

The fully pooled estimator lowers the variance further by imposing a single coefficient across all K cells, at the cost of bias in every group. Under Eq. 11 with orthogonal dummies its probability limit is

$$\mathbb{E}\big[\hat{\tau}_{\mathrm{pool}}\big] = \sum_{k=1}^{K} w_k \, \tau_k = \sum_{p=1}^{m} W_p \, \phi_p, \qquad w_k = \frac{n_k}{\sum_j n_j}, \quad W_p = \sum_{k \in C_p} w_k, \tag{17}$$

where W_p is the total weight of group C_p. For any cell $k \in C_p$ the bias is therefore

$$\mathbb{E}\big[\hat{\tau}_{\mathrm{pool}}\big] - \phi_p = \sum_{q \neq p} W_q \, \big(\phi_q - \phi_p\big), \tag{18}$$

a weight-averaged sum of the gaps between group p and every other group. It is non-zero whenever some other group differs ($\phi_q \neq \phi_p$) and does not vanish as $n \to \infty$. The signals of the other groups bleed into the pooled estimate, so ϕ_p is estimated as a distorted mixture of all m true effects.

The practical challenge is that $\mathcal{P}$ is unknown. A researcher who uses the fully flexible estimator by default leaves $K-m$ degrees of freedom on the table and reports unnecessarily wide intervals, while one who collapses to a single pooled coefficient introduces bias that cannot be detected without further structure. What is needed is a data-driven procedure that recovers $\mathcal{P}$, or a close approximation, from the data. The ℓ_0-penalized estimator and the Bayesian Dirichlet Process approach of the following sections provide two principled answers.[1]

3 A Bayesian Model for Partition Selection

We develop the Dirichlet Process (DP) mixture model for the CATTs, specifying it in Section 3.1, treating its error variance and the ℓ_0 form of its maximum a posteriori partition in Section 3.2, and estimating it with a collapsed Gibbs sampler in Section 3.3.

[1]Partial homogeneity also aids identification under limited overlap. A cohort-time cell with no clean comparison, and hence no independent identifying variation, can inherit the common effect of its group whenever that group contains an identified cell, though this is identification by assumption, untestable for the very cell that lacks overlap. Appendix A gives the formal statement.

3.1 Model Specification

Recall the within-transformed model from (11),

$$\tilde{Y} \;=\; \tilde{D}\,\tau \;+\; \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \mid \tilde{D} \sim N\big(0,\, \sigma^2 I_{NT}\big)\,. \tag{19}$$

We treat the τ_k as exchangeable draws from an unknown mixing distribution G and place a Dirichlet Process prior on G:

$$G \;\sim\; \mathrm{DP}(\alpha,\, G_0), \tag{20}$$

$$\tau_k \mid G \overset{\text{iid}}{\sim} G, \qquad k = 1, \ldots, K, \tag{21}$$

$$G_0 \;=\; \mathcal{N}\big(\mu_0,\, \sigma_0^2\big)\,. \tag{22}$$

Following Ferguson (1973), draws from a Dirichlet Process are almost surely discrete, so realisations of G assign positive probability to ties among the τ_k. These ties induce a random clustering of the K CATTs into $m \le K$ distinct values, which is precisely the partition structure of interest. The base measure G_0 governs the location of cluster effects, and its variance σ_0^2 encodes the analyst's prior dispersion of treatment effects across distinct groups. In the absence of prior information we recommend a diffuse choice (σ_0^2 large relative to σ^2). The concentration parameter $\alpha > 0$ controls the prior expected number of clusters,

$$\mathbb{E}[m \mid \alpha, K] \;\approx\; \alpha \log(1 + K/\alpha)\,, \tag{23}$$

with $\alpha \to 0$ favoring the fully pooled specification and $\alpha \to \infty$ favoring the fully flexible specification. Marginalizing G in (20)–(21) yields a Pólya-urn predictive distribution for the cluster labels $z = (z_1, \ldots, z_K)$, commonly referred to as the Chinese Restaurant Process (CRP). Conditional on the previous $k-1$ assignments,

$$\Pr(z_k = c \mid z_1, \ldots, z_{k-1}) \;=\; \begin{cases} \dfrac{n_c}{k - 1 + \alpha}, & \text{cluster } c \text{ has } n_c \text{ members}, \\[2ex] \dfrac{\alpha}{k - 1 + \alpha}, & c \text{ is a new cluster}. \end{cases} \tag{24}$$

The CRP exhibits the rich-get-richer property, whereby clusters with more members are more likely to attract additional CATTs, but a fresh cluster can always be opened with weight proportional to α. Equivalently, the marginal prior probability of any partition $\mathcal{P} = \{C_1, \ldots, C_m\}$ of $\{1, \ldots, K\}$ takes the closed form

$$\Pr(\mathcal{P}) \;=\; \frac{\alpha^m \,\prod_{l=1}^{m} \big(|C_l| - 1\big)!}{\alpha^{(K)}}, \qquad \alpha^{(K)} = \alpha(\alpha+1)\cdots(\alpha+K-1), \tag{25}$$

which depends on $\mathcal{P}$ only through the cluster sizes and the number of clusters. Equation (25) delivers the partition prior we combine with the data likelihood below.

The complete hierarchical model places a conjugate prior on the error variance as well,

$$\tilde{y} \mid \mathcal{P}, \phi, \sigma^2 \;\sim\; \mathcal{N}\big(\tilde{D}\,\phi,\, \sigma^2 I_{NT}\big)\,, \tag{26}$$

$$\phi \mid \mathcal{P} \;\sim\; \mathcal{N}\big(\mu_0 \mathbf{1}_m,\, \sigma_0^2 I_m\big)\,, \tag{27}$$

$$\sigma^2 \;\sim\; \text{Inv-Gamma}(a_0, b_0), \tag{28}$$

so that no quantity is fixed by plug-in and the specification is a complete Bayesian model. We take this full model, with σ^2 random, as our primary specification throughout, and it is what the simulations and applications report. The inverse-gamma prior is conditionally conjugate, which keeps the collapsed sampler tractable.

At any value of σ^2, the value the sampler conditions on at a given sweep, the prior on ϕ is conjugate to the Gaussian likelihood, so the group effects integrate out analytically. Standard calculations yield

$$\tilde{y} \mid \mathcal{P} \;\sim\; \mathcal{N}\big(\mu_0 \tilde{D}\,\mathbf{1}_m,\; \sigma^2 I_{NT} + \sigma_0^2\,\tilde{D}\,\tilde{D}'\big)\,. \tag{29}$$

Direct evaluation of the density implied by (29) would require inverting the $NT \times NT$ covariance matrix, which is computationally prohibitive in panels of empirical scale. The matrix determinant lemma and the Woodbury identity reduce the calculation to operations on $m \times m$ matrices. Letting $\kappa = \sigma_0^2/\sigma^2$ and dropping additive constants in $\mathcal{P}$, the log marginal likelihood admits the form

$$\log p(\tilde{y} \mid \mathcal{P}) \;\propto\; -\tfrac{1}{2}\,\log\big|\, I_m + \kappa\,\tilde{D}'\tilde{D}\,\big| \;+\; \frac{\kappa}{2\sigma^2}\,\big(\tilde{D}'\tilde{Y}\big)'\big(I_m + \kappa\,\tilde{D}'\tilde{D}\big)^{-1}\big(\tilde{D}'\tilde{Y}\big). \tag{30}$$

The posterior of ϕ given $\mathcal{P}$ and σ^2 is $N(\mu^*, \Sigma^*)$ with

$$\Sigma^* \;=\; \big(\tilde{D}'\tilde{D}/\sigma^2 + I_m/\sigma_0^2\big)^{-1}, \qquad \mu^* \;=\; \Sigma^*\big(\tilde{D}'\tilde{Y}/\sigma^2 + \mu_0 \mathbf{1}_m/\sigma_0^2\big)\,. \tag{31}$$

The collapsed Gibbs sampler (Section 3.3) draws σ^2 from its inverse-gamma full conditional each sweep and evaluates (30)–(31) at that draw.

Remark 1 (The design covariance)**.** *The posterior in* (31) *must use the exact within-design cross-product $\tilde{D}'\tilde{D}$. The cohort-time estimates $\hat{\tau}_k$ are jointly $N(\tau, \sigma^2(\tilde{D}'\tilde{D})^{-1})$ and are correlated through the shared fixed effects. Treating them as independent understates the posterior variance of linear aggregates such as the overall ATT, in our design by a factor of about* 3.5*, and produces overconfident, undercovering intervals. The model above already uses $\tilde{D}'\tilde{D}$; we flag the point because the diagonal shortcut is tempting and its effect on coverage is large.*

3.2 Fixed and Random σ^2

Under the full random-σ^2 specification of Section 3.1, integrating (ϕ, σ^2) out of a partition's marginal likelihood gives a multivariate-t under which the MAP partition minimizes a BIC-type penalty on the log residual sum of squares,

$$\mathcal{P}^{\mathrm{MAP}} \;=\; \arg\min_{\mathcal{P}}\; NT\log\big(\mathrm{RSS}(\mathcal{P})/NT\big) \;-\; 2\log \Pr(\mathcal{P}), \tag{32}$$

so the fit enters on the log scale, σ^2 is profiled out, and the complexity charge grows at the Schwarz $\log(NT)$ rate rather than being a fixed multiple of the RSS. Holding σ^2 fixed at the plug-in $\hat{\sigma}^2 = \mathrm{RSS}_{\mathrm{flex}}/(NT - N - T + 1 - K)$, the residual variance from the fully flexible regression (3), is a useful special case. With σ^2 constant the model is Gaussian, the

group-effect posterior (31) is Normal, and the sampler draws partitions from the Gaussian marginal likelihood (30). Because the residual degrees of freedom $NT - N - T + 1 - K$ are large, the fixed- and random-σ^2 versions deliver essentially the same coverage and interval length in our experiments (Appendix D), so the choice is immaterial for the reported results. The homoskedastic assumption itself can be relaxed, and Appendix C extends the model to a heterogeneous error covariance and augments the sampler with a Gibbs update for group-specific error variances.

The fixed-σ^2 case is also the one under which the MAP partition takes an exact ℓ_0 form, coinciding with the ℓ_0 homogeneity-pursuit estimators of the panel-data and statistics literatures. Under the DP-induced prior $\Pr(\mathcal{P}) \propto \exp(-\lambda \,\#\{\text{cross-group pairs}\})$, the negative log-posterior is, up to a constant, $\mathrm{RSS}(\mathcal{P})/\sigma^2 + 2\lambda \,\#\{\text{cross-group pairs}\}$, so scaling by $\sigma^2/2$ shows that the MAP partition minimizes the penalized residual sum of squares, which trades fit against an ℓ_0 penalty on the number of cross-group CATT pairs, pairs of CATTs placed in different groups and hence estimated with different coefficients,

$$Q(\mathcal{P}) \;=\; \underbrace{\sum_{i}\sum_{t}\Big(Y_{igt} - \alpha_i - \gamma_t - \sum_{g\in\mathcal{G}}\sum_{t\geq g}\tau_{gt}\,D_{igt}\Big)^2}_{\mathrm{RSS}(\mathcal{P})} \;+\; \underbrace{\sum_{g\neq g'}\lambda\,\mathbf{1}[\tau_{gt}\neq\tau_{g't}]}_{\ell_0\text{ penalty}}, \tag{33}$$

with $L = \lambda\sigma^2$, so the penalty is exactly the model's log prior and the ℓ_0-PH estimator is the MAP of the same model rather than a separate procedure. Here $\lambda > 0$ is the penalty parameter and $\mathbf{1}$ is the indicator function. The number of cross-group pairs equals $\frac{K(K-1)}{2} - \sum_{p=1}^{m}\frac{|C_p|(|C_p|-1)}{2}$. The penalty discourages splitting groups (which increases cross-group pairs) unless the resulting reduction in RSS is large enough to compensate. When $\lambda = 0$, $Q(\mathcal{P}) = \mathrm{RSS}(\mathcal{P})$, which is minimized by the fully flexible partition with K groups. As $\lambda \to \infty$, the penalty dominates and $Q(\mathcal{P})$ is minimized by the fully pooled partition with 1 group. For intermediate λ, the optimal partition reflects the data's evidence on which CATTs are equal. Merge groups A and B if and only if the increase in RSS from constraining them equal is smaller than $\lambda\cdot|A|\cdot|B|$. The factor $|A|\cdot|B|$ arises because merging two groups of sizes $|A|$ and $|B|$ eliminates $|A|\cdot|B|$ cross-group pairs.

Proposition 2 (Threshold interpretation)**.** *For two singleton groups $\{j\}$ and $\{k\}$, the merge criterion $\Delta\mathrm{Obj}(j,k) < 0$ is equivalent to*

$$\big(\hat\tau_j - \hat\tau_k\big)^2 \cdot \frac{n_j\, n_k}{n_j + n_k} \;<\; \lambda, \tag{34}$$

where $n_j = \|\tilde D_j\|^2$ is the effective sample size for CATT j. Two CATTs are pooled if and only if their squared difference, scaled by their harmonic-mean sample size, falls below the threshold λ.

This result clarifies why there is a wide range of λ values that recover the true partition $\mathcal{P}$: the RSS cost of a correct merge (two CATTs with the same true effect) is $O(\sigma^2/N)$, while the cost of a wrong merge (different true effects) is $O(N\cdot\Delta\tau^2)$, where $\Delta\tau$ is the effect gap.

For large N, the signal overwhelms the noise and any λ in the interval $(\sigma^2/N,\ N \cdot \Delta\tau^2)$ recovers $\mathcal{P}$.

The penalty in (33) counts the number of distinct pairs across groups, which is the ℓ_0 norm applied to the vector of all $\binom{K}{2}$ pairwise differences $\{\tau_j - \tau_k\}_{j<k}$. The fused LASSO (Tibshirani et al., 2005) instead uses the ℓ_1 norm of pairwise differences (for ordered indices), producing a convex but inexact version of equality. The advantage of the ℓ_0 formulation is that it produces exact equality constraints rather than approximate shrinkage.

Computing the estimator takes two steps, an agglomerative search that returns $\hat{\mathcal{P}}$ in $O(K^3)$ time and an ordinary least squares re-estimation of the grouped model under $\hat{\mathcal{P}}$.[2]

3.3 Estimation and Inference

We sample from the posterior $\Pr(\mathcal{P} \mid \tilde{y}) \propto p(\tilde{y} \mid \mathcal{P}) \Pr(\mathcal{P})$ using the collapsed Gibbs sampler of Neal (2000) (Algorithm 3), which integrates ϕ out analytically at each cluster-assignment step and reinstates it via a conjugate draw at the end of every sweep. At each iteration, we cycle over all K CATTs. For CATT k:

1. Remove k from its current cluster; delete the cluster if empty.
2. For each existing cluster c and for a new cluster, compute the conditional probability

$$\Pr(z_k = c \mid z_{-k}, \tilde{y}) \ \propto \ \begin{cases} n_{-k,c} \times p(\tilde{y} \mid z_k = c,\, z_{-k}), & \text{existing}, \\ \alpha \times p(\tilde{y} \mid z_k = c_{\text{new}},\, z_{-k}), & \text{new}, \end{cases} \tag{35}$$

   where the marginal likelihoods are computed via (30).
3. Draw z_k from the normalized probabilities (35).

After updating all cluster assignments, draw ϕ from the conjugate posterior $N(\mu^*, \Sigma^*)$. The sequence of draws $\{z^{(s)}, \phi^{(s)}\}_{s=1}^{S}$ approximates the posterior.

The posterior mean of τ_k is

$$\hat{\tau}_k^{\text{Bayes}} \ = \ (S - S_B)^{-1} \sum_{s=S_B}^{S} \phi^{(s)}_{z_k^{(s)}}, \tag{36}$$

[2]Appendix B gives the agglomerative merge criterion and the re-estimation, and explains why the second step is needed. The agglomerative search minimizes an orthogonal approximation to the objective, and its pooled coefficients equal the restricted OLS only when the within-transformed cohort-time dummies are orthogonal, so the OLS re-estimation recovers the BLUE in general panels.

where S_B is the burn-in sample. The 95% credible interval is obtained from the empirical quantiles of the draws. The co-clustering probability

$$\hat{\pi}_{jk} \;=\; S^{-1} \sum_{s=1}^{S} \mathbf{1}\{z_j^{(s)} = z_k^{(s)}\} \tag{37}$$

gives the posterior probability that CATTs j and k belong to the same group. The $K \times K$ matrix $\hat{\Pi} = [\hat{\pi}_{jk}]$ is the posterior similarity matrix, summarizing the full uncertainty over the grouping structure; when a single representative partition is desired, it can be obtained from the posterior together with a credible region using the loss-based summaries of Lau and Green (2007) and Wade and Ghahramani (2018). Aggregated estimands (e.g. the overall ATT or event-study coefficients) are computed by applying the linear aggregator of interest to each posterior draw and summarizing across the chain, which delivers uncertainty quantification that incorporates the uncertainty in the partition itself.

The MAP partition of Section 3.2 and the posterior mean differ in what they report. The MAP is a single grouping of the cohort-time effects into a few distinct levels, the minimizer of the penalized residual sum of squares. The posterior mean, by contrast, averages over partitions and returns values that match no single grouping, so extracting a representative partition from it requires a separate loss-based summary of the co-clustering matrix.

We do not claim that the mode is a less biased point estimate than the posterior mean. Holding the regularization fixed, the two nearly coincide, so choosing between the MAP and the posterior mean is about whether a single partition or the full partition-averaged summary is wanted, not about small-sample bias, and inference should come from the posterior in either case.

Remark 2 (Inference after partition selection)**.** *The variance formula for $\hat{\phi}$ in* (47) *treats the partition $\hat{\mathcal{P}}$ as fixed. In finite samples $\hat{\mathcal{P}}$ is itself estimated, so confidence intervals built from $\sigma^2[(\tilde{D}^{*\prime}\tilde{D}^{*})^{-1}]_{pp}$ are valid only conditional on the selected partition and can under-cover unconditionally, as with inference after any model-selection step (Leeb and Pötscher, 2005). When the recovery interval of Proposition 2 is wide, so that the true partition is selected with probability approaching one, this gap is small, and the simulations of Section 4 confirm that the plug-in intervals are then only mildly anti-conservative. The Bayesian estimator of Section 3 avoids the issue entirely by averaging over the partition rather than conditioning on a single one, and attains near-nominal coverage in our experiments; it is our preferred route to honest uncertainty quantification.*[3]

[3] An obvious alternative, sample splitting, selecting the partition on one subsample and estimating the coefficients on another, does not help, because halving the sample degrades recovery and a wrongly merged partition induces a specification bias that no standard-error correction removes. A selective-inference correction that conditions on the selection event (Fithian et al., 2014; Rinaldo et al., 2019) is a further route we do not pursue.

4 Simulation Study

This section reports a Monte Carlo study with three aims: (i) to document, on the variance scale that matters for causal estimands, the efficiency of the Dirichlet-Process (DP) estimator relative to the fully flexible and fully pooled benchmarks under partial homogeneity; (ii) to show how those gains depend on how well separated the distinct effects are; and (iii) to evaluate honest inference by measuring confidence- and credible-interval coverage. We also carry the ℓ_0 MAP estimator through the same experiments to see how the model's mode compares with its posterior, and it behaves much like the DP throughout.

4.1 Design

We simulate a balanced panel with $N = 2{,}000$ units and $T = 10$ periods. Units are split equally across four cohorts, a never-treated group ($g = 0$) and three treated cohorts entering at $t = 3, 5, 7$. This yields $K = 18$ post-treatment cohort-time cells. Outcomes follow

$$Y_{it} = \alpha_i + \lambda_t + \tau^*_{g(i),t}\,\mathbf{1}\{t \geq g(i)\} + \varepsilon_{it}, \qquad \alpha_i \sim \mathcal{N}(0,1),\ \lambda_t = 0.1(t-1),\ \varepsilon_{it} \sim \mathcal{N}(0,1). \tag{38}$$

The unit and time fixed effects are removed by the within transformation, so the estimators' sampling behavior is driven entirely by ε.

The partial-homogeneity truth is parameterized by the number of distinct effects and their separation. For a given number of distinct effects m^*, we partition the $K = 18$ cells into m^* near-equal groups and assign each group a mean on an evenly spaced grid with adjacent gap Δ. We sweep $m^* \in \{1, 3, 6, 9, 18\}$, from fully homogeneous ($m^* = 1$, where pooling is correct) to fully heterogeneous ($m^* = 18$, where the flexible estimator is correct). The gap is calibrated to a unit-free separation $\delta \equiv \Delta/\overline{\mathrm{sd}}(\hat{\tau}_{\mathrm{flex}})$, the distance between adjacent group means in standard errors of the flexible cohort-time estimates; δ governs how easily the groups can be told apart. We report $\delta \in \{3, 6, 12\}$, with $\delta = 6$ as the headline.

We compare five estimators, pooled TWFE (the biased-but-precise benchmark), flexible TWFE (unbiased-but-imprecise, in this balanced design it coincides with the Gardner (2022) two-stage estimator), the infeasible oracle partial-homogeneity estimator that is handed the true partition (the efficiency floor), the ℓ_0-PH estimator, and the Bayes-PH estimator. The ℓ_0 penalty is chosen by BIC (Schwarz, 1978) along the agglomeration path, $\mathrm{BIC}(m) = NT \log(\mathrm{RSS}(\hat{\mathcal{P}}_m)/NT) + m \log(NT)$, requiring no tuning grid. The DP sampler uses a diffuse base measure ($\mu_0 = 0$, $\sigma_0^2 = 100$) and concentration $\alpha = 7$ (so the prior expected number of groups is about $K/2$), run for 150 Gibbs sweeps with a 50-sweep burn-in, drawing σ^2 from its inverse-gamma full conditional each sweep (the full Bayesian model), and its recorded draws use the exact within-design covariance of the cohort-time estimates (Remark 1). The overall ATT is the equal-weight average of the

K CATTs; aggregator choice is orthogonal to the specification question and does not affect the conclusions. All results use 500 Monte Carlo replications.

4.2 Variance and recovery under partial homogeneity

Table 1 reports, for the headline separation $\delta = 6$, the sampling variance of the cohort-time estimates relative to flexible TWFE (lower is more precise), the average absolute CATT bias (a unbiasedness check), and the adjusted Rand index (ARI) measuring how well the true partition is recovered.

Table 1: Sampling variance (ratio to flexible), bias, and partition recovery under partial homogeneity, $\delta = 6$. "Var ratio" is the average per-cell Monte Carlo variance divided by that of flexible TWFE; "|Bias|" is the average absolute CATT bias; ARI is the adjusted Rand index against the true partition. Oracle PH is infeasible (true partition known). 500 replications.

Scenario	Method	Var ratio	|Bias|	ARI
$m^* = 1$ (homogeneous)	Pooled	0.14	0.001	1.00
	Flexible	1.00	0.002	0.00
	Oracle PH	0.14	0.001	1.00
	ℓ_0-PH	0.49	0.002	0.41
	Bayes-PH	0.20	0.001	0.57
$m^* = 3$ (partial)	Pooled	0.13	0.236	0.00
	Flexible	1.00	0.002	0.00
	Oracle PH	0.26	0.001	1.00
	ℓ_0-PH	0.55	0.004	0.94
	Bayes-PH	0.48	0.005	0.89
$m^* = 6$ (partial)	Pooled	0.14	0.490	0.00
	Flexible	1.00	0.002	0.00
	Oracle PH	0.40	0.002	1.00
	ℓ_0-PH	0.74	0.002	0.95
	Bayes-PH	0.69	0.002	0.88
$m^* = 9$ (partial)	Pooled	0.15	0.742	0.00
	Flexible	1.00	0.002	0.00
	Oracle PH	0.61	0.001	1.00
	ℓ_0-PH	1.30	0.005	0.91
	Bayes-PH	0.97	0.004	0.86
$m^* = 18$ (heterogeneous)	Pooled	0.15	1.483	0.00
	Flexible	1.00	0.002	1.00
	Oracle PH	1.00	0.002	1.00
	ℓ_0-PH	3.42	0.008	0.01
	Bayes-PH	1.61	0.004	0.29

Three patterns stand out. First, pooled TWFE is always the most precise (variance ratio ≈ 0.14) but is severely biased under any heterogeneity, with an average CATT bias rising from 0.24 at $m^* = 3$ to 1.48 at $m^* = 18$; it is unusable for cohort-time effects. Second, at the partial-homogeneity scenarios that motivate the paper ($m^* = 3, 6$) the ℓ_0-PH and Bayes-PH estimators cut the sampling variance of the CATTs by 26–52% relative to flexible TWFE while remaining essentially unbiased (average $|\text{bias}| \leq 0.005$), and they recover the partition well (ARI ≈ 0.9). Third, the methods correctly degrade at the extremes, in that at $m^* = 1$ they shrink toward pooling (though not all the way), and at $m^* = 18$, where every cell is genuinely distinct, any pooling hurts and the flexible estimator is preferred. Figure 1 plots the variance ordering across m^*.

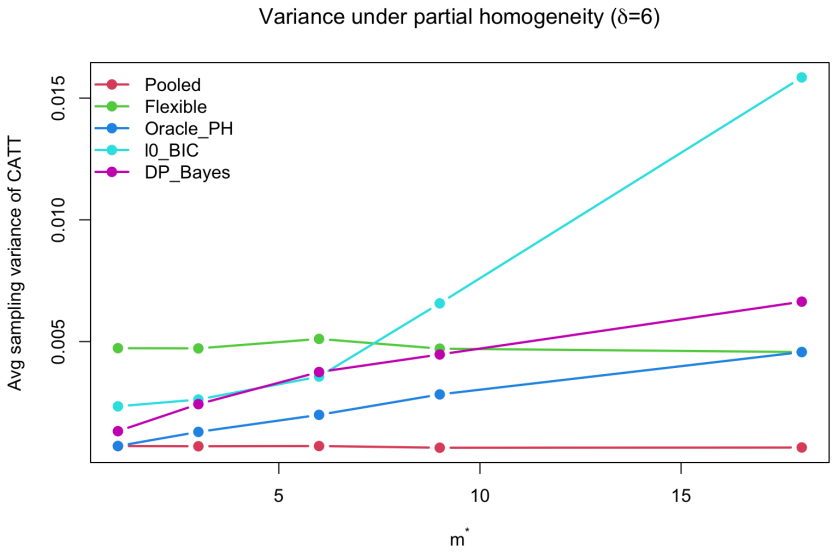

Figure 1: Average sampling variance of the cohort-time estimates against the true number of distinct effects m^* ($\delta = 6$). The ℓ_0 and DP estimators track the infeasible oracle in the partial-homogeneity region and revert to flexible TWFE as $m^* \to K$.

The realized gain depends on how well separated the distinct effects are, exactly as the recovery interval of Proposition 2 suggests. Table 2 fixes $m^* = 6$ and varies δ.

Table 2: Effect of separation δ on precision and recovery, $m^* = 6$. Variance ratio to flexible and ARI. 500 replications.

	$\delta = 3$		$\delta = 6$		$\delta = 12$	
Method	Var ratio	ARI	Var ratio	ARI	Var ratio	ARI
Oracle PH	0.40	1.00	0.40	1.00	0.40	1.00
ℓ_0-PH	1.47	0.54	0.74	0.95	0.41	1.00
Bayes-PH	1.19	0.46	0.69	0.88	0.46	0.93

When the effects are close ($\delta = 3$) the partition cannot be recovered reliably (ARI ≈ 0.5) and the selection noise erases, indeed reverses, the variance advantage, and both feasible estimators are then less precise than flexible TWFE. As separation grows the gap to the oracle closes, and at $\delta = 12$ the feasible ℓ_0-PH estimator essentially attains the oracle variance reduction (0.41 versus 0.40) with near-perfect recovery. Figure 2 shows the monotone relationship. The practical message is that the methods deliver their promised gains when heterogeneity is "clumpy", a modest number of well-separated effect levels, and should not be expected to when the effects form a near-continuum. A secondary but useful observation is that the Bayes-PH estimator is the more robust of the two, since in the hard regimes ($\delta = 3$, or $\delta = 6$ with the fine $m^* = 9$ partition) its variance ratio stays at or below one, whereas the ℓ_0-PH estimator, which commits to a single partition, has a heavier tail and can exceed one.

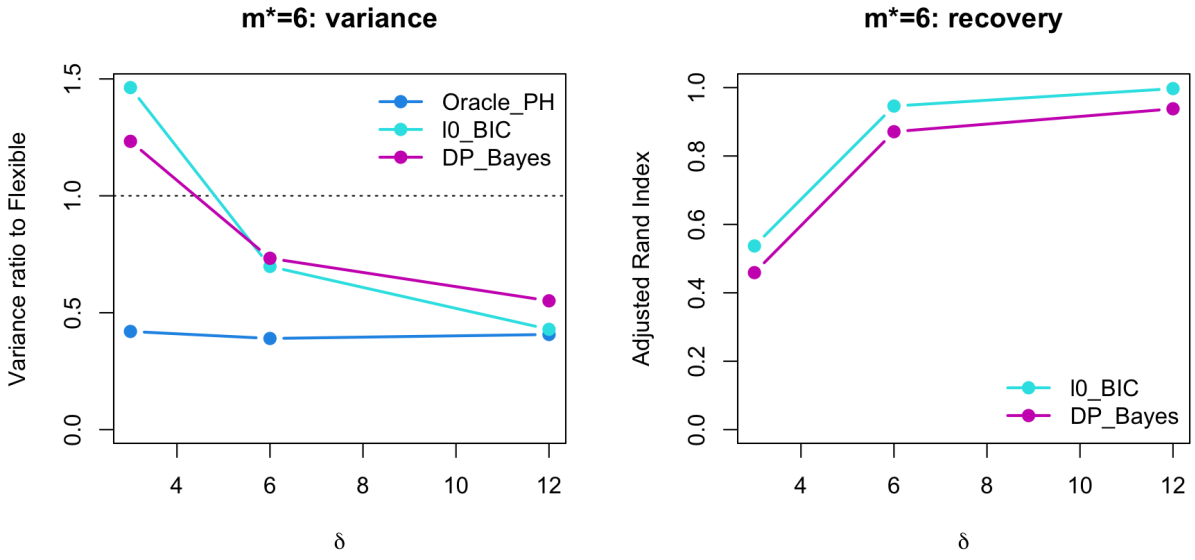

Figure 2: Precision (left) and partition recovery (right) as functions of the separation δ, at $m^* = 6$. The variance advantage over flexible TWFE (values below the dashed line) is realized only once the effects are separated enough to be recovered.

4.3 Honest inference: coverage

Table 3 evaluates the calibration of 95% intervals at $m^* = 6$, $\delta = 6$, reporting coverage and average length for the cohort-time effects and for the overall ATT. For the ℓ_0-PH estimator we report the plug-in interval that treats the selected partition as fixed. The Bayes-PH interval is the 2.5–97.5% posterior credible interval.

Table 3: Coverage and average length of nominal 95% intervals at $m^* = 6$, $\delta = 6$; targets are the true CATTs and the true ATT. Oracle PH is infeasible. 500 replications.

	CATT		ATT	
Method	Coverage	Length	Coverage	Length
Pooled	0.05	0.10	0.05	0.10
Flexible	0.96	0.27	0.95	0.12
Oracle PH	0.95	0.17	0.95	0.11
ℓ_0-PH	0.91	0.17	0.94	0.11
Bayes-PH	0.94	0.20	0.95	0.11

The pooled intervals collapse (coverage 0.04–0.05): they are short but centered on a biased estimand. The flexible and oracle intervals are correctly calibrated, the oracle being much shorter, and they bracket the feasible estimators. The Bayes-PH credible intervals attain near-nominal coverage (0.94 for the CATTs, 0.95 for the ATT) at a length between oracle and flexible, honest and informative uncertainty. The ℓ_0-PH interval is only mildly anti-conservative here (0.91 CATT, 0.94 ATT): at this separation the correct partition is recovered often enough that the conditional-on-selection distortion is small.[4] Honest uncertainty under partition selection thus comes not from adjusting the variance but from averaging over the partition, which is what the DP posterior does. It is also worth noting what does drive the result. The calibration hinges on using the exact cross-cell covariance of the cohort-time estimates (Remark 1), whereas the treatment of σ^2 is immaterial. The full Bayesian sampler draws σ^2 from its inverse-gamma full conditional, and holding it fixed at a plug-in value gives essentially the same coverage and interval length (Appendix D).

4.4 Solution paths

Figure 3 traces the bias-variance trade-off as a function of the regularization strength, for ℓ_0-PH along the agglomeration path indexed by the number of groups m (small m = strong penalty), and for DP across the concentration α. Two features are worth emphasizing. First, the overall ATT is remarkably robust to over-pooling, where its bias stays small ($\lesssim 0.015$) across the entire ℓ_0 path down to $m = 3$ and only spikes once the path collapses toward a single group ($|\text{bias}| = 0.10$ at $m = 1$), while the CATT variance falls as cells are pooled. The aggregate thus tolerates a wide range of penalties, whereas the cohort-time effects require getting the partition approximately right, the CATT variance is minimized near the true $m = 6$ and inflates when the path is forced into wrong merges. Second, the DP path is nearly flat in α: both the ATT bias and variance barely move over $\alpha \in \{1, \ldots, 30\}$, so the practitioner's choice of concentration has little effect on the reported estimates. The

[4] A natural alternative removes this distortion by sample splitting, selecting the partition on one half of the units and computing coefficients and standard errors on the other. It does not help, its coverage is in fact lower (0.74 CATT, 0.90 ATT). Splitting halves the sample, which degrades recovery, and a wrongly merged partition then injects a specification bias that a correctly sized standard error cannot undo.

mean BIC-selected number of groups (≈ 6.2) sits in the flat, low-bias region of the ℓ_0 path, confirming that the plateau in $\hat{m}$ is a serviceable tuning diagnostic.

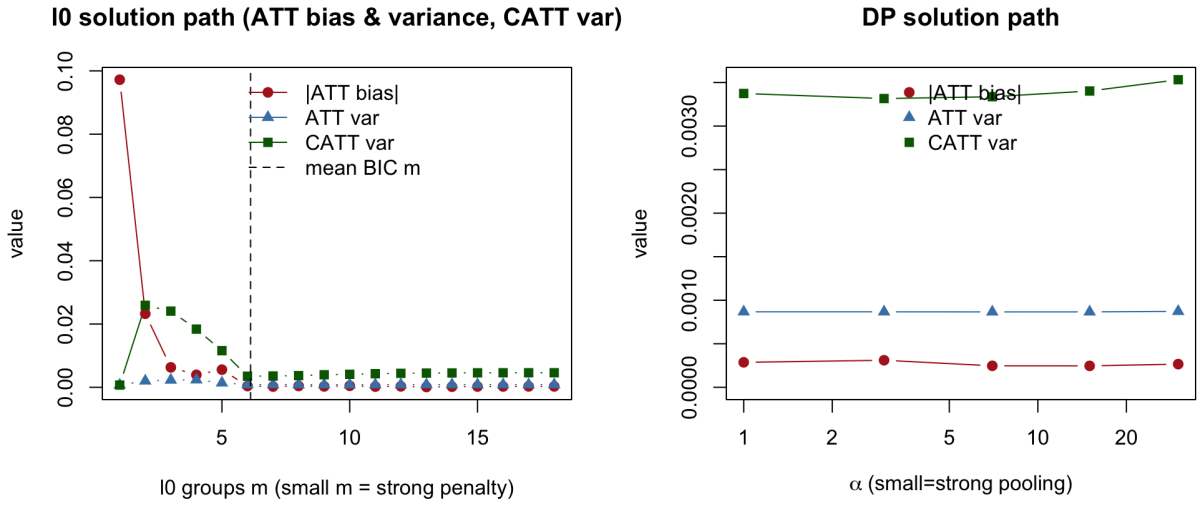

Figure 3: Solution paths at $m^* = 6$, $\delta = 6$. Left: ℓ_0, indexed by the number of groups m (dashed line = mean BIC selection). Right: DP, indexed by the concentration α (log scale). The overall ATT bias (red) stays near zero across a wide range while the CATT variance (green) falls with pooling.

5 Empirical Applications

We present two applications that exhibit the two regimes the method is built for, one where the cohort-time effects carry recoverable heterogeneity and one where they are indistinguishable from a common value.

5.1 Minimum Wages and Teen Employment: Recovering Partial Homogeneity

We revisit the county-level analysis of minimum wages and teen employment from Callaway and Sant'Anna (2021), using their publicly available panel of 500 U.S. counties observed over 2003–2007, with cohorts first raising the minimum wage in 2004, 2006, and 2007 and a not-yet-treated control group. We estimate the $K = 7$ post-treatment cohort-time effects $\mathrm{ATT}(g,t)$ with the Callaway and Sant'Anna (2021) estimator and take their exact joint covariance $\hat{\Sigma}$ from the estimator's influence functions. Because the cells share counties, $\hat{\Sigma}$ is non-diagonal; as in the simulations, we use it exactly in both estimators, and the DP posterior draws use the corresponding Gauss–Markov posterior (Remark 1).

Unlike a homogeneous design, here there is real structure to recover, as the cross-cell dispersion of the estimates (0.050) is more than twice the typical standard error (0.022), a signal-to-noise ratio of 2.25 that places the application in the recoverable region of Section 4. Table 4 reports the seven effects, the ℓ_0 grouping selected by BIC, and the per-cell variance reduction.

Table 4: Cohort-time effects on log teen employment, Callaway and Sant'Anna (2021) data. ATT(g,t) and its standard error; the BIC-selected ℓ_0 group and grouped effect; and the ratio of the ℓ_0-grouped variance to the flexible variance for each cell (lower = more precise).

cohort:year	ATT(g,t)	SE	ℓ_0 group	grouped effect	var. ratio
2004:2004	−0.019	0.022	1	−0.029	0.28
2006:2007	−0.041	0.020	1	−0.029	0.33
2007:2007	−0.026	0.017	1	−0.029	0.49
2004:2005	−0.078	0.030	2	−0.087	0.77
2004:2007	−0.101	0.034	2	−0.087	0.61
2004:2006	−0.136	0.035	3	−0.140	0.78
2006:2006	+0.005	0.016	4	+0.011	0.78

The ℓ_0-PH estimator collapses the seven cells into four effect levels: a near-zero group containing the 2004 impact effect and the recent cohorts' effects, the 2004 cohort's larger mature effects, its peak, and the lone positive cell. Pooling roughly halves the variance of the cells that share a group (the near-zero group's variance ratios are 0.28–0.49), and even the singleton cells gain precision because the Gauss–Markov estimator borrows across the correlated cells. The overall (cohort-size-weighted) average effect is essentially unchanged and slightly more precise, −0.040 (SE 0.012) under the flexible estimator versus −0.039 (SE 0.012) under ℓ_0, so the substantive conclusion of a roughly 4% teen-employment reduction is preserved while the estimates are summarized by a handful of interpretable effect levels. These point estimates agree on the headline, but both condition on a single grouping, and the Dirichlet-Process posterior that follows both gauges how firmly that grouping is identified and, given the small number of cells, shows how much the overall summary can move once that conditioning is relaxed.

The DP posterior sharpens the interpretation by quantifying which groupings are certain. Figure 4 plots the effects colored by ℓ_0 group and the posterior co-clustering matrix. The near-zero cells co-cluster with high probability (0.77–0.90), so the "small effect" group is a robust feature; the 2004 cohort's dynamic cells co-cluster only moderately (0.56–0.70), correctly signaling that the finer split of its smooth ramp is uncertain rather than a firm partition. This is the honest reading the co-clustering is meant to deliver, a reliable coarse structure with calibrated uncertainty about the finer detail.

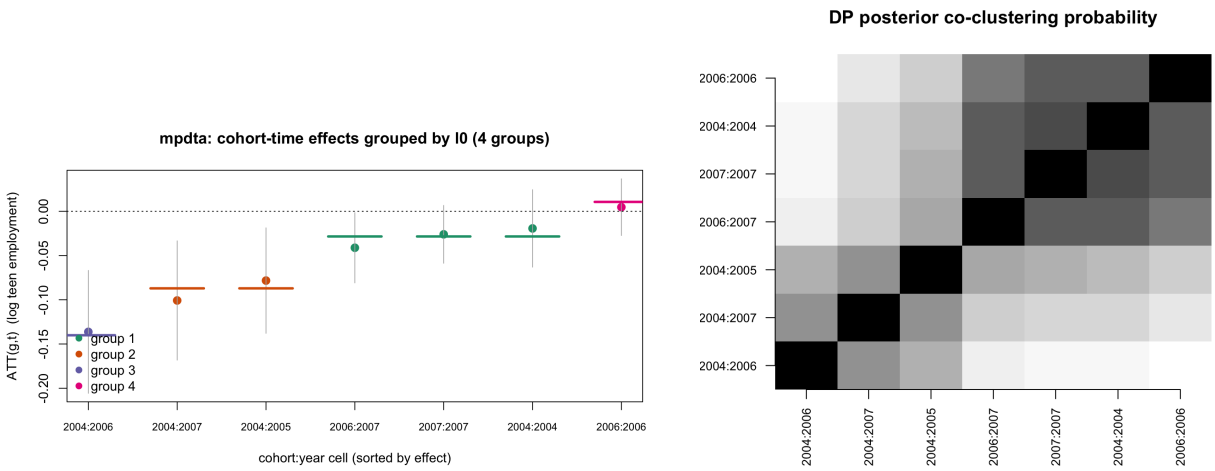

Figure 4: Left: the seven cohort-time effects (with 95% intervals) colored by ℓ_0 group, horizontal bars mark the grouped effects. Right: the DP posterior co-clustering probabilities; darker cells co-cluster more often.

Because the number of cells is small, the DP overall effect, unlike the flexible and ℓ_0-PH point estimates, is sensitive to the concentration α. Figure 5 traces the population-weighted effect and its 95% credible band across α from 0.1 to 100, against the fully pooled (-0.010) and flexible (-0.040) anchors. The effect slides monotonically from about -0.014 under heavy pooling to about -0.037 as α grows and the model approaches the flexible fit, and it is bounded away from zero only for $\alpha \gtrsim 2$. Rather than fix α, we report the whole path. As a reference point, a BIC pass on the agglomeration path selects four groups, which the prior's expected group count matches at $\alpha \approx 13$ (panel b), where the DP effect is about -0.03 and close to the flexible estimate. We offer this only as a suggestion. Choosing α from the data turns the prior into a data-dependent object, with the usual consequences of empirical Bayes (understated posterior uncertainty and a double use of the data), so we prefer to show the sensitivity rather than commit to a single value.

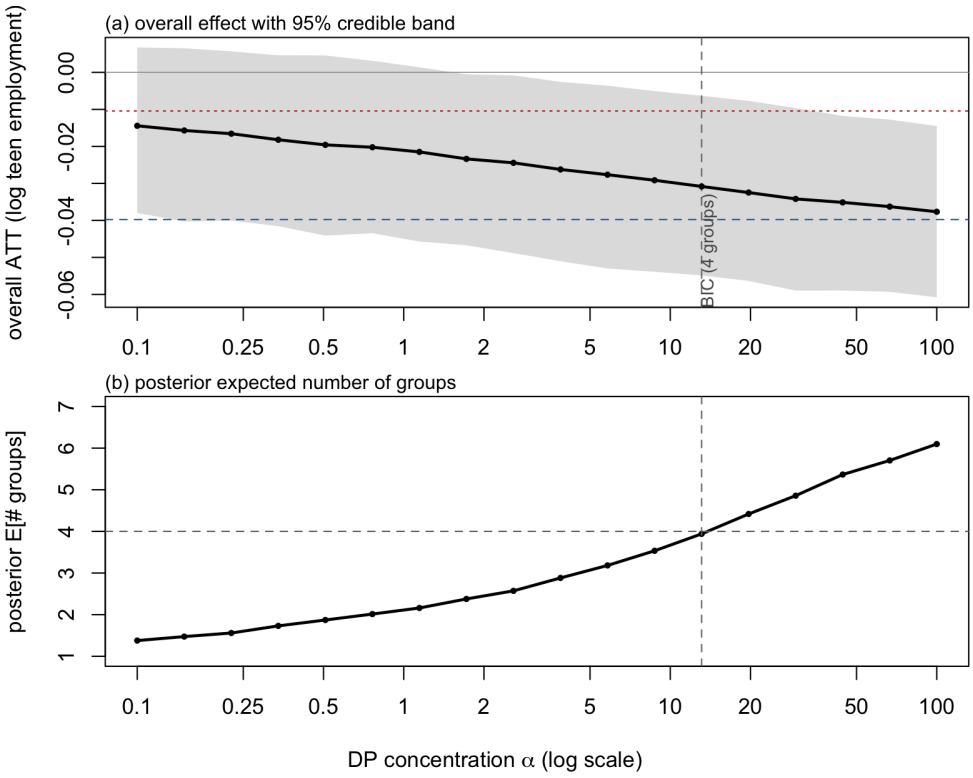

Figure 5: Sensitivity of the Callaway and Sant'Anna (2021) overall effect to the DP concentration α. Panel (a), the population-weighted overall ATT (solid) with its 95% credible band (shaded), against the fully pooled (red dotted, -0.010) and flexible (blue dashed, -0.040) benchmarks. Panel (b), the posterior expected number of groups. The vertical line marks $\alpha \approx 13$, where the prior's expected group count equals the four groups selected by BIC on the agglomeration path.

The dynamic (event-study) aggregation in Figure 6 clarifies what the estimators are grouping. The effect on teen employment is small at impact (event time 0, -0.019) and grows over the horizon to -0.054, -0.136, and -0.101 at event times 1 through 3, the familiar pattern of a minimum-wage effect that accumulates over several years. Because only the 2004 cohort is observed at the longer horizons, its later cells carry the large effects while the impact-year effects are small and similar across cohorts. The partial homogeneity the method recovers is, in this application, largely the homogeneity of the short-run response across cohorts, with the accumulating longer-run effects kept separate. The pre-treatment coefficients are small (event times -2 and -1), with a modest positive value at the longest lead that is identified off a single early cohort.

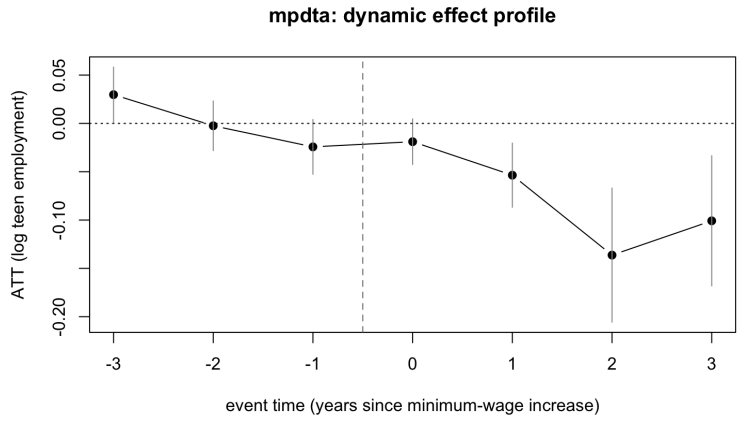

Figure 6: Dynamic (event-study) aggregation of the Callaway and Sant'Anna (2021) effects: the teen-employment effect is near zero at impact and builds over event time. Pre-treatment coefficients (left of the dashed line) are small.

5.2 Minimum Wages and Low-Wage Jobs: A Specification Test

Our second application is the influential study of minimum wages and low-wage jobs by Cengiz et al. (2019), which estimates the employment effect of 138 state-level minimum wage increases between 1979 and 2016 using a "stacked" event-study design. Here the method plays the complementary role of a specification test.

5.2.1 Context and Stacked Event-Study Design

In the stacked design, a separate dataset (a "stack" or "event") is created for each of the $E = 138$ minimum wage hikes. Each stack $g \in \{1, \ldots, E\}$ contains the treated state and a set of clean control states observed over a 5-year window around the policy implementation. The stacks are then appended together.

The authors' baseline specification is a fully pooled TWFE regression on the stacked data, which imposes a single, homogeneous treatment effect across all 138 events:

$$Y_{itg} = \alpha_{ig} + \gamma_{tg} + \theta \cdot D_{itg} + \epsilon_{itg}, \tag{39}$$

where Y_{itg} is the employment outcome for state i at time t in stack g. The terms α_{ig} and γ_{tg} are state-by-stack and time-by-stack fixed effects, ensuring that treated units are only compared to control units within their specific sub-experiment. D_{itg} is an indicator equal to 1 if state i is treated in stack g in the post-treatment period. Under this specification, the authors report a near-zero average effect on net low-wage employment.

However, as established in Section 2, this pooled estimator $\hat{\theta}$ is efficient but biased if the true treatment effects are heterogeneous. To allow for heterogeneity, we relax Equation 39

into a fully flexible Interacted TWFE (ITWFE) model, assigning a distinct Cohort-Average Treatment Effect on the Treated (CATT), τ_g, to each event:

$$Y_{itg} = \alpha_{ig} + \gamma_{tg} + \sum_{g=1}^{E} \tau_g \left(D_{itg} \times \mathbb{I}\{event = g\} \right) + \epsilon_{itg} \tag{40}$$

Because the design matrix of Equation 40 is block-diagonal, the estimation of each $\hat{\tau}_g$ is perfectly separable. While this flexible model is unbiased, relying strictly on state-level variation yields highly noisy estimates for individual events.

To recover the latent structure, we apply our ℓ_0-penalized estimator. We select the optimal partition of the 138 minimum wage effects by minimizing the penalized residual sum of squares:

$$\min_{\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\gamma}} \left\{ \sum_{g=1}^{E} \sum_{i,t} \left(Y_{itg} - \alpha_{ig} - \gamma_{tg} - D_{itg}\tau_g \right)^2 + \lambda \sum_{j<k} \mathbb{I}(\tau_j \neq \tau_k) \right\}, \tag{41}$$

where λ is the hyperparameter determining the threshold for model parsimony. The penalty forces states with sufficiently similar independent estimates to share a single coefficient, achieving partial homogeneity without imposing the strict global restriction of Equation 39.

5.2.2 Baseline Replication and Data Hygiene

Before applying the penalized estimator to explore latent heterogeneity, we first replicate the fully pooled TWFE specification from Cengiz et al. (2019) to establish a baseline. Table 5 demonstrates that our replication sample accurately recovers the primary pooled estimates reported by the authors. The pooled model implies a near-zero average net effect on low-wage employment, driven by the destruction of jobs below the new minimum wage (Δb) being largely offset by the "bunching" of new compliance jobs at or slightly above the new minimum wage (Δa).

Table 5: Baseline Pooled ITWFE Replication

Employment Outcome	**Pooled Estimate**
Missing Jobs (Δb)	-0.016^{***}
Excess Jobs (Δa)	0.019^{***}
Net Employment Change	0.003

Note: Estimates represent the change in employment as a fraction of the affected workforce. These pooled point estimates closely mirror the benchmark findings of the original study, suggesting a negligible average disemployment effect.

While the pooled model provides a stable average, estimating Equation 40 flexibly for all 138 independent events reveals severe localized sampling anomalies. Because the employment effects are normalized by $\bar{b}_{-1}$ (the share of the state's workforce earning below the new minimum wage in the year prior to the hike), regions with very thin survey samples can produce a denominator approaching zero. Dividing the raw job counts by a near-zero denominator causes the resulting percentage estimate to artificially explode to economically impossible magnitudes.

A handful of events yield mathematically anomalous net percentage effects (exceeding several hundred percent of the affected workforce) for exactly this reason. To sidestep the near-zero-denominator problem, our specification analysis below summarizes each event by its average post-treatment event-study coefficient on affected employment rather than by the ratio-normalized percentage effect; the event-study coefficient is well behaved (its cross-event standard deviation is 0.005, with no explosive outliers).

5.2.3 Specification Analysis: Is the Null Effect Homogeneous?

We obtain the $E = 138$ event-level effects from the authors' Harvard Dataverse replication package for Cengiz et al. (2019), summarizing event g by $\hat{\tau}_g$, the average of its post-treatment event-study coefficients on affected employment (event times 0 through 4). Before selecting a specification, we ask whether there is any heterogeneity to recover. The design provides an internal noise gauge, since under no-anticipation the pre-treatment coefficients have expectation zero, so their cross-event dispersion measures the per-event sampling noise. We find a pre-treatment (noise) standard deviation of 0.0065, which exceeds the post-treatment cross-event spread of 0.0050. The implied signal-to-noise ratio is 0.77, and the variance decomposition attributes essentially 0% of the cross-event variation in employment effects to genuine heterogeneity, so the 138 event effects are statistically indistinguishable from a common value.

This places the application squarely in the low-separation regime that Section 4 identifies as the boundary of reliable recovery, and the estimators behave accordingly. The BIC-tuned ℓ_0-PH estimator splits the events into five effect-ordered bands (of sizes 51, 47, 36, 3, 1), and the Bayes-PH estimator returns between 2.7 groups (parsimonious prior, $\alpha = 1$) and 7.6 groups ($\alpha = 5$). As our simulations show, at a signal-to-noise ratio below one these "groups" are quantiles of noise rather than genuine effect clusters, and should not be interpreted as evidence of systematic heterogeneity; indeed, the formal test above finds none. The disagreement between the two estimators is itself a symptom of the regime.

What is robust, and what matters for the applied conclusion, is that the overall average effect is invariant to the specification. Table 6 reports it, showing that the pooled, ℓ_0, and Bayes-PH estimates of the population-weighted average employment effect are all near zero and mutually indistinguishable, and the DP posterior credible interval is a tight band of small positive values that excludes both zero and any economically meaningful disemployment, its narrowness reflecting the pooled precision rather than an economically

important effect. Allowing for arbitrary event-level heterogeneity does not overturn the near-zero net-employment finding of Cengiz et al. (2019); it confirms that the pooled specification is adequate for the aggregate. Figure 7 displays the 138 flexible event effects with the pooled effect superimposed.

Table 6: Overall employment effect under alternative specifications, Cengiz et al. (2019) events. Effects are average post-treatment coefficients on affected employment (population-weighted for the aggregate). The Bayes-PH interval is the 95% posterior credible interval.

Specification	Groups	Overall effect
Pooled (single effect)	1	0.0015
Flexible (per event)	138	0.0018
ℓ_0-PH (BIC)	5	0.0016
Bayes-PH ($\alpha = 1$)	2.7	0.0014 [0.0003, 0.0026]

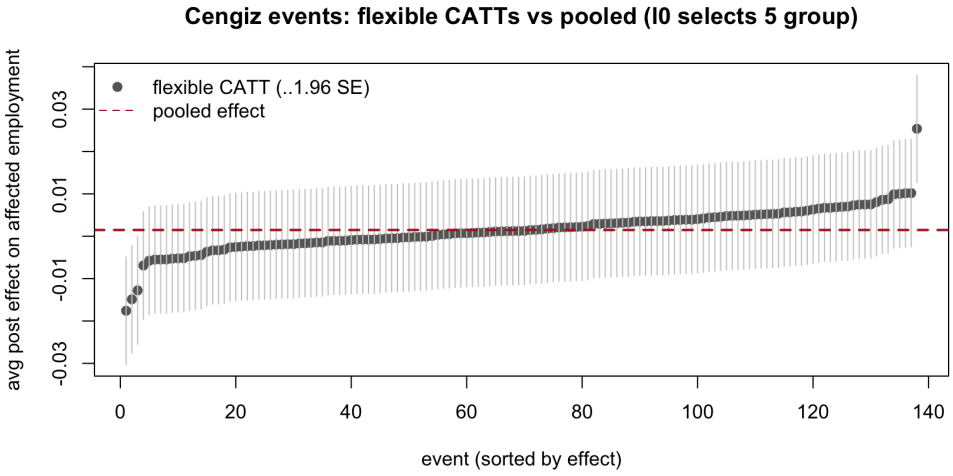

Figure 7: The 138 event-level employment effects (average post-treatment coefficient on affected employment), sorted, with a common permutation-based confidence band; the dashed line is the pooled effect. The cross-event spread is no larger than the pre-trend noise band, and the pooled effect is near zero.

Finally, we note that Cengiz et al. (2019) themselves group events into three "Card–Krueger" bins defined ex ante by how binding the minimum wage is. Our exercise asks, and answers in the negative, a different question, whether the employment effects, taken as estimated, cluster into distinct levels. The two are complementary, in that grouping by a policy covariate is a modeling choice about expected heterogeneity, whereas our data-driven partition tests for realized heterogeneity in the effects and, here, finds none beyond noise.

5.2.4 Diagnosing Canonical TWFE Bias: The Goodman-Bacon Decomposition

To motivate the stacked design, we decompose the canonical (unstacked) two-way fixed effects estimand following Goodman-Bacon (2021). Two adjustments are needed to apply the decomposition here, since we proxy the affected workforce by employment below a fixed \$10 threshold (the true minimum-wage threshold moves across states and years), and we define treatment by the first increase a state enacts (to obtain an absorbing treatment). Both are conservative and, if anything, understate the bias.

Table 7 reports the result. About 20% of the estimand's weight comes from "forbidden" comparisons that use already-treated states as controls, and these carry a negative average estimate (-0.006), whereas the clean treated-versus-untreated comparisons (about 57% of the weight) are positive (0.006). Because the pooled coefficient is the weighted average of these components, the already-treated controls bias it downward. This is the standard rationale for the stacked event-study design used above, and more broadly for taking the cohort-time effects, rather than a single pooled coefficient, as the objects of interest that the specification methods of this paper then group.

Table 7: Goodman–Bacon decomposition of the canonical TWFE estimand (simplified absorbing-treatment panel).

Comparison type	Weight	Average estimate
Earlier vs. later treated	0.236	0.001
Later vs. earlier treated	0.195	-0.006
Treated vs. untreated	0.569	0.006

6 Conclusion

This paper addresses a step in the staggered DiD workflow that has received comparatively little attention, model specification. Once cohort-time effects are identified as the estimands of interest, the researcher must still decide how many distinct parameters to estimate. We frame this as a partition-selection problem and address it with a Dirichlet Process mixture over the cohort-time effects. The DP posterior assigns a probability to every grouping and reports full uncertainty, including the uncertainty in the grouping itself, and is the object we recommend for inference. Holding the error variance fixed, its maximum a posteriori partition coincides with an ℓ_0-penalized estimator, a connection to the homogeneity-pursuit literature. As a secondary benefit, the partial-homogeneity restriction can also supply identifying information for a cell that lacks a clean comparison, though that identification rests entirely on the assumed grouping and is untestable for the cell that needs it.

The simulations and applications make the trade-offs concrete. Under partial homogeneity with well-separated effect levels, both estimators cut the sampling variance of

the cohort-time effects by 26–52% relative to the fully flexible estimator at no bias cost, approaching the infeasible oracle as the separation grows, with gains that vanish once the effects are nearly indistinguishable, and the DP posterior attains near-nominal coverage by marginalizing over the partition where naive ℓ_0-PH intervals are usable but mildly anti-conservative. The two applications show both regimes on real data. In the minimum-wage and teen-employment data of Callaway and Sant'Anna (2021) the effects are genuinely heterogeneous and the method recovers a partially homogeneous structure, collapsing seven effects into a few interpretable levels and roughly halving the variance of the pooled cells while preserving the headline effect, whereas in the low-wage-jobs study of Cengiz et al. (2019) the 138 event effects are statistically indistinguishable from a common value and the method correctly reports that pooling is adequate, functioning as a formal specification test. The same procedure thus both exploits real heterogeneity and declines to manufacture it where none exists.

Several directions merit future work. First, the agglomerative algorithm is a heuristic; developing exact or near-exact algorithms for small K, or theoretical guarantees on approximation quality, would be valuable. Second, honest inference after partition selection deserves a fuller theoretical treatment, since the naive ℓ_0-PH intervals are conditional on the selected grouping (Remark 2), so developing valid selective-inference corrections for them (the Bayesian posterior already propagates partition uncertainty) is an open problem. Third, formal asymptotic theory establishing consistency of the partition estimator, as both N and the effective sample size per CATT grow, would solidify the method's theoretical foundations, and would formalize the separation threshold that our simulations identify as the determinant of the efficiency gain. Finally, whereas covariate adjustment in the outcome equation is immediate (Section 2, equations (8)–(9)), letting observable characteristics inform the grouping itself, through a covariate-dependent partition prior, is a more substantial extension that could improve performance in settings with richer data.

References

Antoniak, Charles E., "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 1974, *2* (6), 1152–1174.

Armstrong, Timothy B. and Michal Kolesár, "Optimal Inference in a Class of Regression Models," *Econometrica*, 2018, *86* (2), 655–683.

_ , Patrick Kline, and Liyang Sun, "Adapting to Misspecification," *Forthcoming at Econometrica*, 2025.

Arora, Parush and Rishabh Bijani, "Estimating Treatment Effects under Staggered Timing and Non-Spherical Errors," 2026. Working paper, SSRN Working Paper No. 6558759.

Barry, Daniel and John A. Hartigan, "Product Partition Models for Change Point Problems," *The Annals of Statistics*, 1992, *20* (1), 260–279.

Bonhomme, Stéphane and Elena Manresa, "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 2015, *83* (3), 1147–1184.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event-Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 2024, *91* (6), 3253–3285.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs," *The Quarterly Journal of Economics*, 2019, *134* (3), 1405–1454.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

_ and _ , "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey," *The Econometrics Journal*, 2023, *26* (3), C1–C30.

Escobar, Michael D. and Mike West, "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 1995, *90* (430), 577–588.

Fan, Jianqing and Runze Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 2001, *96* (456), 1348–1360.

Ferguson, Thomas S., "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1973, *1* (2), 209–230.

Fisher, Walter D., "On Grouping for Maximum Homogeneity," *Journal of the American Statistical Association*, 1958, *53* (284), 789–798.

Fithian, William, Dennis Sun, and Jonathan Taylor, "Optimal Inference After Model Selection," *arXiv:1410.2597*, 2014.

Gardner, John, "Two-Stage Differences in Differences," *arXiv preprint arXiv:2207.05943*, 2022.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

Hartigan, John A., "Partition Models," *Communications in Statistics - Theory and Methods*, 1990, *19* (8), 2745–2756.

Hill, Jennifer L., "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 2011, *20* (1), 217–240.

Ke, Yuan, Jianqing Fan, and Yichao Wu, "Homogeneity Pursuit," *Journal of the American Statistical Association*, 2015, *110* (509), 175–194.

Kwon, Soonwoo and Liyang Sun, "Estimating Treatment Effects Under Bounded Heterogeneity," *Working paper, arXiv:2510.05454*, 2026.

Lau, John W. and Peter J. Green, "Bayesian Model-Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 2007, *16* (3), 526–558.

Leeb, Hannes and Benedikt M. Pötscher, "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 2005, *21* (1), 21–59.

Liu, Licheng, Ye Wang, and Yiqing Xu, "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data," *American Journal of Political Science*, 2024, *68* (1), 160–176.

Lu, Xun and Liangjun Su, "Determining the Number of Groups in Latent Panel Structures with an Application to Income Democracy," *Quantitative Economics*, 2017, *8* (3), 729–760.

Miller, Jeffrey W. and Matthew T. Harrison, "Mixture Models with a Prior on the Number of Components," *Journal of the American Statistical Association*, 2018, *113* (521), 340–356.

Müller, Peter and Fernando A. Quintana, "Nonparametric Bayesian Data Analysis," *Statistical Science*, 2004, *19* (1), 95–110.

Neal, Radford M., "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 2000, *9* (2), 249–265.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, 2023, *90* (5), 2555–2591.

Rinaldo, Alessandro, Larry Wasserman, and Max G'Sell, "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 2019, *47* (6), 3438–3469.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe, "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *Journal of Econometrics*, 2023, *235* (2), 2218–2244.

Schwarz, Gideon, "Estimating the Dimension of a Model," *The Annals of Statistics*, 1978, *6* (2), 461–464.

Shen, Xiaotong and Hsin-Cheng Huang, "Grouping Pursuit Through a Regularization Solution Surface," *Journal of the American Statistical Association*, 2010, *105* (490), 727–739.

Su, Liangjun, Zhentao Shi, and Peter C. B. Phillips, "Identifying Latent Structures in Panel Data," *Econometrica*, 2016, *84* (6), 2215–2264.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight, "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society: Series B*, 2005, *67* (1), 91–108.

Wade, Sara and Zoubin Ghahramani, "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)," *Bayesian Analysis*, 2018, *13* (2), 559–626.

Wang, Wuyi, Peter C. B. Phillips, and Liangjun Su, "Homogeneity Pursuit in Panel Data Models: Theory and Application," *Journal of Applied Econometrics*, 2018, *33* (6), 797–815.

Wooldridge, Jeffrey M., "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," *Empirical Economics*, 2025, *69*, 2545–2587.

A Identification under Limited Overlap

In some staggered designs a cohort-time cell lacks a clean comparison group, for example when every unit is eventually treated and the last-treated cells have no not-yet-treated or never-treated controls. Such a cell k then contributes no independent identifying variation, its within-transformed column $\tilde{D}_k$ is collinear with the others, $\tilde{D}'\tilde{D}$ is singular, and the flexible estimator $\hat{\tau}_{k,\text{flex}}$ is undefined for that coordinate. This is the staggered-DiD counterpart of the lack-of-overlap problem in cross-sectional designs, where the fully interacted ("long") regression is not well defined (Kwon and Sun, 2026). Partial homogeneity offers a route through it, by letting the unidentified cell inherit its group's effect. Let $\mathcal{U} \subseteq \{1, \ldots, K\}$ be the set of cells not identified by the flexible model and $\mathcal{U}^c$ its complement.

Proposition 3 (Identification via partial homogeneity)**.** *Suppose the true partition is $\mathcal{P} = \{C_1, \ldots, C_m\}$ and each group contains at least one identified cell, $C_p \cap \mathcal{U}^c \neq \emptyset$ for every p. Then under the restricted model* (11) *the group effect ϕ_p is identified for every p, and hence $\tau_k = \phi_p$ is identified for every $k \in C_p$, including the unidentified cells $k \in C_p \cap \mathcal{U}$.*

Proof. Under $\mathcal{P}$ the group column is $\tilde{D}_p = \sum_{k \in C_p} \tilde{D}_k$, so $\|\tilde{D}_p\|^2 \geq \sum_{k \in C_p \cap \mathcal{U}^c} n_k > 0$ whenever $C_p \cap \mathcal{U}^c \neq \emptyset$. Under the orthogonal balanced-panel design $\tilde{D}^{*\prime}\tilde{D}^*$ is then diagonal with strictly positive entries, hence invertible, and each ϕ_p is estimable from (12). Assigning $\tau_k = \phi_p$ identifies every cell in the group. □

Two caveats bound the result's practical reach. First, the identifying content comes entirely from the restriction, not the data. An unidentified cell carries no independent variation, so which group it belongs to cannot be learned from the outcomes, and Proposition 3 is an identification result conditional on the grouping whose recovered effect is only as credible as the (untestable) assignment. Second, the argument is clean only in the orthogonal case, where an unidentified cell is a zero column. Under general collinearity, whether pooling restores identification depends on the partition in ways the proposition does not settle. The Bayesian estimator's co-clustering probabilities describe this assignment, but for a genuinely unidentified cell they are governed by the prior rather than by evidence, so they should be read as prior belief rather than as data.

B Computing the ℓ_0-PH Estimator

The ℓ_0-PH estimator is computed in two steps, an agglomerative search (a greedy sequence of pairwise merges) for the partition and an ordinary least squares re-estimation of the grouped model under it. This appendix gives both and shows why the re-estimation is necessary.

Step 1: Partition recovery. Exact minimization of $Q(\mathcal{P})$ over all partitions is NP-hard (Fan and Li, 2001); we solve it with an agglomerative algorithm that runs in $O(K^3)$ time, in the spirit of classical optimal-grouping procedures (Fisher, 1958). Initialise with K singleton groups, each assigned the flexible CATT estimate $\hat{\tau}_{k,\text{flex}}$ and effective sample size $n_k = \|\tilde{D}_k\|^2$. At each iteration, compute for every pair of current groups (A, B)

$$\Delta\text{Obj}(A, B) \;=\; \frac{n_A\, n_B}{n_A + n_B}\big(\hat{\tau}_A - \hat{\tau}_B\big)^2 \;-\; \lambda \cdot |A| \cdot |B|, \tag{42}$$

where n_A, n_B and $\hat{\tau}_A$, $\hat{\tau}_B$ are the current effective sample sizes and pooled estimates for groups A and B respectively. If $\min_{A \neq B} \Delta\text{Obj}(A, B) \geq 0$, terminate and return the current partition $\hat{\mathcal{P}}$. Otherwise merge the minimizing pair, updating

$$\hat{\tau}_{A\cup B} \;=\; \frac{n_A\,\hat{\tau}_A + n_B\,\hat{\tau}_B}{n_A + n_B}, \qquad n_{A\cup B} \;=\; n_A + n_B. \tag{43}$$

The algorithm terminates after at most $K - 1$ merges.

Step 2: Re-estimation under the discovered partition. Let $\hat{\mathcal{P}} = \{\hat{C}_1, \ldots, \hat{C}_{\hat{m}}\}$ denote the partition returned by Step 1. For each group $\hat{C}_p$, define the group regressor $\tilde{D}_p = \sum_{k \in \hat{C}_p} \tilde{D}_k$ and collect the $\hat{m}$ group regressors into the restricted within-transformed design matrix $\tilde{D}^* = [\tilde{D}_1, \ldots, \tilde{D}_{\hat{m}}]$. Estimate the restricted model

$$\tilde{Y} \;=\; \tilde{D}^*\phi + \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \sim \big(0,\, \sigma^2 I_{NT}\big), \tag{44}$$

by OLS to obtain $\hat{\phi} = (\tilde{D}^{*\prime}\tilde{D}^*)^{-1}\tilde{D}^{*\prime}\tilde{Y}$. The ℓ_0-PH CATT estimate for cell k is $\hat{\tau}_k^{\ell_0} = \hat{\phi}_p$ for all $k \in \hat{C}_p$. Equivalently, (44) corresponds to the TWFE regression

$$Y_{igt} \;=\; \alpha_i + \gamma_t + \sum_{p=1}^{\hat{m}} \phi_p \sum_{k \in \hat{C}_p} D_{k,igt} + \varepsilon_{it}, \tag{45}$$

estimated on the full panel. We refer to the complete two-step procedure as the ℓ_0-PH estimator.

Why re-estimation is essential. A natural one-step alternative uses the greedy-pooled values (43) directly as CATT estimates. We show that this approach is generically inefficient and that the OLS re-estimation in Step 2 is a necessary correction, not a refinement.

The greedy update $\hat{\tau}_{A\cup B}$ in (43) is the weighted mean of the flexible estimates with weights $n_k = \|\tilde{D}_k\|^2$. Under the orthogonality condition $\tilde{D}_j'\tilde{D}_k = 0$ for $j \neq k$, which holds approximately in balanced panels in which each unit belongs to exactly one cohort, $\tilde{D}^{*\prime}\tilde{D}^*$ is diagonal with p-th diagonal entry $\|\tilde{D}_p\|^2 = \sum_{k \in \hat{C}_p} n_k$. In this case, the OLS estimate from

Step 2 reduces to

$$\hat{\phi}_p \;=\; \frac{\tilde{D}_p'\tilde{Y}}{\|\tilde{D}_p\|^2} \;=\; \frac{\displaystyle\sum_{k\in\hat{C}_p} n_k\,\hat{\tau}_{k,\text{flex}}}{\displaystyle\sum_{k\in\hat{C}_p} n_k}, \tag{46}$$

which coincides with the greedy-pooled value (43). By Proposition 1, $\hat{\phi}_p$ is therefore the BLUE of the group treatment effect, and Steps 1 and 2 deliver the same answer when the design is orthogonal.

In general panels, however, cells belonging to the same treatment cohort share unit fixed effects, and cells falling on the same calendar period share time fixed effects. Both sources of co-occurrence produce non-zero off-diagonal elements in $\tilde{D}^{*\prime}\tilde{D}^{*}$, so the OLS estimate from Step 2,

$$\hat{\phi} \;=\; \left(\tilde{D}^{*\prime}\tilde{D}^{*}\right)^{-1}\tilde{D}^{*\prime}\tilde{Y}, \tag{47}$$

no longer reduces to the weighted mean (43). Three consequences follow. First, the greedy-pooled estimator does not satisfy the normal equations of (44) and therefore fails to be BLUE. Second, the ΔObj criterion (42) is an approximation, in that it treats $\tilde{D}^{*\prime}\tilde{D}^{*}$ as diagonal when computing the RSS cost of each candidate merge, so partition selection in Step 1 rests on an approximate objective. Third, the greedy-pooled mean converges to $(n_A\tau_A^{*} + n_B\tau_B^{*})/(n_A + n_B)$ rather than to the common group effect ϕ_p targeted by (44), which undermines consistency whenever cells are not orthogonal.

Step 2 corrects all three shortcomings simultaneously. Given any partition $\hat{\mathcal{P}}$, the estimator $\hat{\phi}$ in (47) is the PH estimator of Proposition 1 applied to $\hat{\mathcal{P}}$. It is BLUE with variance $\sigma^2[(\tilde{D}^{*\prime}\tilde{D}^{*})^{-1}]_{pp}$, which under orthogonality simplifies to $\sigma^2/\sum_{k\in\hat{C}_p} n_k$ and is strictly less than σ^2/n_k for every $k \in \hat{C}_p$ whenever $|\hat{C}_p| \geq 2$, as established in (16). The efficiency gain is guaranteed by Gauss–Markov and grows with the size of the pooled group, and in the balanced case the variance ratio equals $n_k/\sum_{j\in\hat{C}_p} n_j < 1$, matching the ratio in (16). The one-step greedy-pooled estimator attains this gain only under exact orthogonality; Step 2 attains it unconditionally.

The Bayesian estimator underscores the same point. Given a partition, its posterior mean of ϕ in (31) solves the same grouped normal equations as (44) and, in the diffuse limit $\sigma_0^2 \to \infty$, equals $\hat{\phi}$ in (47). Both target the PH estimator under the discovered partition, whereas the one-step greedy-pooled mean does not.

C Heterogeneous Errors

The model of Section 3 assumes a scalar error variance, $\tilde{\varepsilon} \mid \tilde{D} \sim N(0, \sigma^2 I_{NT})$ in (19). This appendix relaxes that assumption to a heterogeneous error covariance and records the few quantities that change, together with the augmented Gibbs sampler. Neither the partition

prior (25) nor the sampler's cluster-assignment logic is affected. The only changes are the design moments that enter the marginal likelihood and the group-effect posterior, and one additional Gibbs block that updates the variance parameters.

The modified model. Replace the scalar covariance in (19) with a structured positive-definite matrix,

$$\tilde{\varepsilon} \mid \tilde{D} \sim N(0, \Omega), \qquad \Omega = \bigoplus_{j=1}^{J} \sigma_j^2 I_{n_j}, \tag{48}$$

where the NT within-transformed observations are partitioned into J variance blocks $\mathcal{B}_1, \ldots, \mathcal{B}_J$ of sizes $n_1, \ldots, n_J$, indexed by a known label $v(i) \in \{1, \ldots, J\}$ for observation i. The blocks collect observations expected to share a noise level, for instance by cohort, by treated versus comparison status, or by an observed cluster. The homoskedastic model is the special case $J = 1$. A fully general non-diagonal Ω, such as the clustered influence-function covariance already used in the empirical applications (Remark 1), enters the formulas below in the same way through Ω^{-1}, and only the update (53) is specific to the block-diagonal parameterization. We complete the model with an independent conjugate prior on each variance level,

$$\sigma_j^2 \sim \text{Inv-Gamma}(a_0, b_0), \qquad j = 1, \ldots, J, \tag{49}$$

the natural generalization of (28).

What changes in the theory. Every appearance of the precision $\sigma^{-2} I_{NT}$ becomes Ω^{-1}. Writing the two design moments in the Ω^{-1} metric,

$$S \;=\; \tilde{D}'\Omega^{-1}\tilde{D}, \qquad u \;=\; \tilde{D}'\Omega^{-1}\tilde{Y}, \tag{50}$$

the group effects still integrate out in closed form. Conditional on Ω, the marginal likelihood (29) becomes $\tilde{y} \mid \mathcal{P} \sim N(\mu_0 \tilde{D}\mathbf{1}_m,\ \Omega + \sigma_0^2 \tilde{D}\tilde{D}')$, and the Woodbury reduction (30), dropping terms constant in $\mathcal{P}$, reads

$$\log p(\tilde{y} \mid \mathcal{P}, \Omega) \;\propto\; -\tfrac{1}{2} \log\big| I_m + \sigma_0^2 S \big| \;+\; \tfrac{\sigma_0^2}{2}\, u' \big(I_m + \sigma_0^2 S\big)^{-1} u. \tag{51}$$

The group-effect posterior (31) becomes $\phi \mid \mathcal{P}, \Omega \sim N(\mu^*, \Sigma^*)$ with

$$\Sigma^* \;=\; \big(S + I_m/\sigma_0^2\big)^{-1}, \qquad \mu^* \;=\; \Sigma^* \big(u + \mu_0 \mathbf{1}_m/\sigma_0^2\big), \tag{52}$$

which is the ridge-regularized GLS estimator of the group effects. Setting $\Omega = \sigma^2 I_{NT}$ gives $S = \tilde{D}'\tilde{D}/\sigma^2$ and $u = \tilde{D}'\tilde{Y}/\sigma^2$ and recovers (30) and (31) exactly. The computational cost is unchanged, the $m \times m$ matrix S still governs the calculation, and forming it costs one weighting of the design by Ω^{-1}, which is $O(NT)$ for the diagonal case (48).

One feature of the fixed-σ^2 theory does not survive. The exact ℓ_0 correspondence of Section 3.2, in which the MAP partition minimizes the residual sum of squares plus a fixed

penalty $L = \lambda\sigma^2$ per group, relied on the scalar variance factoring out of the objective. Under (48) the fit term is the Ω^{-1}-weighted sum of squares and the variance levels also enter through $\log|\Omega|$, so the penalty is no longer a fixed multiple of a residual sum of squares. Heterogeneous errors are therefore handled inside the sampler, which averages over the variance levels along with the partition, rather than by the fixed-penalty point estimator.

The augmented Gibbs sampler. The collapsed sampler of Section 3.3 gains one block. Given a current Ω, one sweep runs the following three steps.

1. For each CATT k, remove k from its cluster, delete the cluster if it is empty, and draw z_k from the conditional probabilities (35) with the marginal likelihoods now evaluated from (51).
2. Draw the group effects $\phi \sim N(\mu^*, \Sigma^*)$ from (52).
3. Form the within-transformed residuals $\hat{\tilde{\varepsilon}} = \tilde{Y} - \tilde{D}_{\mathcal{P}}\phi$ under the current partition, and for each variance block draw

$$\sigma_j^2 \mid \phi, \mathcal{P}, \tilde{Y} \;\sim\; \text{Inv-Gamma}\Big(a_0 + \tfrac{n_j}{2},\; b_0 + \tfrac{1}{2}\sum_{i\,:\,v(i)=j} \hat{\tilde{\varepsilon}}_i^{\,2}\Big), \tag{53}$$

   then assemble $\Omega = \bigoplus_j \sigma_j^2 I_{n_j}$ for the next sweep.

Reinstating ϕ in Step 2 before the variance draw is what makes Step 3 conjugate, exactly as the original sampler reinstates ϕ to report the group effects. All posterior summaries of Section 3.3, the CATT posterior mean (36), the co-clustering matrix (37), and the aggregated estimands, are formed from the augmented draws $\{z^{(s)}, \phi^{(s)}, \Omega^{(s)}\}$ without further change, so the reported credible intervals now propagate uncertainty in the error variances alongside the partition. When a plug-in covariance is preferred, fixing Ω at an estimate $\hat{\Omega}$ and omitting Step 3 recovers a single GLS-posterior sampler, the direct analog of the fixed-$\hat{\sigma}^2$ special case of Section 3.2.

D Fixed versus Random σ^2

The main text draws σ^2 from its inverse-gamma full conditional each sweep, the full Bayesian model of Section 3.1. Holding σ^2 fixed at the plug-in $\hat{\sigma}^2$ is the special case that yields the exact ℓ_0 correspondence (Section 3.2). This appendix shows the two give the same inference. We rerun the coverage experiment of Section 4.3 at $m^* = 6$ and $\delta = 6$ over 500 replications and, for each draw, form the DP credible intervals both ways from the same partition sampler.

Table 8: Coverage and average length of nominal 95% DP credible intervals under a random (full Bayesian) and a fixed (plug-in) error variance, $m^* = 6$, $\delta = 6$, 500 replications.

	CATT		ATT	
σ^2 treatment	Coverage	Length	Coverage	Length
Random (full Bayesian)	0.94	0.20	0.95	0.11
Fixed (plug-in)	0.93	0.20	0.93	0.11

Table 8 reports the result. The two samplers deliver essentially the same coverage and interval length, for the cohort-time effects and for the overall ATT. The reason is the one given in Section 3.2, the residual degrees of freedom $NT - N - T + 1 - K$ are large, so σ^2 is pinned down and the estimation error a prior on it would propagate is negligible. The choice is therefore immaterial for the reported results, and we use the fixed-σ^2 case in the main text only to make the ℓ_0 connection concrete.