Finitely Heterogeneous Treatment Effect in Event-study*

Myungkou Shin†

January 6, 2025

Abstract

A key assumption of the differences-in-differences designs is that the average evolution of untreated potential outcomes is the same across different treatment cohorts: parallel trend assumption. In this paper, we relax the parallel trend assumption by assuming a latent type variable and developing a *type-specific* parallel trend. With a finite support assumption on the latent type and long pre-treatment time periods, an extremum classifier consistently estimates the type assignment. Based on the classification, we propose a type-specific DiD estimator for type-specific ATT. By estimating the type-specific ATT, we study heterogeneity in treatment effect, in addition to heterogeneity in baseline outcomes.

Keywords: event-study, difference-in-differences, panel data, heterogeneity, classification, K-means clustering

JEL classification codes: C13, C14, C23

*I am deeply grateful to Stéphane Bonhomme, Christian Hansen and Azeem Shaikh, who have provided me invaluable support and insight. I would also like to thank Manasi Deshpande, Ali Hortaçsu, Guillaume Pouliot, Max Tabord-Meehan, Alex Torgovitsky, Martin Weidner and the participants of the metrics advising group at the University of Chicago and the Nuffield Econometrics Seminar at the University of Oxford for their excellent feedback. I acknowledge the support from the European Research Council through the grant ERC-2018-CoG-819086-PANEDA. Any and all errors are my own.

†School of Social Sciences, University of Surrey. email: m.shin@surrey.ac.uk

1 Introduction

The event-study design, which utilizes the variation in the timing of an event, is an empirical framework whose popularity among empirical researchers has risen tremendously over the time. In applied microeconomics, the event-study design is most often implemented with the difference-in-differences (DiD) approach. A key identifying assumption of the DiD style event-study research design is the parallel trend assumption: temporal differences of untreated potential outcomes are mean independent of treatment timing.

The traditional parallel trend assumption that assumes parallel trends with unconditional means is limited in its scope; it is directly violated when there exist heterogeneous patterns of time trends across units. A widely used alternative to this unconditional parallel trend assumption is to assume a conditional parallel trend with observable conditioning covariates (see Abadie (2005); Callaway and Sant'Anna (2021) among others). Following the same spirit, we also assume conditional parallel trend assumption to relax the unconditional parallel trend assumption, but assume that the conditioning variable is latent. Throughout the paper, we call the conditioning variable a latent 'type' variable and call the conditional parallel trend a 'type-specific parallel trend' assumption. To illustrate the assumption, let us consider a simple two time periods case: with a latent type variable k_i,

$$\mathbf{E}[Y_{i2}(\infty) - Y_{i1}(\infty)|k_i, D_i = 1] = \mathbf{E}[Y_{i2}(\infty) - Y_{i1}(\infty)|k_i, D_i = 0]. \tag{1}$$

$Y_{it}(\infty)$ is the untreated potential outcome for unit i at time t and D_i is an indicator for being treated at time $t = 2$. When k_i is observable, Equation (1) is the conditional parallel trend assumption as in Abadie (2005); Callaway and Sant'Anna (2021).

In this paper, allowing for the conditioning variable k_i to be latent comes at three additional requirements on the model and the data. Firstly, we assume that the latent type variable k_i has a finite support. Secondly, we assume that the types are sufficiently separated in terms of pretreatment outcomes. Lastly, we assume that the number of pretreatment pe-

riods grows to infinity. Thus, the type-specific parallel trend assumption can be understood as an alternative to the unconditional parallel trend or the conditional parallel trend with observable conditioning covariates; it allows for heterogeneity patterns that may not be captured by the observable information, at the price of assuming that the heterogeneity is finite and the pretreatment time series is long.

The finite support assumption gives us unique merits. Firstly, it motivates a simple within-type DiD style estimator, which is potentially less susceptible to the overlap problem compared to the conditional DiD estimators with a continuous conditioning variable. Secondly, the finite support assumption helps us analyze the patterns of treatment effect heterogeneity. While the type-specific parallel trend framework of this paper does not put any restriction on the treatment effect heterogeneity, thus allowing for fully flexible treatment effect heterogeneity, the finite support assumption provides a model-based stratifying structure that is helpful in summarizing the pattern of treatment effect heterogeneity. Let $Y_{i2}(2)$ be the treated potential outcome of unit i at time $t = 2$. The existing literature mostly focuses on the conditional expectation of $Y_{i2}(2) - Y_{i2}(\infty)$ given some observable information: e.g., $\mathbf{E}\left[Y_{i2}(2) - Y_{i2}(\infty)|X_i, D_i = 1\right]$. With the type-specific parallel trend framework, we document treatment effect heterogeneity along the latent type variable k_i:

$$\mathbf{E}\left[Y_{i2}(2) - Y_{i2}(\infty)|k_i, D_i = 1\right].^{1}$$

The stratifying structure induced by the finite type can be helpful for policy recommendation when policymakers have little to zero constraints in policy assignment but do not have perfect information for individual-level responsiveness; it provides a feasible alternative to individual-level treatment assignment, by identifying the treatment effect at the type level.

To apply the type-specific parallel trend framework to datasets, we propose a two-step estimation procedure. In the first step, we classify units into the finite number of types,

[1]The treatment effect parameter is comparable to those defined in Abadie (2005); Sant'Anna and Zhao (2020); Callaway and Sant'Anna (2021) when k_i is observable.

using the K-means clustering algorithm on the long pretreatment periods. Given the classification result, the second step of the estimation procedure estimates the treatment effect, using the estimated types as given.[2] In the estimation step, a variety of existing estimation strategies can be employed by treating the type as a categorical variable: De Chaisemartin and d'Haultfoeuille (2020); Borusyak et al. (2024); Callaway and Sant'Anna (2021); Sun and Abraham (2021). As with De Chaisemartin and d'Haultfoeuille (2020) and Callaway and Sant'Anna (2021), our 'type-specific DiD' estimator takes an average of the canonical DiD estimates with two time periods and two treatment timings.

To discuss asymptotic properties of the treatment effect estimators, we first show that the probability of first-step misclassification goes to zero when the number of pretreatment time periods grows at a polynomial rate of the number of units. Given that the number of pretreatment time periods grows sufficiently fast compared to the number of units, the type-specific DiD estimators are consistent and asymptotically normal under some regularity conditions. These asymptotic results are supported by Monte Carlo simulations.

This paper contributes to the large literature of panel data models where interactive fixed-effects models are used to control for unit heterogeneity across treatment cohorts, relaxing the two-way fixed-effect (TWFE) specification used in the parallel trend approach: see Abadie et al. (2010); Arkhangelsky et al. (2021); Athey et al. (2021); Hsiao et al. (2012); Freyaldenhoven et al. (2019); Xu (2017); Callaway and Karami (2023); Janys and Siflinger (2024) among others. The interactive fixed-effect model often assumes a factor model for the interactive fixed-effects; our framework can be thought of as a special case of the factor model with a finite support on the factor. Janys and Siflinger (2024) is closest to this paper in that it takes the same approach; however, Janys and Siflinger (2024) neither explores

[2]This two-step property of the estimation procedure closely relates to the stratification exercise used in estimating subpopulation treatment effect (see Abadie et al. (2018)). The goal of the stratification (i.e. classification in this paper's terminology) is to find groups of units whose (estimated) counterfactual untreated outcomes are similar. The type-specific parallel trend assumption directly relates to this since under the type-specific parallel trend assumption, units with the same type share the same time trend.

treatment effect heterogeneity nor develop a full asymptotic theory. As with this paper, most of the papers in the interactive fixed-effect model literature rely on large pretreatment periods. Notable exceptions are Callaway and Karami (2023) and Freyaldenhoven et al. (2019). In both papers, informative external variables or instruments are used, instead of assuming long pretreatment periods.

This paper also closely relates to the rapidly growing literature on heterogeneous treatment effect: see De Chaisemartin and d'Haultfoeuille (2020); Sun and Abraham (2021); Callaway and Sant'Anna (2021); Goodman-Bacon (2021); Borusyak et al. (2024); Baker et al. (2022); Goldsmith-Pinkham et al. (2024), among others. Callaway and Sant'Anna (2021) is particularly close to this paper in the sense that they also consider a conditional parallel trend assumption. This literature discusses the negative weighting problem that arises in the standard TWFE specification when there is treatment effect heterogeneity across units. We build upon this literature and construct the type-specific DiD estimator to be robust to the treatment effect heterogeneity.

The rest of the paper is organized as follows. In Section 2, we formally discuss the type-specific parallel trend assumption. In Section 3, we propose the two-step estimation procedure for treatment effect estimation. In Section 4-5, we discuss the asymptotic properties of the estimator. In Section 6, we present Monte Carlo simulation results on the finite-sample performance of the estimator. In Section 7, we provide an empirical illustration of the type-specific DiD estimator by revisiting Lutz (2011).

2 Model

In the main model of the paper, an econometrician observes a panel data with a binary treatment: $\left\{\{Y_{it}, D_{it}\}_{t=-T_0-1}^{T_1-1}\right\}_{i=1}^{n}$. Y_{it} is the outcome variable for unit i at time t and $D_{it} \in \{0,1\}$ is the binary treatment variable for unit i at time t. D_{it} follows the staggered adoption scheme; $D_{it} \leq D_{it+1}$. $E_i = \min\{t : D_{it} = 1\}$ denotes the treatment timing of unit

i; this is the same notation as in Sun and Abraham (2021) and equivalent with treatment 'group' notation G in Callaway and Sant'Anna (2021). For never-treated units, let $E_i = \infty$. There are n units and $T_0 + T_1 + 1$ time periods, with the unit index ranging $i = 1, \cdots, n$ and the time index ranging $t = -T_0 - 1, \cdots, T_1 - 1$. $t < 0$ denotes pretreatment periods at the population level where no unit is treated and $t \geq 0$ denotes treatment periods at the population level where at least one unit is treated. Thus, $T_0 + 1$ is the number of population pretreatment periods that grows and T_1 is the fixed number of population treatment periods. Throughout the paper, we use the potential outcome framework to discuss treatment effect:

$$Y_{it} = Y_{it}(E_i).$$

$Y_{it}(e)$ is the time t potential outcome of unit i when they are treated at time e. Thus, $Y_{it}(e)$ is a pretreatment potential outcome if and only if $t < e$.

The key assumption of the paper assume a unit-level latent type variable, conditioning on which the parallel trend assumption and the no anticipation assumption hold.

Assumption 1. (Type-Specific Parallel Trend) *There exists a latent type variable k_i such that for any t, s, $\mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, E_i\right] = \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i\right]$.*

Assumption 2. (No Anticipation) *For any $t < e$, $\mathbf{E}\left[Y_{it}(e) - Y_{it}(\infty) | k_i, E_i\right] = 0$.*

Assumption 1 assumes that there is some latent type variable k_i, conditioning on which the time trend of the never-treated potential outcome $Y_{it}(\infty) - Y_{is}(\infty)$ is mean independent of the treatment timing E_i.[3] Assumption 2 assumes that the never-treated potential outcome $Y_{it}(\infty)$ and the pretreatment potential outcome $Y_{it}(e)$ with $t < e$ have the same conditional mean, given the latent type k_i and the treatment timing E_i.

[3]The parallel trend is assumed for every time period, across every treatment timing; when $K = 1$, this is equivalent to the conventional parallel trend assumption assumed in De Chaisemartin and d'Haultfoeuille (2020); Sun and Abraham (2021), but different from Callaway and Sant'Anna (2021), which restricts the window of time periods to apply the parallel trend to.

In addition to Assumptions 1-2, we assume the following for type classification.

Assumption 3. (Finite Support) $k_i \in \{1, \cdots, K\}$.

Assumption 4. (Well-Separated Types) *Whenever* $k \neq k'$,

$$\lim_{T_0 \to \infty} \frac{1}{T_0} \sum_{t=-T_0}^{-1} \Big(\mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k\right] - \mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k'\right] \Big)^2 =: c(k, k') > 0.$$

Assumption 3 assumes that the latent type variable k_i has a finite support, as in the group fixed-effect literature or the finite mixture literature. Assumption 4 assumes that the types are well-separated in terms of $\mathbf{E}[Y_{it}(\infty) - Y_{it-1}(\infty)|k_i = k]$; for any two different types, the l_2 norm of the difference between their conditional means is strictly nonzero. Note that the separation assumption is in relation to time trends of the never-treated potential outcomes $Y_{it}(\infty)$. From Assumptions 1-2, we have

$$\mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k\right] = \mathbf{E}\left[Y_{it}(e) - Y_{it-1}(e) | k_k = k, E_i = e\right] \tag{2}$$

whenever $t < e$. Thus, Assumption 4 can be extended to the pretreatment outcomes of all units regardless of their treatment timings.

Given the finite type structure, let us define type-specific treatment effect parameters. Fix two time periods (s, t) and a treatment cohort $\{i : E_i = e\}$ such that $s < e \leq t$. The type-cohort-and-time specific average treatment effect on treated units (ATT) for time t, type k and treatment timing e can be written as follows:

$$\begin{aligned} ATT_t(k, e) &= \mathbf{E}\left[Y_{it}(e) - Y_{it}(\infty) | k_i = k, E_i = e\right] \\ &= \mathbf{E}\left[Y_{it} - Y_{is} | k_i = k, E_i = e\right] - \mathbf{E}\left[Y_{it} - Y_{is} | k_i = k, E_i > t\right]. \end{aligned} \tag{3}$$

The second equality holds from Assumptions 1-2 and thus $ATT_t(k, e)$ becomes a function of the conditional moments of $\{Y_{it}\}_t$ given the latent type variable k_i and the treatment timing

variable E_i. $ATT_t(k,e)$ is a type-specific version of the treatment effect estimands in the DiD literature: see Callaway and Sant'Anna (2021); Sun and Abraham (2021) for more.

The type-cohort-and-time specific ATT parameter in (3) takes treatment timing E_i as a conditioning variable and focuses on a specific time period t. The full-fledgedness of the type-cohort-and-time specific ATT is useful when the researcher is interested in treatment effect heterogeneity across both time periods and types. Though both dimensions of the treatment effect heterogeneity may be of interest depending on contexts, we focus on an aggregated ATT parameter in this paper, to highlight the treatment effect heterogeneity across types. To aggregate, we take the average of (3) across (t,e) while maintaining the relative treatment timing $t-e$ fixed: for some $r \geq 0$,

$$\beta_r(k) := \sum_{e=0}^{T_1-1-r} \frac{\Pr\{E_i = e | k_i = k\}}{\Pr\{E_i \leq T_1 - r | k_i = k\}} \cdot \mathbf{E}\left[Y_{i,e+r}(e) - Y_{i,e+r}(\infty) | k_i = k, E_i = e\right]. \tag{4}$$

$\beta_r(k)$ is the r-times lagged ATT for type k. In aggregating $ATT_t(k,e)$, we use the weights discussed in Callaway and Sant'Anna (2021); Sun and Abraham (2021). We call $\beta_r(k)$ as 'dynamic' type-specific ATT since the parameter compares instantaneous treatment effect and long-run treatment effect, as we compare $\beta_r(k)$ with $r \approx 0$ and $\beta_r(k)$ with $r \gg 0$.

It is straightforward to see that $ATT_t(k,e)$ and $\beta_r(k)$ are identified from the distribution of $\left(\{Y_{it}\}_{t=-T_0-1}^{T_1-1}, E_i, k_i\right)$ whenever their respective overlap conditions discussed in Section C.1 are satisfied and $\{k_i\}_{i=1}^n$ is observed for a given T_0. Though the type assignment $\{k_i\}_{i=1}^n$ is not directly observed from data, the following lemma shows that the treatment effect parameters are identified when $T_0 \to \infty$ and the pretreatment time series shows weak dependence.

Lemma 1. *Assumptions 1-4 hold. Also,* $\Pr\{k_i = k\} > 0$ *and*

$$\lim_{T_0 \to \infty} \frac{1}{T_0} \sum_{t=-T_0}^{-1} \left(\mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k\right]\right)^2$$

exists for each k and

$$\operatorname*{plim}_{T_0\to\infty} \frac{1}{T_0} \sum_{t=-T_0}^{-1} a_t\big(Y_{it}(E_i) - Y_{it-1}(E_i) - \mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|k_i\right]\big) = 0$$

for any $\{a_t\}_{t=-T_0}^{-1}$ such that $\lim_{T_0\to\infty} \frac{1}{T_0}\sum_{t=-T_0}^{-1} {a_t}^2$ is finite. Then, $ATT_t(k,e)$ and $\beta_r(k)$ are identified from the distribution of $\left(\{Y_{it}\}_{t=-\infty}^{T_1-1}, E_i\right)$ when their respective overlap conditions are satisfied.

Specifics of the overlap condition are given in Section C.1 of the Appendix. The proof for Lemma 1 is found in Section B of the Supplementary Appendix. Lemma 1 is an identification counterpart of the type classification consistency result in Section 4: Theorem 1.

3 Estimation

The estimation procedure is two-step. The first step is to estimate the type using the K-mean minimization problem. The second step is to take the estimated type as given and estimate ATT. To describe the estimation procedure, let us adopt the following notations:

$$\gamma = (k_1, \cdots, k_n) \in \Gamma, \qquad \text{and} \qquad \delta = \{\delta_t(k)\}_{t,k} \in \mathcal{D}.$$

γ is a $n \times 1$ vector of a type assignment while k_i denotes the type of unit i. Its support is $\Gamma = \{1, \cdots, K\}^n$, a set of all possible type assignments. δ is a collection of $\delta_t(k)$, the type-specific time trend given time t and type k:

$$\delta_t(k) = \mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|k_i = k\right].$$

Its support is $\mathcal{D} = [-M, M]^{T_0 K}$ with some $M > 0$; time trends are bounded. In the classification step, we only use a subset of the given data: population pretreatment periods

where every ever-treated unit is untreated.[4] With the population pretreatment periods, we construct an objective function and the resulting first-step classifier:

$$\widehat{Q}(\delta,\gamma) = \frac{1}{nT_0}\sum_{i=1}^{n}\sum_{t=-T_0}^{-1}\left(Y_{it} - Y_{it-1} - \delta_t(k_i)\right)^2 \tag{5}$$

$$\left(\hat{\delta},\hat{\gamma}\right) = \arg\min_{(\delta,\gamma)\in\mathcal{D}\times\Gamma}\widehat{Q}\left(\delta,\gamma\right). \tag{6}$$

The minimization problem in (5) is called K-mean minimization problem; the solution to the K-means minimization problem is a grouping structure with K groups, defined with K centeroids. In the minimization (5), the centeroids are denoted with $\{\delta_t(1)\}_{t<0}, \cdots, \{\delta_t(K)\}_{t<0}$ and the grouping structure with $\{k_i\}_{i=1}^n$.

The algorithm that we use to obtain (6) is a conventional K-means clustering algorithm. Given an initial type assignment $\gamma^{(0)} = \left(k_1^{(0)}, \cdots, k_n^{(0)}\right)$,

1. **(update δ)** Given the type assignment $\gamma^{(s)}$ from the s-th iteration, estimate $\hat{\delta}_t^{(s)}(k)$ by letting
$$\hat{\delta}_t^{(s)}(k) = \frac{\sum_{i=1}^n \left(Y_{it} - Y_{it-1}\right)\mathbf{1}\{k_i^{(s)} = k\}}{\sum_{i=1}^n \mathbf{1}\{k_i^{(s)} = k\}}$$
whenever the denominator is not zero.

2. **(update γ)** Update $k_i^{(s)}$ for each i by letting $k_i^{(s+1)}$ be the solution to the following minimization problem: for $i = 1, \cdots, N$,
$$\min_{k\in\{1,\cdots,K\}}\sum_{t=-T_0}^{-1}\left(Y_{it} - Y_{it-1} - \hat{\delta}_t^{(s)}(k)\right)^2.$$

3. Repeat Step 1-2 until Step 2 does not update $\hat{\gamma}$, or some stopping criterion is met. For stopping criterion, one can set a maximum number of iteration or a minimum update

[4]The performance of the K-means classifier depends on estimating the type-specific time trends $\{\delta_t(k)\}_{t,k}$ consistently. Using pretreatment outcomes for $t \geq 0$ where many of the ever-treated units are already treated may impair the classification result since $\hat{\delta}_t(k)$ for such t will have larger variance. Thus, we only use the population pretreatment periods for type classification.

in $\hat{\delta}^{(s)}$: set S and ε such that the iteration stops when $s \geq S$ or $\left\|\hat{\delta}^{(s)} - \hat{\delta}^{(s-1)}\right\|_\infty \leq \varepsilon$.

The iterative algorithm proposed here has two stages. In the first stage, the algorithm estimates δ by taking sample means. In the second stage, the algorithm reassigns a type for each unit, by finding the type that minimizes the distance between $\{Y_{it} - Y_{it-1}\}_{t<0}$ and $\{\delta_t(k)\}_{t<0}$. The algorithm quickly attains a local minimum of the minimization problem (5). In the application we used in Section 7, the algorithm mostly converged within 20 iterations.

Since the iterative algorithm does not conduct an exhaustive search, it may not converge to a global minimum; the computational burden of the exhaustive search is extremely heavy since the space for the type assignment has cardinality of n^K. Thus, we recommend that a random initial type assignment be drawn multiple times and the local minima be compared. Another concern is the choice of K. So far, the number of types K has been treated as known. When there is no natural choice for K, an information criterion can be used to estimate the number of type K (see Bai and Ng (2002); Bonhomme and Manresa (2015); Janys and Siflinger (2024)). To construct the information criterion, first we set $K_{\max}$, an upper bound on the number of types. Then, we estimate the error variance using the classification with $K_{\max}$: with $\Gamma(K) = \{1, \cdots, K\}^n$,

$$\hat{\sigma}^2 = \min_{\delta,\gamma \in \Gamma(K_{\max})} \widehat{Q}(\delta, \gamma)$$

Given the estimate, we estimate K to be the minimizer of the Bayesian information criterion:

$$\hat{K} = \arg\min_{K \leq K_{\max}} \left(\min_{\delta,\gamma \in \Gamma(K)} \widehat{Q}(\delta, \gamma) + \hat{\sigma}^2 \frac{KT_0 + n}{nT_0} \log nT_0 \right).$$

Given the first-step result, the type-specific DiD estimator for $ATT_t(k, e)$ defined in (3) can be constructed by taking sample means for each type:

$$\widehat{ATT}_t(k, e) = \sum_{i=1}^{n} (Y_{it} - Y_{i,e-1}) \left(\frac{\mathbf{1}\{\hat{k}_i = k, E_i = e\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = e\}} - \frac{\mathbf{1}\{\hat{k}_i = k, E_i > t\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i > t\}} \right).^{5}$$

In the case of the dynamic ATT parameter $\beta_r(k)$ defined in (4), the estimator is

$$\hat{\beta}_r(k) = \sum_{e \leq T_1 - 1 - r} \frac{\hat{\mu}(k,e)}{\sum_{e' \leq T_1 - 1 - r} \hat{\mu}(k,e')} \cdot \widehat{ATT}_{e+r}(k,e) \tag{7}$$

where $\hat{\mu}(k,e) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = e\}$. $\hat{\mu}(k,e)$ is the estimator for the probability $\mu(k,e) = \Pr\{k_i = k, E_i = e\}$.

Similarly, we can extend $\hat{\beta}_r(k)$ for $r < -1$ and construct estimators for

$$\beta_r(k) = \sum_{e=0}^{T_1 - 1} \frac{\Pr\{E_i = e | k_i = k\}}{\Pr\{E_i \leq T_1 - r | k_i = k\}} \cdot \mathbf{E}\left[Y_{i,e+r}(e) - Y_{i,e+r}(\infty) | k_i = k, E_i = e\right]$$

for some $r < -1$. From Assumptions 1-2, $\beta_r(k) = 0$ whenever $r < -1$. Thus, we can use $\hat{\beta}_r(k)$ for $r < -1$ to test Assumptions 1-2, equivalent to the widely used 'no pretreatment test' in the event-study literature.

4 Asymptotic Theory

In this section, we discuss the asymptotic properties of the estimator proposed in Section 3. As we extend the main model by adding control covariates in Section 5 and Sections C.2-4 of the Appendix, the proofs of the results for the main model are implied by the more general results. The proofs for the general results and how they connect to the asymptotic results of this section are discussed in Sections C-D of the Supplementary Appendix.

Firstly, to derive the classification result for the type estimator defined in (6), let us adopt following assumptions.

[5] As discussed in Callaway and Sant'Anna (2021), there is no straightforward choice in which time differences to use in a DiD type approach. Though the estimator described above takes one period before the treatment timing to construct a time difference, other choices are equally valid as long as the parallel trend assumption holds for every time period. Roth and Sant'Anna (2023*a*) discusses efficiency of these DiD type estimates when the treatment timing is truly random. More discussion on this is given the Appendix. Also, the estimator uses the not-yet-treated units as control units; alternatively, one can use never-treated units or any other subset of $\{i : E_i > t\}$.

Assumption 5. *With some $M > 0$,*

a. *(iid across units)* $\left(\{Y_{it}(e)\}_{e,t}, E_i, k_i\right) \overset{iid}{\sim} F$.

b. *(finite moments) For every e, t and k,* $\mathbf{E}\left[Y_{it}(e)^4 \middle| k_i = k\right] \leq M$.

c. *(long pretreatment)* $T_0 \to \infty$ *as* $n \to \infty$.

d. *(no measure zero types) For all* $k \in \{1, \cdots, K\}$, $\Pr\{k_i = k\} > 0$

e. *(weakly dependent, thin-tailed errors) With some positive constant d_1 and a,*

$$\left\{Y_{it}(e) - Y_{it-1}(e) - \mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|k_i\right]\right\}_{t=-T_0}^{-1}$$

is strongly mixing with mixing coefficient $\alpha[t]$ such that $\alpha[t] \leq \exp(-at^{d_1})$ uniformly over e. Also, with some positive constant d_2 and b, $Y_{it}(e)$ satisfies the following tail probability: for any $y > 0$,

$$\Pr\left\{|Y_{it}(e) - \mathbf{E}\left[Y_{it}(\infty)|k_i\right]| \geq y\right\} \leq \exp\left(1 - (y/b)^{d_2}\right)$$

uniformly over e and $t < 0$.

Assumption 5-c assumes that the number of population pretreatment periods T_0 grows to infinity as n goes to infinity. Assumption 5-d assumes that each type realizes with positive probability. Assumption 5-e assumes that for $t < 0$, tail probability of $Y_{it}(e) - \mathbf{E}[Y_{it}(\infty)|k_i]$ goes to zero exponentially and the first difference of $Y_{it}(e) - \mathbf{E}[Y_{it}(\infty)|k_i]$ is weakly dependent in the sense that it is strongly mixing with mixing coefficient decreasing exponentially in t.

Theorem 1. *Let Assumptions 1-5 hold. Then, up to some permutation on* $\{1, \cdots, K\}$,

$$\Pr\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i^0\} > 0\right\} = o(n{T_0}^{-\nu}) + o(1) \quad \forall \nu > 0$$

as $n \to \infty$.

Theorem 1 puts a bound on the misclassification probability; the rate is identical to the rate found in the group fixed-effect literature.

The classification of n units into K types is a crucial part of the estimation procedure that the performance of the treatment effect estimators depends on. Consider a very simple case where $K = 2$ and model the untreated potential outcomes as follows: for $t \leq 0$,

$$Y_{it}(\infty) = \delta(k_i) + U_{it}, \qquad U_{it} \overset{iid}{\sim} \mathcal{N}(0,1).$$

WLOG let $\delta(1) < \delta(2)$. Find that $\bar{Y}_i(\infty) = \frac{1}{T_0+1}\sum_{t=-T_0-1}^{-1} Y_{it}(\infty) \sim \mathcal{N}\left(\delta(k_i), \frac{1}{T_0+1}\right)$. It is easy to see that for any fixed T_0,

$$\begin{aligned}\Pr\left\{\bar{Y}_i(\infty) \geq \bar{Y}_j(\infty) | k_i = 1, k_j = 2\right\} &= \Pr\left\{\bar{U}_i - \bar{U}_j \geq \delta(2) - \delta(1) | k_i = 1, k_j = 2\right\} \\ &= \Phi\left(\sqrt{\frac{T_0+1}{2}}\Big(\delta(2) - \delta(1)\Big)\right)\end{aligned}$$

is nonzero, with Φ being the distribution function of the standard normal $\mathcal{N}(0,1)$; the probability of imperfect classification is nonzero. Thus, we need large pretreatment periods ($\Leftrightarrow T_0 \gg 0$), in addition to the strong separation ($\Leftrightarrow \delta(2) - \delta(1) > 0$).[6] When we do not have both conditions satisfied and thus units are potentially misclassified, the treatment effect estimator suffers from a non-classical measurement error problem.[7]

When T_0 is large, the bound on the misclassification probability from Theorem 1 can be used to derive asymptotic properties of the type-specific DiD estimator from (7).

Corollary 1. *Let Assumptions 1-5 hold. Suppose that for any t, s and e,* $\mathrm{Var}\left(Y_{it}(e) - Y_{is}(e) | k_i, E_i\right) >$

[6]By evaluating the CDF function Φ, we can see that ${T_0}^{\nu}\Phi\left(\sqrt{\frac{T_0+1}{2}}\left(\delta(2) - \delta(1)\right)\right)$ goes to zero for any $\nu > 0$ as T_0 grows, as stated in Theorem 1.

[7]The measurement error problem may be severe for a type-specific ATT even when the misclassification rate is small and the tail of the potential outcome distribution is thin, if the within-type overlap is weak. On the other hand, when the misclassification rate is negligible compared to the within-type shares of ever-treated units and control units, the bias will be small.

0 and that there exists some $\nu^* > 0$ such that $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. Then, for any k and r with some permutation on $\{1, \cdots, K\}$,

$$\sqrt{n}\left(\hat{\beta}_r(k) - \beta_r(k)\right) \xrightarrow{d} \mathcal{N}\left(0, {\sigma_{\beta_r(k)}}^2\right)$$

with some ${\sigma_{\beta_r(k)}}^2 > 0$, as $n \to \infty$, when k and r satisfy the corresponding overlap condition.

Remark 1. The asymptotic variance has a consistent estimator, whose expression is given in the Supplementary Appendix, along with the proof of Corollary 3.

Remark 2. In formulating the dynamic ATT parameter $\beta_r(k)$, treatment timing distribution $\mu(k, e)$ is used as weights. Similar asymptotic results as in Corollary 1 hold for many other choices of weights: e.g. uniform weights across treatment timing.

5 Extension to the Model with Covariates

5.1 Introducing control covariates X_{it}

In this section, we extend our main model by adding observed covariates $X_{it} \in \mathbb{R}^p$ to the model. The control covariates X_{it} gives us an extra source of heterogeneity in outcomes across different units and different times. For the classification to be successful, we need to decompose the variation in the outcome variable into the variation from the control covariates X_{it} and the variation from the latent type k_i. For that end, we assume the following linear model for untreated potential outcome: for $t < 0$,

$$Y_{it}(E_i) - Y_{it-1}(E_i) = \delta_t(k_i) + {X_{it}}^{\intercal}\theta + U_{it}. \tag{8}$$

Note that the interpretation of $\delta_t(k)$ is changed. In the linear model (8), $\delta_t(k)$ is not the conditional mean of first-differenced potential outcome anymore since there exists ${X_{it}}^{\intercal}\theta$. Thus, we call $\delta_t(k)$ the type-specific time fixed-effects. $\delta_t(k)$ explains heterogeneity across

units that is not linearly explained by the observable control covariates X_{it}.

Given the model (8), we can construct a similar objective function from before and solve the K-means minimization problem for classification:

$$(\theta, \delta, \gamma) = \arg\min_{\theta,\delta,\gamma} \frac{1}{nT_0} \sum_{i=1}^{n} \sum_{t=-T_0}^{-1} \left(Y_{it} - Y_{it-1} - \delta_t(k_i) - X_{it}{}^{\intercal}\theta \right)^2.$$

The objective function includes X_{it}. Given an initial type assignment $\gamma^{(0)} = \left(k_1^{(0)}, \cdots, k_N^{(0)}\right)$,

1. **(update θ and δ)** Given the type assignment $\gamma^{(s)}$ from the s-th iteration, construct indicator variables for each time s and the assigned type k: $\mathbf{1}\{t = s, k_i^{(s)} = k\}$ for $s = -T_0 \cdots, -1$ and $k = 1, \cdots, K$. By running OLS regression of $Y_{it} - Y_{it-1}$ on X_{it} and the indicators, update $\hat{\delta}_t^{(s)}(k)$ and $\hat{\theta}^{(s)}$.

2. **(update γ)** Update $k_i^{(s)}$ for each i by letting $k_i^{(s+1)}$ be the solution to the following minimization problem: for $i = 1, \cdots, N$,

$$\min_{k \in \{1,\cdots,K\}} \sum_{t=-T_0}^{-1} \left(Y_{it} - Y_{it-1} - \hat{\delta}_t^{(s)}(k) - X_{it}{}^{\intercal}\hat{\theta}^{(s)} \right)^2.$$

3. Repeat Step 1-2 until Step 2 does not update $\hat{\gamma}$, or some stopping criterion is met. For stopping criterion, one can set a maximum number of iteration or a minimum update in $\hat{\theta}^{(s)}$ and $\hat{\delta}^{(s)}$: set S and ε such that the iteration stops when

$$s \geq S \quad \text{or} \quad \max\left\{ \left\| \hat{\theta}^{(s)} - \hat{\theta}^{(s-1)} \right\|_\infty, \left\| \hat{\delta}^{(s)} - \hat{\delta}^{(s-1)} \right\|_\infty \right\} \leq \varepsilon.$$

In Appendix, we discuss Assumption 7 which extends Assumptions 4-5. Under Assumption 7, we have the following classification result.

Theorem 2. *Let Assumptions 3 and 7 hold. Then, up to some permutation on* $\{1, \cdots, K\}$,

$$\Pr\left\{ \sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0 \right\} = o\left(nT_0{}^{-\nu}\right) + o(1) \quad \forall \nu > 0$$

as $n \to \infty$.

Remark 3. When X_{it} is time-invariant, i.e. $X_{it} = X_i$, the linear model (8) and Assumption 7 can be understood as a parametric special case of the conditional parallel trend assumption: for $t < 0$,

$$\mathbf{E}\left[Y_{it}(E_i) - Y_{it-1}(E_i) | k_i, X_i\right] = \delta_t(k_i) + X_i{}^{\intercal}\theta.$$

Remark 4. Instead of assuming a linear structure on the first difference as in (8), we can consider a linear model on the level of the outcome:

$$Y_{it}(E_i) = \alpha_i + \sum_{s=-T_0}^{t} \delta_s(k_i) + X_{it}{}^{\intercal}\theta + U_{it}.$$

The assumptions for the linear model in level are discussed in the Appendix.

Theorem 2 finds the same rate on the misclassification probability as Theorem 1. The key part of the proof utilizes the linear separability of k_i and X_{it}. The proof firstly shows that θ is consistently estimated. Then, $Y_{it} - Y_{it-1} - X_{it}{}^{\intercal}\hat{\theta}$ is sufficiently close to $Y_{it} - Y_{it-1} - X_{it}{}^{\intercal}\theta$ so that the classification using $\hat{\theta}$ and the one using the true θ are the same.

5.2 Implementing treatment effect estimation with X_{it}

Theorem 2 implies that we can take the estimated types as given and apply the available treatment effect estimation methods when the rate given in Corollary 1 is satisfied. There are largely two ways to incorporate the control covariate X_{it} in the treatment effect estimation. Firstly, we can follow an outcome model approach and impose a parametric model for the post-treatment outcome as we do for the pretreatment outcome in (8). Given the parametric model, we plug in the estimated types as true types and estimate the model. A large variety of parametric models with a finite grouping structure can be used for the outcome model approach. For example, we can assume a linear model with type-specific coefficients for the

treatment effect: for $t \geq 0$,

$$Y_{it}(E_i) = \alpha_i + \delta_t(k_i) + \sum_{r \geq 0} \beta_r(k_i) \mathbf{1}\{t = E_i + r\} + X_{it}^{\intercal}\theta(k_i) + U_{it}. \tag{9}$$

The model (9) is merely one example and can be further relaxed; for example, we can let θ depend also on the treatment timing E_i. The outcome model approach has the merit of developing a parsimonious model for treatment effect; using the outcome model, we can impose some structure over how the latent type k_i and the observable information X_{it}, E_i interact in terms of the treatment effect.

Alternatively, we can use an assignment model approach. In the assignment model approach, we impose a parametric model on the treatment timing rather than on the post-treatment outcomes. Suppose that we are given a time-invariant control covariate X_i and that the conditional parallel trend assumption holds with X_i: for every $t, s \geq -1$,

$$\mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, X_i, E_i\right] = \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, X_i\right].$$

Then, we can apply the results of Callaway and Sant'Anna (2021) by assuming an assignment model and estimating the propensity to be treated given the type k_i and the control covariate X_i. For example, when $E_i \in \{0, \infty\}$, the logistic model can be used:

$$\Pr\left\{E_i = 0 | k_i, X_i\right\} = \frac{\exp\left(X_i^{\intercal}\theta + \delta(k_i)\right)}{1 + \exp\left(X_i^{\intercal}\theta + \delta(k_i)\right)}.$$

The benefit of the assignment model approach is that we allow for flexible interaction between the observable control covariates X_i and the latent type k_i, in terms of the treatment effect heterogeneity. The assignment model approach is an extension of Corollary 1 since the type-specific DiD estimator defined in Section 2 is what we would get when we assume the propensity score to be a trivial function of X_i: $\Pr\left\{E_i = e | k_i, X_i\right\} = \Pr\left\{E_i = e | k_i\right\}$. More discussions on the extension are given in Sections C.2-C.4 of the Appendix.

6 Simulation

In this section, we present simulation results to discuss the finite-sample performance of the type-specific DiD estimator, compared to some existing estimators in the literature. For that, we constructed a random sample using the following data generating process: for $t = -T_0 - 1, \cdots, 0$,

$$Y_{it} = \alpha_i + \delta(k_i)(t+1) + \beta(k_i) D_i \mathbf{1}\{t = 0\} + U_{it},$$
$$U_{it} = \rho U_{it-1} + V_{it}.$$

$D_i, \alpha_i, U_{i,-T_0-1}, \{V_{it}\}_{t \leq 0}$ are mutually independent given k_i. $D_i \big| k_i \sim \text{Bernoulli}\big(\pi(k_i)\big)$ and

$$(\alpha_i, U_{i,-T_0-1}) \big| k_i \sim \mathcal{N}\left(\begin{pmatrix} \alpha(k_i) \\ 0 \end{pmatrix}, \begin{pmatrix} 17 & 0 \\ 0 & \sigma \end{pmatrix} \right),$$
$$V_{it} \big| k_i \overset{iid}{\sim} \mathcal{N}\Big(0, \sigma^2(1-\rho^2)\Big).$$

There are two treatment cohorts—$\{i : E_i = 0\}$ and $\{i : E_i = \infty\}$—and only one population treatment period $t = 0$. The values of the DGP parameters that pertain the classification step are taken from the empirical moments of the dataset used in the next section: $\sigma = 1.85$ and $\rho = 0.60$ for the error distribution and $\delta(k)$ for the type separation. The propensity to be treated $\pi(k)$, the treatment effect $\beta(k)$ and the pretreatment time trend $\delta(k)$ are set to vary in k; the unconditional parallel trend assumption does not hold and the unconditional DiD estimator is biased. For the number of types, we considered $K = 2, 3$ and $\delta_t(k)$ are assumed strongly separated in both cases.[8]

[8]The specifics of the simulation parameters are as follows. For a two-types DGP where $K = 2$, we set $(\pi(1), \pi(2)) = (1/3, 2/3)$, $(\alpha(1), \alpha(2)) = (37, 39)$, $(\delta(1), \delta(2)) = (1.66, 0)$ and $(\beta(1), \beta(2)) = (4, 1)$. $\Pr\{k_i = 1\} = \Pr\{k_i = 2\} = 1/2$. For a three-types DGP where $K = 3$, we set $(\pi(1), \pi(2), \pi(3)) = (1/3, 1/2, 1/2)$, $(\alpha(1), \alpha(2), \alpha(3)) = (37, 39, 35)$, $(\delta(1), \delta(2), \delta(3)) = (2.74, 1.42, 0)$, $(\beta(1), \beta(2), \beta(3)) = (5, 1, 0)$, and $(\Pr\{k_i = 1\}, \cdots, \Pr\{k_i = 3\}) = (2/5, 2/5, 1/5)$.

In the classification step, two different specifications for the type-specific time trend $\delta_t(k)$ were used. Firstly, we used the most flexible specification where $\delta_t(k)$ is allowed to vary across every t: $\{\delta_t(k)\}_{t\leq 0}$. Secondly, we assumed that the researcher has a priori knowledge on the DGP and imposes a constant slope restriction $\delta_t(k) = \delta_{t'}(k)$ for every t, t', estimating only one parameter for each type: $\delta(k)$. Given the type classifications, we estimated the ATT using the type-specific DiD estimators. For comparison, we considered the unconditional DiD, the synthetic control and the synthetic DiD estimators; see Section B of the Appendix for more discussion.

[Table 1 and Table 2 here]

Table 1 and Table 2 contain the simulated bias and the simulated MSE from 500 random samples. Also, they contain the type classification success rate: the first and the second columns of the bottom panel denote probability of a random sample having less than 5% misclassification and the third and the fourth columns denote probability of perfect classification. For large T_0, both the type-specific DiD estimator and the synthetic DiD estimator perform well since there are many pretreatment outcomes to control for the unit-level heterogeneity. However, in the case of the two-types DGP, the type-specific DiD estimator outperforms the other estimators for small $T_0 = 10$ since it best reflects the finite type structure in dataset. As for the classification, long pretreatments ($T_0 = 20, 30$) always achieves perfect classification, except when no smoothness restriction is imposed to the three-types DGP; even then, perfect classification probability is 0.80 for $T_0 = 20$ and 0.94 for $T_0 = 30$.

In addition to the simulation specifications discussed above, we consider two additional DGPs to discuss misspecification in classification: firstly, we let $K = 5$;[9] secondly, we let the latent type be continuous.[10] For both DGPs, we applied the two-types type-specific DiD

[9]We set $\pi(k) = 1/3 + 1/3 \cdot \mathbf{1}\{k \geq 4\}$, $(\alpha(1), \cdots, \alpha(5)) = (37, 39, 35, 36, 38)$, $\beta(k) = 4 - 3 \cdot \mathbf{1}\{k \geq 4\}$, $(\delta(1), \cdots, \delta(5)) = (2.06, 1.66, 1.26, 0.4, 0)$, and $(\Pr\{k_i = 1\}, \cdots, \Pr\{k_i = 5\}) = (1/10, 1/5, 1/5, 1/4, 1/4)$. The DGP is 'fair' to the type-specific DiD estimator with $K = 2$, in the sense that the DGP can be reduced to a two-types DGP, when limited to $\pi(k)$ and $\beta(k)$. If the classification step with $K = 2$ successfully separate Type 1-3 and Type 4-5, the type-specific DiD estimator will be unbiased

[10]We set $\pi(k) = 1/3 \cdot \mathbf{1}\{k \leq 0.5\} + 2/3 \cdot \mathbf{1}\{k > 0.5\}$, $\alpha(k) = 37 + 2k$, $\beta(k) = 4 \cdot \mathbf{1}\{k \leq$

estimator. The first DGP is a case of a misspecified number of types and the second DGP is a case of weakly separated types.

[Table 3 and Table 4 here]

Table 3 and Table 4 contain the simulation results. In the case of misspecified number of types, the overall performance of the type-specific DiD estimator is on par with that of the synthetic DiD estimator, which is a more flexible method. In the case of the continuous latent type, the synthetic DiD estimator outperforms the type-specific DiD estimator for large pretreatment cases ($T_0 = 20, 30$). This is because the two-types structure cannot successfully capture the continuous pattern of heterogeneity from the pretreatment outcomes. The performance of the type-specific DiD estimator may improve as we use more types in the classification step when n and T_0 are large: see Bonhomme et al. (2022) for how we can use a grouping structure with growing number of groups to approximate a continuous pattern of heterogeneity. Larger K approximates the latent heterogeneity better but requires more information, i.e. large n and T_0.

7 Application

For an empirical illustration of the type-specific DiD estimator, we revisit Lutz (2011). Since the Supreme Court ruling on Brown v. Board of Education of Topeka in 1954 that found state laws in the United States enabling racial segregation in public schools unconstitutional, various efforts have been made to desegregate public schools, including court-ordered desegregation plans. After several decades, another important Supreme Court case was made in 1991; the ruling on Board of Education of Oklahoma City v. Dowell in 1991 stated that school districts could terminate the court-ordered plans once it successfully removed the effects of the segregation. Since the second Supreme Court ruling, school districts started to

$0.5\} + \mathbf{1}\{k > 0.5\}$, $\delta(k) = 1.66 \cdot (1 - k)$. $k_i \sim \text{unif}[0, 1]$. Again, the DGP is 'fair' to the type-specific DiD estimator with $K = 2$, in the sense that the type-specific DiD estimator is unbiased when the classification step with $K = 2$ successfully separate $\{i : k_i \leq 0.5\}$ and $\{i : k_i > 0.5\}$.

file for dismissal of court-ordered desegregation plans, mostly in southern states.

Lutz (2011) used the variation in timing of the district court rulings on the desegregation plan to estimate the effect on racial composition and education outcomes in public schools. The paper uses annual data on mid- and large-sized school districts from 1987 to 2006, obtained from the Common Core of Data (CCD), which contains data on school districts from 1987 to 2006, and the School District Databook (SDDB) of the US census, which contains data on school districts in 1990 and in 2000. To document if a school district was under a court-ordered desegregation plan at the time of the Supreme Court ruling in 1991 and when and if the school district got the desegregation plan dismissed at the district courts, Lutz (2011) collected data from various published and unpublished sources, including a survey by Rosell and Armor (1996) and the Harvard Civil Rights Project.

Though Lutz (2011) looks at several outcome variable, we focus on one outcome variable, the dissimilarity index: the dissimilarity index for school district i is

$$Y_i = \frac{1}{2} \sum_{j \in J_i} \left| \frac{b_j}{B_i} - \frac{w_j}{W_i} \right| \times 100$$

where b_j is the number of black students in school j, w_j is the number of white students in school j and J_i is the set of school in school district i. Also, $B_i = \sum_{j \in J_i} b_j$ and $W_i = \sum_{j \in J_i} w_j$. The dissimilarity index ranges from 0 to 100, with 100 being perfectly segregated schools and 0 being perfectly representative schools.[11]

We followed the data cleaning process in the paper and chose the timespan of 1988-2007 to form a balanced panel of school districts that were under a court-ordered desegregation plan in 1988-1999, which gave us 50 school districts.[12] We use the following linear model for

[11] In Lutz (2011), the dissimilarity index ranges from zero to one but we rescaled the index for visibility.

[12] Out of 50 school districts, 3 school districts were treated in 2000, 5 in 2001, 5 in 2002, 2 in 2003, 1 in 2004, 2 in 2005, 1 in 2006 and 2 in 2007, totaling 21 ever-treated school districts. 29 school districts were not treated until 2007.

the pretreatment outcomes: for $t = 1989, \cdots, 1999$,

$$Y_{it}(E_i) - Y_{it-1}(E_i) = \delta_t(k_i) + X_{it}{}^{\intercal}\theta + U_{it}.$$

The effective number of pretreatment outcomes is 11. The control covariates X_{it} contain a central city indicator variable, percentage of students who are white, percentage of students who are hispanic, percentage of students with free/reduced-priced lunch and number of students. For comparison, we present the main empirical specification of Lutz (2011):

$$Y_{it} - Y_{it-1} = \delta_{jt} + \sum_{r=-4}^{10} \beta_r \left(\mathbf{1}\{t = E_i + r\} - \mathbf{1}\{t = E_i + r + 1\}\right) + X_i{}^{\intercal}\theta_t + U_{it}$$

Though two specifications look alike, there are some differences. Firstly, though Lutz (2011) and we use the same control covariates, Lutz (2011) only used their values from the first year, with time-varying coefficient θ_t: $X_i = X_{i,-T_0-1}$.[13] On the other hand, we use time-varying control covariates X_{it}, with time-invariant coefficient θ. Secondly, Lutz (2011) uses time fixed-effects δ_{jt} based on census region, which assigns every school district into one of the four regions. In the terminology of the model used in this paper, Lutz (2011) took the census region as the true type assignment whereas we used the data to estimate the type assignment. Lastly, the regression specification in Lutz (2011) has a dynamic treatment effect β_r whereas we only impose linearity on pretreatment outcomes and therefore do not have any treatment effect term. Since $n = 50$ is relatively small, we imposed an additional smoothness restriction on the type-specific time fixed-effects: $\delta_t(k) = \delta(k)$.[14] Then, we applied the K-means clustering classifier with $K = 2$.[15] The first-step classification assigns

[13]Also, Lutz (2011) used three additional variables: squared number of students, cubed number of students and squared percentage of students with free/reduced-price lunch.

[14]The constant slope restriction was chosen out of five specifications—constant slope, constant slope with a break, constant slope with two breaks, linear, linear with one break—, based on cross-validated mean-squared forecasting error. For more discussion, see the Supplementary Appendix.

[15]As robustness check, we also considered $K = 3$ and $K = 4$. The Bayesian information criterion selected $K = 3$ and the qualitative result remains the same for both $K = 2$ and

8 ever-treated units and 14 never-treated units to Type 1 and 13 ever-treated units and 15 never-treated units to Type 2; Type 2 school districts are slightly more likely to be treated.

[Table 5 here]

Given the first-step classification result, we conducted a within-type balancedness test; Table 5 contains within-type balancedenss test results using control covariates from 1988. Within the two types, the control covariates are well-balanced across treatment status: treated v. never-treated. Thus, we did not use X_{it} in the estimation step and use the unconditional type-specific DiD estimator from (7); the target parameter captures the type-specific dynamic ATT assigning equal weights to each ever-treated unit within the type.

[Figure 1 here]

Figure 1 contains the type-specific DiD estimates for Type 1 and Type 2 school districts. From Figure 1, we see that the treatment effect is bigger for Type 1 and smaller for Type 2; the termination of court-ordered desegregation plans exacerbated racial segregation more severely for Type 1. The type-specific estimation provide a sensible aggregate estimate; averaging the type-specific DiD estimates across types gives us $(8\hat{\beta}_r(1) + 13\hat{\beta}_r(2))/21 = 4.37$ at $r = 4$ whereas the pooled regression estimate from Lutz (2011) is around 4-5 at $r = 4$, depending on specifications, and the conditional DiD estimate from Callaway and Sant'Anna (2021) is 4.85 at $r = 4$. For reference on the magnitude, the mean of the dissimilarity index was 34 and its standard deviation was around 16 in 1988. Also, Figure 1 contains 95% confidence intervals; the 'no pretreatment' test did not reject at 0.05 significance level.

[Table 6 here]

So, estimates on treatment effect suggest that Type 1 and Type 2 are different; the Type 1 school districts are more responsive to the treatment. How are these types different in other regards? Firstly, Table 6 shows us some descriptive statistics on the outcome variable and other control covariates for each type, using year 1988 data. The null hypothesis that the entire vector of mean differences between Type 1 and Type 2 is zero is rejected

$K = 3$. For more discussion, see Supplementary Appendix.

with a t-test at size 0.05; the Type 1 school districts are different from the Type 2 school districts in terms of their observable characteristics. For instance, Type 1 school districts have higher proportion of white students and lower proportion of hispanic students. Secondly, in terms of the unobserved heterogeneity captured by the latent type variable, Type 1 has seen a steeper increase in the dissimilarity index while the slope was smaller for Type 2: $\left(\hat{\delta}(1), \hat{\delta}(2)\right) = (3.59, 1.93)$. This implies that the dismissal of desegregation plans had a bigger impact on Type 1, where the dissimilarity index was already rising faster. This observation presents future research questions: e.g., why do the school districts that were getting more segregated also get affected more from the dismissal of the desegregation plan?

8 Conclusion

In this paper, we introduce a type-specific parallel trend assumption in a panel data model with a latent type. By assuming the latent type variable has a finite support and is well-separated in the long pretreatment time series, the K-means classifier estimates the true types consistently. Also, based on the estimated types, we estimate the type-specific treatment effect. The type-specific DiD estimator is useful when we suspect heterogeneity in time trends across units and want to explore the associated treatment effect heterogeneity. By applying the estimation method to an empirical application, we find some interesting empirical results where the estimates on the type-specific treatment effects and those on the type-specific time trend tell a story: the effect of terminating court-mandated desegregation plans were bigger for school districts where the dissimilarity index was growing.

References

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

Abadie, Alberto, Alexis Diamond and Jens Hainmueller (2010), 'Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program', *Journal of the American statistical Association* **105**(490), 493–505.

Abadie, Alberto, Alexis Diamond and Jens Hainmueller (2015), 'Comparative politics and the synthetic control method', *American Journal of Political Science* **59**(2), 495–510.

Abadie, Alberto, Matthew M Chingos and Martin R West (2018), 'Endogenous stratification in randomized experiments', *Review of Economics and Statistics* **100**(4), 567–580.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens and Stefan Wager (2021), 'Synthetic difference-in-differences', *American Economic Review* **111**(12), 4088–4118.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi (2021), 'Matrix completion methods for causal panel data models', *Journal of the American Statistical Association* **116**(536), 1716–1730.

Bai, Jushan and Serena Ng (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

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

Bonhomme, Stéphane and Elena Manresa (2015), 'Grouped patterns of heterogeneity in panel data', *Econometrica* **83**(3), 1147–1184.

Bonhomme, Stéphane, Thibaut Lamadon and Elena Manresa (2022), 'Discretizing unobserved heterogeneity', *Econometrica* **90**(2), 625–643.

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

Callaway, Brantly and Pedro HC Sant'Anna (2021), 'Difference-in-differences with multiple time periods', *Journal of Econometrics* **225**(2), 200–230.

Callaway, Brantly and Sonia Karami (2023), 'Treatment effects in interactive fixed effects models with a small number of time periods', *Journal of Econometrics* **233**(1), 184–208.

De Chaisemartin, Clément and Xavier d'Haultfoeuille (2020), 'Two-way fixed effects estimators with heterogeneous treatment effects', *American Economic Review* **110**(9), 2964–96.

Ding, Peng and Fan Li (2019), 'A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment', *Political Analysis* **27**(4), 605–615.

Freyaldenhoven, Simon, Christian Hansen and Jesse M Shapiro (2019), 'Pre-event trends in the panel event-study design', *American Economic Review* **109**(9), 3307–38.

Ghanem, Dalia, Pedro HC Sant'Anna and Kaspar Wüthrich (2022), 'Selection and parallel trends', *arXiv preprint arXiv:2203.09001* .

Goldsmith-Pinkham, Paul, Peter Hull and Michal Kolesár (2024), 'Contamination bias in linear regressions', *American Economic Review* **114**(12), 4015–4051.

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

Hsiao, Cheng, H Steve Ching and Shui Ki Wan (2012), 'A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china', *Journal of Applied Econometrics* **27**(5), 705–740.

Janys, Lena and Bettina Siflinger (2024), 'Mental health and abortions among young women: Time-varying unobserved heterogeneity, health behaviors, and risky decisions', *Journal of Econometrics* **238**(1), 105580.

Lutz, Byron (2011), 'The end of court-ordered desegregation', *American Economic Journal: Economic Policy* **3**(2), 130–68.

Roth, Jonathan and Pedro HC Sant'Anna (2023*a*), 'Efficient estimation for staggered rollout designs', *Journal of Political Economy Microeconomics* .

Roth, Jonathan and Pedro HC Sant'Anna (2023*b*), 'When is parallel trends sensitive to functional form?', *Econometrica* **91**(2), 737–747.

Sant'Anna, Pedro HC and Jun Zhao (2020), 'Doubly robust difference-in-differences estimators', *Journal of Econometrics* **219**(1), 101–122.

Sun, Liyang and Sarah Abraham (2021), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *Journal of Econometrics* **225**(2), 175–199.

Xu, Yiqing (2017), 'Generalized synthetic control method: Causal inference with interactive fixed effects models', *Political Analysis* **25**(1), 57–76.

APPENDIX

A Parallel trend v. design-based approach

The type-specific parallel trend assumption used in this paper do not impose restrictions on the assignment process for the treatment timing and rather directly impose restrictions on the outcome model. Though the parallel trend type assumptions have their own advantages of being concise and straightforward, the parallel trend assumption hinges on an arbitrary choice of what to compare: the temporal differences in level. For example, when a researcher is interested in estimating the treatment effect as a percentage change, they may want to use a parallel trend assumption with logged outcome variables:

$$\mathbf{E}\left[\log Y_{it}(\infty) - \log Y_{is}(\infty) | k_i, E_i\right] = \mathbf{E}\left[\log Y_{it}(\infty) - \log Y_{is}(\infty) | k_i\right].$$

On the other hand, a design-based approach such as an unconfoundedness assumption would be free of this commitment to a functional form. When

$$\{Y_{it}(e)\}_{t,e} \perp\!\!\!\perp E_i | k_i, \tag{10}$$

a parallel trend assumption with any functional form would hold,[16] at the cost of assuming distributional independence which is stronger than the parallel trend type assumption.

There are some benefits to the design-based approach in addition to being robust to the choice of functional form. Under the parallel trend type assumption, there was no clear choice in which temporal differences to use. However, when we assume random treatment timing or the unconfoundedness assumption, we have some theoretical guidance on this choice. Roth and Sant'Anna (2023*a*) assume random treatment timing and find an efficient estimator among DiD type estimators that uses different weights across different temporal differences. Given the same classification result from Theorem 1, the unconfoundedness assumption (10) can be used to find an efficient type-specific DiD estimator following their procedure.

When a researcher uses a design-based approach, the question of great interest is how much more restrictions are imposed when assuming the unconfoundedness, compared to the conditional parallel trend. Roth and Sant'Anna (2023*b*); Ghanem et al. (2022) provide insights to this question. Roth and Sant'Anna (2023*b*) show that an equivalent condition for the parallel trend assumption to hold for any monotone transformation of the outcome variable is that the population is divided into two subgroups where the treatment is random for the first subgroup and the untreated potential outcome has time-invariant distribution for the second subgroup. In this sense, the unconfoundedness assumption (10) is indeed strictly stronger than the type-specific parallel trend assumption holding for every monotone

[16]This statement is only true for the unconfoundedness assumption as given in (10). Ding and Li (2019) discuss a simple two period case ($t = 1, 2$) where no one is treated at $t = 1$ and show that the parallel trend assumption and the unconfoundedness assumption do not nest each other when the unconfoundedness assumption is applied sequentially: $Y_{i1}(\infty) = Y_{i1}(2)$ and $Y_{i2}(\infty) \perp\!\!\!\perp D_{i2} | \big(k_i, Y_{i1}\big)$

transformation of the outcome. Ghanem et al. (2022) provide insight in understanding the cost of assuming an additional parallel trend type assumption incrementally. Given a functional form, Ghanem et al. (2022) provide necessary conditions and sufficient conditions for that specific parallel trend assumption in terms of restrictions on the assignment model.

B Weighted outcomes as counterfactual

In this section, we discuss how the type-specific DiD estimator compares to other estimators that assign weights over control units to construct a counterfactual outcome. In specific, we discuss the conventional DiD estimator and the synthetic control estimator. Find that the classification result from (6) satisfy that

$$\hat{\delta}_t(k) = \frac{\sum_{i=1}^{n} (Y_{it} - Y_{it-1})\, \mathbf{1}\{\hat{k}_i = k\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k\}}.$$

$\hat{\delta}_t(k)$ puts equal weights over $Y_{it} - Y_{it-1}$ for units with the same estimated type k. Suppose that there is only one post-treatment period and only one ever-treated unit: $T_1 = 1$, $E_i = \infty$ for every $i \leq N_0 = n - 1$ and $E_n = 0$. Consider a treatment effect estimator $\hat{\beta}$ which can be written as a weighted sum of Y_{it}: $\hat{\beta} = \sum_{i,t} w_{it} Y_{it}$. For a simple DiD estimator, the weight is

$$w_{it}^{did} = \mathbf{1}\{i = n, t = 0\} - \mathbf{1}\{i = n, t = -1\} - \frac{\mathbf{1}\{i \leq N_0, t = 0\}}{N_0} + \frac{\mathbf{1}\{i \leq N_0, t = -1\}}{N_0}.$$

In the case of the synthetic control (see Abadie et al. (2010, 2015),

$$w_{it}^{sc} = \mathbf{1}\{i = n, i = 0\} - \sum_{j=1}^{N_0} w_j^* \mathbf{1}\{i = j, t = 0\}$$

where $\{w_j^*\}_{j \leq N_0}$ are solution to the following minimization:

$$\min_{w} \sum_{t=-T_0-1}^{-1} \left(Y_{nt} - \sum_{i=1}^{N_0} w_i Y_{it} \right)^2.$$

subject to $\sum_{i=1}^{N_0} w_i = 1$ and $w_i \geq 0$. In the case of the type-specific DiD,

$$w_{it}^{tdid} = \mathbf{1}\{i = n, t = 0\} - \mathbf{1}\{i = n, t = -1\} - \sum_{j=1}^{N_0} w_j^{**} \Big(\mathbf{1}\{i = j, t = 0\} - \mathbf{1}\{i = j, t = -1\} \Big)$$

where $\{w_j^{**}\}_{j \leq N_0}$ are (a function of) the solution to the following minimization:

$$\min_w \sum_{i=1}^{n} \sum_{t=-T_0}^{-1} \left((Y_{it} - Y_{it-1}) - \sum_j w_{ij} \left(Y_{jt} - Y_{jt-1} \right) \right)^2$$

subject to $w_{ij} = \mathbf{1}\{k_i = k_j\} \big/ \sum_l \mathbf{1}\{k_l = k_i\}$ for some $\{k_i\}_{i=1}^n \in \{1, \cdots, K\}^n$. Based on the optimized w_{ij}, we get $w_j^{**} = w_{nj} \big/ \sum_{j' \neq n} w_{nj'}$.

Compared to the DiD estimator, the type-specific DiD estimator admits more flexible cross-sectional weights by using a subset of the never-treated (or not-yet-treated) units. Compared to the synthetic control estimator, the type-specific DiD estimator is less flexible in terms of the cross-sectional weights since it is dichotomous cross-sectionally; a never-treated unit gets a positive weight if and only if it shares the same type with the ever-treated unit and the weight is uniform. However, the synthetic control estimator assigns nonzero weights only to contemporaneous outcomes while the type-specific DiD estimator takes temporal difference. Lastly, the synthetic DiD estimator from Arkhangelsky et al. (2021) also uses a weighted sum type estimator and assigns flexible weights both cross-sectionally and intertemporally, therefore nesting all of the methods discussed above.

C Additional assumptions and asymptotic results

C.1 Overlap condition

Assumption 6. *Fix some* $t \in \{0, \cdots, T_1 - 1\}$, $k \in \{1, \cdots, K\}$, $e \in \{0, \cdots, T_1 - 1\}$ *and* $r \in \{0, \cdots, T_1 - 1\}$.

a. $\Pr\{E_i = e | k_i = k\} \cdot \Pr\{E_i > t | k_i = k\} > 0$.

b. $\Pr\{E_i = e'|k_i = k\} \cdot \Pr\{E_i > e' + r|k_i = k\} > 0$ *for some* $e' \leq T_1 - 1 - r$.

Assumption 6-a assumes nonzero treatment probability at time e and at some time later than t, for a unit with type k. In Lemma 1 and Corollary 1, Assumption 6-a is the overlap condition for $ATT_t(k, e)$.

Assumption 6-b assumes there exists two treatment timings $e' \in \{0, \cdots, T_1 - 1 - r\}$ and $s \in \{e' + r + 1, \cdots, T_1 - 1, \infty\}$ with nonzero treatment probability for a unit with type k; i.e., the treatment timings of the two treatment cohorts $\{i : E_i = e'\}$ and $\{i : E_i = s\}$ differ by at least r and both of the treatment cohorts have nonzero measure. In Lemma 1 and Corollary 1, Assumption 6-b is the overlap condition for $\beta_r(k)$.

For Assumption 6-b, it helps to think of some $\bar{r}_k \geq 0$ such that

$$\bar{r}_k = \underbrace{\min\{T_1, \max\{e : \Pr\{E_i = e|k_i = k\} > 0\}\} - 1}_{\text{latest time period with untreated outcomes}} - \underbrace{\min\{e : \Pr\{E_i = e|k_i = k\} > 0\}}_{\text{earliest time period with treated outcomes}}.$$

$\bar{r}_k$ is the upper bound on how far the dynamic treatment effect can be identified; for any $r \leq \bar{r}_k$, the dynamic ATT $\beta_r(k)$ is identified.

C.2 Type classification given control covariates X_{it}

To have the classification result under the linear outcome model with control covariates (8), we assume the following assumption.

Assumption 7. *With some* $M, \tilde{M} > 0$,

a. *(iid across units)* $\left(\{X_{it}, U_{it}\}_{t<0}, E_i, k_i\right) \overset{iid}{\sim} F$.

b. *(compact parameter space) For every* t *and* k, $|\delta_t(k)| \leq M$. $\|\theta\|_2 \leq M$.

c. *(well-separated types) Whenever* $k \neq k'$,

$$\frac{1}{T_0} \sum_{t=-T_0}^{-1} \left(\delta_t(k) - \delta_t(k')\right)^2 \to c(k, k') > 0$$

as $n \to \infty$.

d. *(strict exogeneity and finite moments)*

For every $t < 0$, $\mathbf{E}\left[U_{it} | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\right] = 0$ *and* $\mathbf{E}\left[{U_{it}}^4 | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\right] \leq M$.

For every $t, s < 0$, $\mathbf{E}\left[\left|{X_{it}}^\intercal X_{is}\right|\right] \leq M$. *For any* $\nu > 0$,

$$\Pr\left\{\frac{1}{T_0}\sum_{t=-T_0}^{-1} \|X_{it}\|_2 \geq \tilde{M}\right\} = o\left({T_0}^{-\nu}\right)$$

as $n \to \infty$.

e. *(long pretreatment)* $T_0 \to \infty$ *as* $n \to \infty$.

f. *(no measure zero types) For all* $k \in \{1, \cdots, K\}$, $\Pr\{k_i = k\} > 0$

g. *(weakly dependent, thin-tailed errors) With some positive constant* d_1 *and* a, $\{U_{it}\}_{t=-T_0}^{-1}$ *is strongly mixing with mixing coefficient* $\alpha[t]$ *such that* $\alpha[t] \leq \exp(-at^{d_1})$. *Also, with some positive constant* d_2 *and* b, U_{it} *satisfies the following tail probability: for any* $u > 0$,

$$\Pr\{|U_{it}| \geq u\} \leq \exp\left(1 - (u/b)^{d_2}\right)$$

uniformly over $t < 0$.

h. *(no multicollinearity) Given an arbitrary type assignment* $\tilde{\gamma} = \left(\tilde{k}_1, \cdots, \tilde{k}_n\right)$, *let* $\bar{X}_{k \wedge \tilde{k}, t}$ *denote the mean of* X_{it} *among units such that* $k_i = k$ *and* $\tilde{k}_i = \tilde{k}$. *Let* $\rho_n(\tilde{\gamma})$ *denote the minimum eigenvalue of the following matrix:*

$$\frac{1}{nT_0}\sum_{i=1}^{n}\sum_{t=-T_0}^{-1}\left(X_{it} - \bar{X}_{k_i \wedge \tilde{k}_i, t}\right)\left(X_{it} - \bar{X}_{k_i \wedge \tilde{k}_i, t}\right)^\intercal.$$

Then, $\min_{\tilde{\gamma} \in \Gamma} \rho_n(\tilde{\gamma}) \xrightarrow{p} \rho$ *as* $n \to \infty$.

Assumption 7-c replaces Assumption 4 in the context of (8). Assumption 7-c assumes that the residual unobserved heterogeneity across units after regressing out X_{it} has finite

types and is well-separated in the l_2 norm. Assumption 7-d replaces Assumption 5-b and additionally assumes that for large enough $\tilde{M}$, the probability of $\frac{1}{T_0}\sum_{t=-T_0}^{-1}\|X_{it}\|_2$ being larger than $\tilde{M}$ goes to zero exponentially. Moreover, Assumption 7-d combined with (8) replaces the parallel trend assumption given in Assumptions 1-2, by imposing

$$\mathbf{E}\left[Y_{it} - Y_{it-1} - X_{it}{}^{\intercal}\theta | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\right] = \delta_t(k_i).$$

Assumption 7-g replaces Assumption 5-e. Assumption 7-h assumes sufficient variation in X_{it} within each type. When an outcome model is assumed for the pretreatment outcome in level as in Remark 4, the same conditions from Assumption 7-d,g and a modification of Assumption 7-h by replacing X_{it} with $X_{it} - X_{it-1}$ give us Theorem 2.

C.3 Outcome model approach given control covariates X_{it}

Once $\{k_i\}_{i=1}^n$ is estimated, we may use the estimated types to estimate various models on post-treatment periods with type-specific parameters. Directly modelling the outcome model with the observable information X_{it} as in (9) can be helpful when we are interested in treatment effect heterogeneity and we would like to impose some restrictions on the heterogeneity because of the structure of X_{it}. For example, when X_{it} is continuous and multidimensional, a linearity assumption on the treatment effect $\beta_r(X_{it}, k_i) = \beta_r(k_i)^{\intercal}X_{it}$ can be helpful in summarizing how X_{it} interacts with the type k_i in terms of the treatment effect.

Consider a generalized outcome model for post-treatment outcomes: for $t \geq 0$,

$$Y_{it}(E_i) - Y_{it-1}(E_i) = m(X_{it}, E_i, k_i; \xi) + U_{it}.^{17}$$

An important restriction we need to assume for the outcome model m is that m treats the latent type k_i as a categorical variable. In the example (9), $\xi = \left\{\{\delta_t(k)\}_{t\geq 0}, \{\beta_r(k)\}_{r\geq 0}, \theta(k)\right\}_k$. Note that the dimension of ξ is fixed; the dimension is $(2T_1 + p)K$ and T_1 and K are fixed.

Alternatively, we can consider a richer control specification: for example,

$$Y_{it}(E_i) = \alpha_i + \delta_t(k_i) + \sum_{r \geq 0} \beta_r(k_i)^\intercal X_{it} \mathbf{1}\{t = E_i + r\} + X_{it}{}^\intercal \theta(k_i, E_i) + U_{it}.$$

The treatment effect $\beta_r(k_i)^\intercal X_{it}$ depends on the type and the control covariates and the baseline trend $\delta_t(k) + X_{it}{}^\intercal \theta(k_i, E_i)$ on the type, the treatment cohort and the control covariates.

Let $\tilde{\xi}$ be the infeasible least-square estimator for ξ and $\hat{\xi}$ be the plug-in estimator:

$$\tilde{\xi} = \arg\min_{\xi \in \Xi} \frac{1}{nT_1} \sum_{i=1}^{n} \sum_{t=0}^{T_1 - 1} \Big(Y_{it} - Y_{it-1} - m(X_{it}, E_i, k_i; \xi) \Big)^2,$$
$$\hat{\xi} = \arg\min_{\xi \in \Xi} \frac{1}{nT_1} \sum_{i=1}^{n} \sum_{t=0}^{T_1 - 1} \Big(Y_{it} - Y_{it-1} - m(X_{it}, E_i, \hat{k}_i; \xi) \Big)^2.$$

Assumption 8. *The latent type k_i enters the outcome model m as a categorical variable. Ξ, the parameter space for ξ, is bounded: with some $M > 0$,*

$$\sup_{\xi \in \Xi} \|\xi\|_2 \leq M.$$

The true value ξ lies in the interior of Ξ. Also, the infeasible estimator $\tilde{\xi}$ satisfies that

$$\sqrt{n} \left(\tilde{\xi} - \xi \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$$

with some $\Sigma > 0$ as $n \to \infty$.

Corollary 2. *Let Assumptions 3 and 7-8 hold. Suppose that there exists some $\nu^* > 0$ such that $n/T_0{}^{\nu^*} \to 0$ as $n \to \infty$. Then, up to some permutation on $\{1, \cdots, K\}$,*

$$\sqrt{n} \left(\hat{\xi} - \xi \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$$

[17] Though the first-differenced outcome variables are used in the post-treatment outcome model for internal consistency with (8), we can also consider models with outcome variable in level. In that case, one could use unit fixed-effects or type-by-treatment-cohort fixed-effects to address unit-level heterogeneity in level.

with $\Sigma > 0$ from Assumption 8 as $n \to \infty$.

Proof. The result holds directly from Theorem 2. Find that

$$\sqrt{n}\left\|\tilde{\xi} - \hat{\xi}\right\|_2 \leq 2\sqrt{n}M\mathbf{1}\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0\right\} = o_p(1)$$

since for any $\varepsilon > 0$,

$$\Pr\left\{2\sqrt{n}M\mathbf{1}\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0\right\} > \varepsilon\right\} \leq \Pr\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0\right\} = o(1)$$

from $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. □

Note that Assumption 8 does not discuss whether the true parameter ξ has sensible causal interpretation as we did for $ATT_t(k, e)$ in Section 3. In the example of (9), it is well known that the linear coefficients $\beta_r(k)$ may suffer from the bias that comes from the dependence structure in $\mathbf{1}\{t = E_i + r\}$, given treatment effect heterogeneity.[18] Thus, we consider an alternative approach in the next subsection.

C.4 Assignment model approach given control covariates X_{it}

Directly modelling the outcome model may be too restrictive in some empirical contexts where the treatment effect depends on the control covariates X_{it} and the type k_i in a more flexible way. A similar concern is addressed in Callaway and Sant'Anna (2021) where the authors consider a conditional parallel trend assumption where the conditioning set is X_i. In Callaway and Sant'Anna (2021), authors impose restriction on the assignment model while not imposing any restriction on the treatment effect heterogeneity in terms of X_i.

[18] It has been discussed that treatment effect estimators from TWFE specification are biased under the parallel trend type assumption (see De Chaisemartin and d'Haultfoeuille (2020); Goodman-Bacon (2021); Borusyak et al. (2024); Sun and Abraham (2021) among others) and potentially distort hypothesis testing (see Baker et al. (2022)). Also, Goldsmith-Pinkham et al. (2024) show that even when the treatment timing is random, treatment effect estimators still suffer from contamination bias when dynamic treatment effect specification is used.

With some finite-dimensional parameter ξ, we use a parametric function π_e to model the conditional distribution of the treatment timing E_i given X_i and k_i:[19]

$$\Pr\{E_i = e | k_i, X_i\} = \pi_e(X_i, k_i, \xi).$$

Again, the type variable k_i is treated as a categorical variable in the assignment model π_e. An example of such a model is an ordered logit model with type-specific coefficients:

$$\Pr\{E_i \leq e | k_i = k, X_i = x\} = \frac{\sum_{e'=0}^{e} \exp(x^\intercal \theta(k) + \delta_{e'}(k))}{\sum_{e'=0}^{T_1-1} \exp(x^\intercal \theta(k) + \delta_{e'}(k)) + \exp(x^\intercal \theta + \delta_\infty(k))}.$$

In this example, $\xi = (\theta, \delta_e(k))_{k,e}$ and its dimension is fixed: $T_1 K + p$.

Let $\tilde{\xi}$ be the infeasible maximum likelihood estimator for ξ and $\hat{\xi}$ be the plug-in estimator:

$$\tilde{\xi} = \arg\min_{\xi \in \Xi} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{e=0}^{T_1-1} \mathbf{1}\{E_i = e\} \log \pi_e(X_i, k_i, \xi) + \mathbf{1}\{E_i = \infty\} \log \pi_\infty(X_i, k_i, \xi) \right),$$
$$\hat{\xi} = \arg\min_{\xi \in \Xi} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{e=0}^{T_1-1} \mathbf{1}\{E_i = e\} \log \pi_e(X_i, \hat{k}_i, \xi) + \mathbf{1}\{E_i = \infty\} \log \pi_\infty(X_i, \hat{k}_i, \xi) \right).$$

Using $\hat{\xi}$, the type-specific DiD estimators can be constructed as follows:

$$\hat{\beta}_r(k) = \sum_{e \leq T_1-1-r} \frac{\hat{\mu}(k, e)}{\sum_{e' \leq T_1-1-r} \hat{\mu}(k, e')} \cdot \widehat{ATT}_{e+r}(k, e)$$

[19]The conditional distribution of E_i given (k_i, X_i) captures how treatment timing depends on the type and the control covariate. However, it does not contain any information on the dependence between k_i and X_i. For that end, we could consider the conditional distribution of k_i given X_i. Given a new draw of X_i, we cannot know k_i; however, we can look at the (estimated) distribution of $k_i | X_i$. Moreover, based on the distribution, a prediction on treatment effect can also be made. A close parallel with the IV literature exists here; we cannot know if a newly drawn unit with covariate X_i is a complier or not, but we can identify the conditional probability of them being a complier given X_i.

where

$$\widehat{ATT}_t(k,e) = \frac{\sum_{i=1}^n (Y_{it} - Y_{i,e-1})\, \mathbf{1}\{\hat{k}_i = k, E_i = e\}}{\sum_{i=1}^n \mathbf{1}\{\hat{k}_i = k, E_i = e\}} - \frac{\sum_{i=1}^n (Y_{it} - Y_{i,e-1})\, \mathbf{1}\{\hat{k}_i = k, E_i > t\}\pi_e(X_i,k,\hat{\xi})/\pi_{t+}(X_i,k,\hat{\xi})}{\sum_{i=1}^n \mathbf{1}\{\hat{k}_i = k, E_i > t\}\pi_e(X_i,k,\hat{\xi})/\pi_{t+}(X_i,k,\hat{\xi})}$$

$$\hat{\mu}(k,e) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\hat{k}_i = k, E_i = e\} \quad \text{and} \quad \pi_{e+}(x,k,\xi) = \sum_{e' \geq e+1} \pi_e(x,k,\xi).$$

For the asymptotic results on the type-specific DiD estimator, we assume the following:

Assumption 9. *With some constant $M > 0$,*

a. *(finite moments) For every e and $t \geq -1$, $\mathbf{E}\left[Y_{it}(e)^4 | k_i, X_i\right] \leq M$.*

b. *(type-specific parallel trend) For every $t, s \geq -1$ and e,*

$$\mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, X_i, E_i\right] = \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, X_i\right]$$

$$\mathbf{E}\left[Y_{it}(e) - Y_{it}(\infty) | k_i, X_i, E_i\right] = 0$$

c. *There exists some $\varepsilon^\pi > 0$ such that $\mu(k,e) > 0 \Rightarrow \Pr\left\{\varepsilon^\pi \leq \inf_{w \in \Xi} \pi_e(X_i,k,w)\right\} = 1$.*

d. *Fix some t, e and k such that $\mu(k,e) > 0$ and $t \geq e$ and define a function $g : \Xi \to \mathbb{R}$ such that*

$$g(w; X_i) = \frac{\pi_e(X_i,k,w)}{\pi_{t+}(X_i,k,w)}.$$

There is a small neighborhood B_ξ around ξ with regard to $\|\cdot\|_2$ such that

i. *g is almost surely twice continuously differentiable on B_ξ;*

ii. *$\frac{\partial}{\partial w} g(w)$ and $\frac{\partial^2}{\partial w \partial w^\intercal} g(w)$ are almost surely bounded by M on B_ξ.*

Assumption 9-c and 9-d correspond to Assumption 7 of Callaway and Sant'Anna (2021). Note that Assumption 9-b assumes a conditional parallel trend assumption for treatment

periods, which is more relaxed the outcome model approach. That being said, the assignment model approach still assumes a linear parametric model for pretreatment periods; the relaxation is not on the the pretreatment trends, but on the treatment effect heterogeneity.

With Assumption 9, we have the following corollary of Theorem 2.

Corollary 3. *Let Assumptions 3, 7-9 hold after replacing X_{it} with X_i and m with π_e. Suppose that there exists some $\nu^* > 0$ such that $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. Then, up to some permutation on $\{1, \cdots, K\}$,*

$$\sqrt{n}\left(\hat{\xi} - \xi\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma\right)$$

with $\Sigma > 0$ from Assumption 8 as $n \to \infty$. In addition, suppose that the infeasible estimator $\tilde{\xi}$ admits an asymptotic linear approximation as follows:

$$\sqrt{n}\left(\tilde{\xi} - \xi\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l^{\pi}(X_i, k_i, E_i) + o_p(1)$$

where $\mathbf{E}[l^{\pi}(X_i, k_i, E_i)] = 0$ and $\mathbf{E}\left[l^{\pi}(X_i, k_i, E_i) l^{\pi}(X_i, k_i, E_i)^{\intercal}\right] > 0$. Then, up to some permutation on $\{1, \cdots, K\}$,

$$\sqrt{n}\left(\hat{\beta}_r(k) - \beta_r(k)\right) \xrightarrow{d} \mathcal{N}\left(0, {\sigma_{\beta_r(k)}}^2\right)$$

with some ${\sigma_{\beta_r(k)}}^2 > 0$, as $n \to \infty$, when k and r satisfy the corresponding overlap condition.

D Tables and figures

Table 1: Simulation Results, $K = 2$

Bias					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(50, 10)$	-0.540	-0.107	-0.183	-0.008	-0.017
$(50, 20)$	-0.594	-0.129	-0.110	-0.027	-0.027
$(50, 30)$	-0.571	-0.140	-0.090	-0.035	-0.035
$(100, 10)$	-0.603	-0.218	-0.233	-0.049	-0.043
$(100, 20)$	-0.531	-0.026	-0.051	0.009	0.009
$(100, 30)$	-0.535	0.029	-0.001	0.025	0.025
Constant slope	-	-	-	NO	YES

MSE					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(50, 10)$	0.696	0.620	0.435	0.370	0.367
$(50, 20)$	0.754	0.642	0.371	0.342	0.342
$(50, 30)$	0.753	0.662	0.380	0.363	0.363
$(100, 10)$	0.576	0.425	0.251	0.185	0.184
$(100, 20)$	0.491	0.312	0.170	0.165	0.165
$(100, 30)$	0.521	0.300	0.185	0.187	0.187
Constant slope	-	-	-	NO	YES

Classification success probability				
(n, T_0)	$\leq 5\%$ misclass.		No misclass.	
$(50, 10)$	0.984	1.000	0.748	0.904
$(50, 20)$	1.000	1.000	1.000	1.000
$(50, 30)$	1.000	1.000	1.000	1.000
$(100, 10)$	0.998	1.000	0.678	0.808
$(100, 20)$	1.000	1.000	1.000	1.000
$(100, 30)$	1.000	1.000	1.000	1.000
Constant slope	NO	YES	NO	YES

Table 2: Simulation Results, $K = 3$

Bias					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(50, 10)$	-0.290	-0.057	-0.068	-0.079	-0.002
$(50, 20)$	-0.325	-0.020	-0.038	-0.017	-0.012
$(50, 30)$	-0.317	-0.068	-0.040	-0.027	-0.029
$(100, 10)$	-0.304	-0.046	-0.077	-0.084	-0.024
$(100, 20)$	-0.321	-0.050	-0.039	-0.018	-0.017
$(100, 30)$	-0.329	-0.005	-0.031	-0.029	-0.029
Constant slope	-	-	-	NO	YES

MSE					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(50, 10)$	0.703	0.617	0.463	0.503	0.420
$(50, 20)$	0.855	0.727	0.511	0.494	0.491
$(50, 30)$	0.799	0.660	0.483	0.467	0.465
$(100, 10)$	0.443	0.434	0.261	0.274	0.232
$(100, 20)$	0.428	0.330	0.215	0.211	0.211
$(100, 30)$	0.445	0.316	0.224	0.224	0.224
Constant slope	-	-	-	NO	YES

Classification success probability				
(n, T_0)	$\leq 5\%$ misclass.		No misclass.	
$(50, 10)$	0.134	0.948	0.036	0.508
$(50, 20)$	0.934	1.000	0.804	1.000
$(50, 30)$	0.956	1.000	0.946	1.000
$(100, 10)$	0.316	0.988	0.028	0.250
$(100, 20)$	0.996	1.000	0.970	1.000
$(100, 30)$	1.000	1.000	1.000	1.000
Constant slope	NO	YES	NO	YES

Table 3: Simulation Results, Misspecified Number of Types

Bias					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(100, 10)$	-0.469	-0.177	-0.212	-0.165	-0.116
$(100, 20)$	-0.482	-0.069	-0.079	-0.033	-0.014
$(100, 30)$	-0.467	-0.027	-0.019	0.012	0.014
Constant slope	-	-	-	NO	YES

MSE					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(100, 10)$	0.437	0.478	0.254	0.241	0.230
$(100, 20)$	0.433	0.276	0.178	0.180	0.174
$(100, 30)$	0.442	0.298	0.195	0.200	0.200
Constant slope	-	-	-	NO	YES

Table 4: Simulation Results, Weak Separation

Bias					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(100, 10)$	-0.292	-0.203	-0.251	-0.289	-0.217
$(100, 20)$	-0.266	-0.060	-0.087	-0.189	-0.087
$(100, 30)$	-0.264	-0.042	-0.046	-0.113	-0.054
Constant slope	-	-	-	NO	YES

MSE					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(100, 10)$	0.258	0.376	0.256	0.260	0.242
$(100, 20)$	0.257	0.301	0.186	0.225	0.197
$(100, 30)$	0.230	0.262	0.166	0.177	0.168
Constant slope	-	-	-	NO	YES

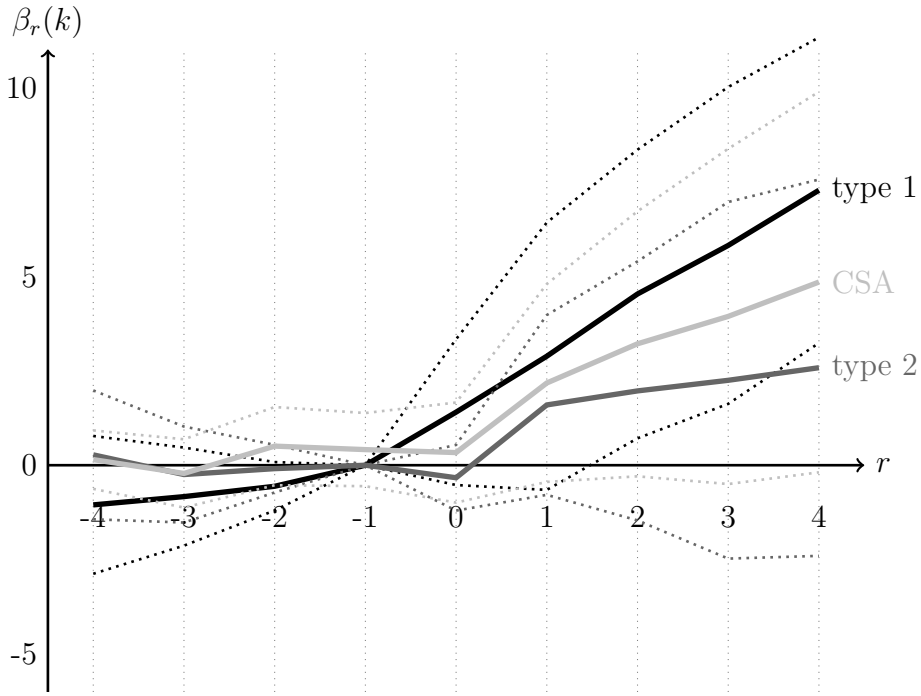
Table 5: Within-type Balancedness Test, $t = 1988$

Type 1	treated	never-treated	Diff
1{central city}	0.38	0.50	-0.13
	(0.52)	(0.52)	(0.23)
% (white)	59.12	63.26	-4.15
	(17.83)	(20.59)	(8.37)
% (hispanic)	7.26	4.21	3.05
	(12.81)	(7.04)	(4.91)
% (free/reduced-price lunch)	39.36	35.65	3.71
	(10.07)	(17.49)	(5.87)
# (student)	56604	62254	-5650
	(38316)	(103167)	(30721)
N	8	14	-
p-value			0.827

Type 2	treated	never-treated	Diff
1{central city}	0.62	0.73	-0.12
	(0.51)	(0.46)	(0.18)
% (white)	46.87	48.14	-1.27
	(22.47)	(21.42)	(8.33)
% (hispanic)	16.43	16.88	0.46
	(16.61)	(19.68)	(6.86)
% (free/reduced-price lunch)	37.27	39.80	-2.53
	(15.25)	(17.87)	(6.26)
# (student)	74862	73790	1072
	(71857)	(154583)	(44612)
N	13	15	-
p-value			0.983

The table reports means of the school district characteristics and their differences across treatment status within each type. The p-value is for the null hypothesis that the means of differences between ever-treated units and never-treated units are all zeros.

Figure 1: Type-specific ATT



The graph reports estimates for the dynamic effect of dismissing court-mandated desegregation plan on the dissimilarity index of a school district. The dissimilarity index ranges from 0 to 100. In 1988, the average dissimilarity index was 34 and the standard deviation was 16.

The red line and the blue line are type-specific DiD estimates for the type-specific dynamic ATT. Type 1 is the type where the dissimilarity index was rising faster and Type 2 is the type where the dissimilarity index was rising slower. The green line is the estimates from Callaway and Sant'Anna (2021); for Callaway and Sant'Anna (2021), only the central city indicator and the percentage of students who are white in 1988 are used, to avoid multicollinearity.

The dotted lines denote the 95% confidence intervals and are computed with asymptotic standard errors.

Table 6: Type-specific Descriptive Statistics, $t = 1988$

	Type 1	Type 2	Diff
dissimilarity index	29.94	37.63	-7.69
	(13.39)	(18.53)	(4.52)
1{central city}	0.45	0.68	-0.22
	(0.51)	(0.48)	(0.14)
% (white)	61.75	47.55	14.20
	(19.30)	(21.51)	(5.78)
% (hispanic)	5.32	16.67	-11.35
	(9.36)	(17.99)	(3.94)
% (free/reduced-price lunch)	37.00	38.63	-1.63
	(15.05)	(16.45)	(4.47)
# (student)	60199	74288	-14089
	(84178)	(121184)	(29096)
N	22	28	-
p-value			0.017

The table reports the group means of the school district characteristics and their differences. The p-value is for the null hypothesis that the means of differences between Type 1 and Type 2 are all zeros.