Finitely Heterogeneous Treatment Effect in Event-study[*]

Myungkou Shin[†]

February 11, 2024

Abstract

The key assumption of the differences-in-differences approach in the event-study design is that untreated potential outcome differences are mean independent of treatment timing: the 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 assumption. With a finite support assumption on the latent type variable, we show that an extremum classifier consistently estimates the type assignment. Based on the classification result, we propose a type-specific diff-in-diff estimator for type-specific CATT. By estimating the CATT with regard to the latent type, 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.

[†] Department of Economics, University of Oxford. email: myungkou.shin@economics.ox.ac.uk

1 Introduction

The event-study design 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 (diff-in-diff) approach, using never-treated units (or last-to-be-treated units) as 'control units.' The key identifying assumption of the diff-in-diff style event-study research design is the parallel trend assumption: temporal differences of untreated potential outcomes are mean independent of treatment status/treatment timing. The parallel trend assumption is a concise and powerful assumption that identifies treatment effects on treated units, while allowing for unobserved unit-level heterogeneity in outcome level. However, when the unit-level heterogeneity goes beyond the heterogeneity in outcome level, an estimator using the parallel trend assumption is susceptible to bias.

The goal of this paper is to relax the parallel trend assumption and to model the unit-level heterogeneity in a more flexible way. For that purpose, we assume that there exists a latent type variable at the unit level. Using the latent type variable, we assume that the usual parallel trend assumption holds, but only within units of the same type: 'type-specific parallel trend.' In a simple two time periods case, the type-specific parallel trend assumption can be written as follows: with some latent type variable k_i,

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

$Y_{it}(\infty)$ denotes untreated potential outcome for unit i at time t and D_i an indicator denoting if unit i is treated at time $t = 2$.

The type-specific parallel trend assumption can be understood as an extension of a conditional parallel trend assumption, replacing the observable pretreatment covariate X_i in the conditioning set of the conditional parallel trend assumption, with the latent type variable k_i (see Abadie (2005); Sant'Anna and Zhao (2020); Callaway and Sant'Anna (2021) among others). The conditional parallel trend assumption is used when the observable covariates associated with the dynamics of the untreated potential outcomes are not balanced across treatment timings. Following the same spirit, we use the latent type variable to model the unit-level unobserved heterogeneity that affects the dynamics of the untreated potential outcomes, while allowing for it to be not balanced across treatment timing.

The difference between the conditional parallel trend assumption with observable covariate and our setup is that the type-specific parallel trend assumption in this paper uses a latent variable which is not observed by the econometrician; the types need to be estimated. For that end, we assume two additional assumptions. Firstly, we assume that the latent type variable k_i has a finite support; hence, the unit-level heterogeneity varies only finitely. Secondly, we assume that the types are sufficiently separated in the domain of the pretreatment outcomes. When the number of pretreatment periods grows to infinity, the separation assumption allows us to classify each unit into their own types consistently.

The finite support assumption gives our framework a unique merit; it helps us in analyzing 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 flexibly treatment effect heterogeneity, the finite support assumption provides a model-based stratifying structure that allows us to summarize the treatment effect heterogeneity. Let $Y_{i2}(2)$ denote the treated potential outcome of unit i at time $t = 2$. The existing literature mostly focuses on estimating 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:

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

The treatment effect parameter shows us how the treatment effect changes along with the latent type variable k_i, which captures unobserved heterogeneity across units, while not imposing any restrictions on the distribution of $Y_{i2}(2) - Y_{i2}(\infty)$.

To apply the type-specific parallel trend framework to datasets, we propose a two-step estimation procedure. In the first step, we use pretreatment outcomes to classify units into the finite number of types; we assume (a dynamic version of) (1) for pretreatment outcomes and apply the K-means clustering algorithm to first-differenced pretreatment outcomes. Given the classification result, the second step of the estimation procedure is to estimate the conditional treatment effect on treated units, using the estimated types as given.[1] In the estimation step, a variety of existing estimation strategies can be used by treating the type as a given categorical variable: De Chaisemartin and d'Haultfoeuille (2020); Borusyak et al. (2021); 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 diff-in-diff' estimator estimates treatment effect by averaging the canonical diff-in-diff 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 diff-in-diff estimators are consistent and asymptotically normal under some regularity conditions. These asymptotic results are supported by Monte Carlo simulations.

To provide an empirical illustration of our method, we revisit Lutz (2011) that studies the effect of dismissing school desegregation plans on racial dissimilarity index at the school district level. Lutz (2011) uses the variation in the timing of the district court ruling that dismisses court-mandated school desegregation plans and uses the first-differenced outcomes

[1] 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.

with census region time fixed-effects. By applying the type-specific parallel trend assumption to Lutz (2011), we find interesting patterns between the pretreatment trend in school dissimilarity index and the treatment effect of dismissing school desegregation plans. Specifically, we find strong segregation effect from dismissing school desegregation plans in school districts where dissimilarity index was worsening even before the dismissal, whereas we find smaller and insignificant segregation effect in school districts where dissimilarity index was rising slower.

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 timings: see Abadie et al. (2010); Arkhangelsky et al. (2021); Athey et al. (2021); Hsiao et al. (2012); Freyaldenhoven et al. (2019); Xu (2017); Chernozhukov et al. (2019); Callaway and Karami (2023); Janys and Siflinger (2024) among others. The interactive fixed-effect model often assumes that the error term is mean zero conditioning on the unit-level factor and therefore nest the type-specific parallel trend assumption by treating the unit-level factor as the type variable; our framework can be thought of as a special case of the interactive fixed-effect model with a finite support on the factor. Janys and Siflinger (2024) takes the same approach; however, Janys and Siflinger (2024) neither explores treatment effect heterogeneity nor develop a full asymptotic theory. As discussed in Athey et al. (2021), one way to compare various estimation procedures suggested in the literature is to compare weights on untreated outcomes that the estimators use in constructing a counterfactual untreated outcome. The type-specific diff-in-diff estimator in this paper applies uniform weights to the untreated observations within the same type to construct a counterfactual outcome. In that sense, the set of weights we consider in this paper is larger than that of the canonical diff-in-diff, but smaller than that of, e.g., synthetic diff-in-diff from Arkhangelsky et al. (2021). Lastly, most of the interactive fixed-effect model literature do not provide a model-based summary of the treatment effect heterogeneity in a way that the latent type structure of this paper does.

As with the type-specific diff-in-diff estimator, most of the papers in the event-study and interactive fixed-effect model literature rely on large pretreatment periods. Notable exceptions are Callaway and Karami (2023) and Freyaldenhoven et al. (2019). Callaway and Karami (2023) do not use a long pretreatment periods by using control variables with time-invariant coefficients in the outcome model as instruments. Freyaldenhoven et al. (2019) also do not require a long pretreatment by using external variables to control for the unit-by-time unobserved heterogeneity. The need for this extra information is the cost of using small pretreatment periods.

Outside of the literature that uses the interactive fixed-effect model, Rambachan and Roth (2022) suggests an alternative framework that relaxes the parallel trend assumption and derives a partial identification result.

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. (2021); Baker et al. (2022); Goldsmith-Pinkham et al. (2022), 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 and provides treatment effect estimators that are robust to this problem. We build upon this literature and construct the type-specific diff-in-diff estimator to be robust to the treatment effect heterogeneity. While doing so, we introduce a new element to the literature: the treatment effect heterogeneity along the dimension of the unobserved unit-level 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 diff-in-diff estimator by revisiting Lutz (2011).

2 Model

For the main model of the paper, we consider a setup where 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. There are $n = N_0 + N_1$ units and $T+1 = T_0+T_1+1$ time periods, with the unit index ranging $i = 1, \cdots, n$ and the time index ranging $t = -T_0-1, \cdots, T_1-1$. N_0 denotes the number of units that are never treated and N_1 denotes the number of units that are treated at some time $0 \leq t \leq T_1 - 1$. For never-treated units, let $E_i = \infty$. WLOG let $\{1, \cdots, N_0\}$ be the set of the never-treated units. $T_0 + 1$ denotes the number of population pretreatment periods and T_1 denotes the number of population treatment periods; $\sum_{i=1}^{n} D_{it} = 0$ for all $t < 0$. $t < 0$ denotes pretreatment periods at the population level and $t \geq 0$ denotes population treatment periods at the population level. T_1 is fixed. Throughout the paper, we use the potential outcome framework to discuss treatment effect:

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

$Y_{it}(e)$ is the potential outcome of unit i at time t when their treatment timing is e. Thus, for some $Y_{it}(e)$, $t < e$ means untreated potential outcome and $t \geq e$ means treated potential outcome.

The key assumption of this paper is that there exists a unit-level latent type variable. Conditional upon the latent type, 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 $t < e$*

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

Assumptions 1-2 identify treatment effect when the types are known. Fix two time periods (s, t) and a treatment timing e such that $s < e \leq t$. The conditional average treatment effect on treated units (CATT) for time t, type k and treatment timing e can be written as follows:

$$\begin{aligned} CATT_t(k, e) &= \mathbf{E}\left[Y_{it}(e) - Y_{it}(\infty)|k_i = k, E_i = e\right] \qquad (2) \\ &= \mathbf{E}\left[Y_{it}(e) - Y_{is}(\infty)|k_i = k, E_i = e\right] - \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty)|k_i = k, E_i = e\right] \\ &= \mathbf{E}\left[Y_{it} - Y_{is}|k_i = k, E_i = e\right] - \mathbf{E}\left[Y_{it} - Y_{is}|k_i = k, E_i > t\right]. \end{aligned}$$

Thus, given $\{k_i\}_{i=1}^n$ is known and $\Pr\{E_i = e|k_i = k\} \cdot \Pr\{E_i > t|k_i = k\} > 0$, $CATT_t(k, e)$ is identified.

Note that the CATT parameter in (2) takes treatment timing E_i as a conditioning variable and focuses on a specific time period t. The full-fledgedness of $CATT_t(k, e)$ 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 CATT parameter in this paper, to highlight the treatment effect heterogeneity across types. To construct the (aggregated) dynamic CATT parameter, we take the average of (2) across (t, e) while maintaining the relative treatment timing $t - e$ fixed: for some $r \geq 0$,

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

$\beta_r(k)$ is the r-times-lagged conditional average treatment effect on treated units. Note that $\beta_r(k)$ is dynamic and type-specific. $\beta_r(k)$ is identified when $\{k_i\}_{i=1}^n$ is known and $\Pr\{E_i \leq T_1 - r|k_i = k\} \cdot \Pr\{E_i = \infty|k_i = k\} > 0$ holds.

Now that we have established identification results for CATT when the types are known, let us adopt two additional assumptions for type classification.

Assumption 3. (Finite Support)

$$k_i \in \{1, \cdots, K\}.$$

The finiteness of the type k_i from Assumption 3 allows us to use the readily available literature of unsupervised partitioning methods to estimate the type. In particular, we use the K-means minimization problem, which will be discussed in detail in Section 3. Once we apply the conventional K-means clustering algorithm to dataset and solve the K-means minimization problem, we take the classification result as our 'estimated' types.

For the classification result to be consistent, we assume that the types are well-separated.

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

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

as $T_0 \to \infty$.

To discuss separation of types, Assumption 4 uses

$$\mathbf{E}[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k],$$

the conditional mean of the first-differenced never-treated potential outcomes. Assumption 4 assumes that 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. From Assumptions 1-2, we have

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

whenever $t < e$. Thus, Assumption 4 can be applied not only to the never-treated units, but also to the pretreatment outcomes of the treated units.

Under Assumptions 3-4, the types of units are identified. Consider the simple case where $\mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|k_i\right] = \delta_i$. δ_i is identified from the time series of $\{Y_{it}\}_{t=-T_0-1}^{-1}$ for each unit $i = 1, \cdots, n$. Assumption 3 assumes δ_i takes only K values—$\delta(1), \cdots, \delta(K)$—and Assumption 4 assumes that $\delta(1), \cdots, \delta(K)$ are distinct values. Thus, the types of units are identified. In the general case of $\mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|k_i = k\right] = \delta_t(k)$, the separation of $\{\delta_t(1)\}_t, \cdots, \{\delta_t(K)\}_t$ on a $\mathbb{R}^{T_0}$ space plays a crucial role in obtaining the consistency of the type classification. More discussion on the type classification result is given in Section 4.

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 CATT. To describe the estimation procedure, let us adopt the following notations:

$$\begin{aligned}
&\gamma := (k_1, \cdots, k_n) \in \Gamma, \\
&\Gamma := \{1, \cdots, K\}^n, \\
&\delta := \{\delta_t(k)\}_{t,k}
\end{aligned}$$

γ is a $n \times 1$ vector of a type assignment. Γ is a set of all possible type assignments where n units are assigned to K different types. δ is a collection of $\delta_t(k)$, the type-specific time trend given time t and type k:

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

In the classification step, we only use a subset of the given data: population pretreatment periods. With the population pretreatment periods, we construct an objective function with mean squared error:

$$\widehat{Q}(\delta, \gamma) = \frac{1}{nT_0} \sum_{i=1}^{n} \sum_{t=-T_0}^{-1} (Y_{it} - Y_{it-1} - \delta_t(k_i))^2 \tag{3}$$

and the resulting first-step classifier is

$$\left(\hat{\delta}, \hat{\gamma}\right) = \arg\min_{(\delta,\gamma)\in\mathcal{D}\times\Gamma} \widehat{Q}\left(\delta, \gamma\right). \tag{4}$$

$\mathcal{D} = [-M, M]^{T_0}$ with some $M > 0$. The minimization problem in (3) 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 our minimization problem (3), the centeroids are denoted with $\{\delta_t(1)\}_{t<0}, \cdots, \{\delta_t(K)\}_{t<0}$ and the grouping structure is denoted with $k_1, \cdots, k_n$.

The algorithm that we use to obtain (4) 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
$$\hat{\delta}_t^{(s)}(k) = \frac{\sum_{i=1}^{n} (Y_{it} - Y_{it-1}) \mathbf{1}\{k_i^{(s)} = k\}}{\sum_{i=1}^{n} \mathbf{1}\{k_i^{(s)} = k\}}$$
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

in $\hat{\delta}^{(s)}$: set S and ε such that the iteration stops when

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

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 (3). 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 associated 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: refer to Bai and Ng (2002); Bonhomme and Manresa (2015); Janys and Siflinger (2024). More discussion on the choice of K is given in the Supplementary Appendix.

Given the first-step classification result, the type-specific diff-in-diff estimator for the full-fledged CATT parameter $CATT_t(k,e)$ can be constructed by taking sample means for each type:

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

In the case of the dynamic CATT parameter $\beta_r(k)$, the type-specific diff-in-diff estimator is

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

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

$$\mu(k,e) := \Pr\{k_i = k, E_i = e\}.$$

Note that there exist multiple ways of constructing an estimator for the dynamic CATT parameter $\beta_r(k)$. As discussed in Callaway and Sant'Anna (2021), there is no straightforward choice in picking which time differences to use in a diff-in-diff type approach. Though the estimator described above takes one period before the treatment timing to construct a time difference, other choices such as two periods before the treatment timing are equally valid as long as the parallel trend assumption holds for every time period.[2] Also, the estimator uses the never-treated units to use as control units. When there is no never treated units, the latest treatment cohort can take up the same role. In that case, the definition of the dynamic CATT $\beta_r(k)$ will be adjusted in a way that it does not include the latest treatment cohort anymore.

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

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

for some $r < -1$. From Assumption 2, $\mathbf{E}\left[Y_{i,e+r}(e) - Y_{i,e+r}(\infty) | k_i = k, E_i\right] = 0$ whenever $r < -1$. Thus, though Assumption 1 does not have a testable implication, we can use $\hat{\beta}_r(k)$ for $r < -1$ to test Assumption 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. Firstly, to derive the classification result for the type estimator defined in (4), let us adopt

[2] Roth and Sant'Anna (2023a) discusses efficiency of these diff-in-diff type estimates when the treatment timing is truly random. More discussion on this is given the Appendix.

following assumptions.

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 $n \to \infty$.

Proof. Theorem 1 is nested in Theorem 2 by connecting Assumption 5 to Assumption 7. Assumption 5-b induces parts of Assumption 7-b,d concerning U_{it} and $\delta_t(k)$, by letting $U_{it} = Y_{it}(E_i) - \mathbf{E}[Y_{it}(\infty)|k_i]$. Assumption 5-e provides the weak dependence conditions for which the proof for Theorem 2 uses Assumption 7-g. □

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 $\mathcal{N}(0, 1)$; the probability of imperfect classification is nonzero. Thus, we need large pretreatment periods ($\Leftrightarrow T_0 \gg 0$), in addition to the strong separation ($\Leftrightarrow \delta(2) - \delta(1) > 0$).[3] 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.

Given the long pretreatment, the bound on the misclassification probability from Theo-

[3] 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.

rem 1 can be used to derive asymptotic properties of the type-specific diff-in-diff estimator. For that, let us adopt the additional assumption below. Recall $\mu(k,e) = \Pr\{k_i = k, E_i = e\}$.

Assumption 6. *For each $k = 1, \cdots, K$, there exists some $\bar{r}_k \geq 0$ such that*

$$\bar{r}_k = \max\left\{r \geq 0 : \mu(k, T_1 - 1 - \bar{r}_k) \cdot \mu(k, \infty) > 0\right\}.$$

For any t, s and e, $\mathrm{Var}\left(Y_{it}(e) - Y_{is}(e) | k_i, E_i\right) > 0$.

Assumption 6 assumes that each type has nonzero measure of never-treated units and finds an upper bound $\bar{r}_k$ on how far the dynamic treatment effects can be estimated. Note that

$$\bar{r}_k \geq r \quad \Rightarrow \quad \mu(k,e) > 0 \text{ for some } e \text{ such that } e + r \leq T_1 - 1.$$

For every $0 \leq r \leq \bar{r}_k$, r-times-lagged treatment effect can be estimated for type k.

Corollary 1. *Let Assumptions 1-6 hold. There exists some $\nu^* > 0$ such that $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. Then, for any k and $r \leq \bar{r}_k$ with some permutation on $\{1, \cdots, K\}$,*

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

with some $\sigma^2 > 0$, as $n \to \infty$.

Proof. Corollary 1 is nested in Corollary 3. □

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 CATT parameter $\beta_r(k)$, treatment timing distribution 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 $t < 0$,

$$Y_{it} - Y_{it-1} = \delta_t(k_i) + X_{it}^\intercal \theta + U_{it}. \tag{5}$$

Note that the interpretation of $\delta_t(k)$ is changed. Within the linear model, $\delta_t(k)$ in (5) 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. The type-specific time fixed-effects explains heterogeneity across units that is not explained by the (linear) observable control covariates X_{it}.

Given the model (5), we can construct a similar objective function from before and solve the K-means minimization problem for classification:

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

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

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, we 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 $n \to \infty$.

Proof. See Supplementary Appendix. □

Remark 3. When X_{it} is time-invariant, i.e. $X_{it} = X_i$, the linear model (5) and Assumption 7 can be understood as a 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 (5), we can consider a linear model on the level of the outcome:

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

and thus

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

The assumptions for the linear model in level is discussed in the Appendix along with Assumption 7.

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 (5). 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. A most straightforward example is to use type-specific coefficient for the treatment variable: for $t \geq 0$,

$$Y_{it} = \alpha_i + \delta_t(\hat{k}_i) + \sum_{r\geq 0} \beta_r(\hat{k}_i)\mathbf{1}\{t = E_i + r\} + X_{it}^{\intercal}\theta + U_{it}. \tag{6}$$

A more discussion on the outcome model approach is discussed in Section C.2 of the Appendix. The outcome model approach has the merit of developing a parsimonious model for treatment effect; using the outcome model approach, we can impose some structure over how the latent type k_i and the observable control covariate X_{it} interact in terms of the treatment effect heterogeneity.

Alternatively, we can abstract away from imposing restriction on the outcome variable and use an assignment model approach. In the assignment model approach, instead of imposing a parametric model for the post-treatment outcomes, we impose a parametric model for the treatment timing. 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. The assignment model approach is an extension of Corollary 1 since the type-specific diff-in-diff estimator defined in Section 2 is what we get when we assume the propensity score to be a trivial function of X_i: $\Pr\left\{E_i = e | k_i, X_i\right\} = \Pr\left\{E_i = e | k_i\right\}$. A more discussion on the assignment model approach is discussed in Section C.3.

6 Simulation

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

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

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

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

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: $\sigma = 2.02$ and $\rho = 0.68$ for the error distribution and $\min_{k \neq k'} |\delta(k) - \delta(k')| = 1.32$ for the type separation.[4] Note that a simple mean comparison of the treated units and untreated units is a biased estimator for the treatment effect when $\pi(k)$ is not constant in k.

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: $\{\delta_t(k)\}_{t \le 0}$. Secondly, we imposed a constant slope restriction $\delta_t(k) = \delta_{t'}(k)$ for every t, t' and estimated only one parameter for each type: $\delta(k)$. Given the two type classifications,

[4] The rest of the simulation parameters are as follows. For $K = 2$, we set $(\pi(1), \pi(2)) = (1/3, 2/3)$, $(\alpha(1), \alpha(2)) = (37, 39)$, $(\delta(1), \delta(2)) = (1.32, 0)$ and $(\beta(1), \beta(2)) = (3, 0)$. $\Pr\{k_i = 1\} = \Pr\{k_i = 2\} = 1/2$. For $K = 3$, we set $(\pi(1), \pi(2), \pi(3)) = (1/2, 3/4, 1/4)$, $(\alpha(1), \alpha(2), \alpha(3)) = (37, 39, 35)$, $(\delta(1), \delta(2), \delta(3)) = (2.74, 1.42, 0)$ and $(\beta(1), \beta(2), \beta(3)) = (2.75, 0, 0)$. $\Pr\{k_i = 1\} = \Pr\{k_i = 2\} = 2/5$ and $\Pr\{k_i = 3\} = 3$. Except for π and β, all numbers are taken from the empirical results in Section 7.

we estimated the ATT using the type-specific diff-in-diff estimators; for comparison, we considered the diff-in-diff, the synthetic control and the synthetic diff-in-diff estimators.

Table 1 and Table 2 contain the simulated bias and the simulated MSE from 500 random samples. Also, they contain some summary statistics for the finite-sample performance of the classification step. For large $T_0 = 30$, both the type-specific diff-in-diff estimator and the synthetic diff-in-diff estimator perform well since there are many pretreatment outcomes to be used to control for the unit-level heterogeneity. However, for small $T_0 = 10$, the type-specific diff-in-diff estimator outperforms the other estimators since it best reflects the finite type structure in dataset. As for the classification with $K = 2$, we see near-perfect classification in more than 90% of the random samples for small $T_0 = 10$, when the correct smoothness restriction is imposed. Even for the flexible time trend specification where $\delta_t(k)$ varies across every t, a relatively small $T_0 = 20$ attains perfect classification in more than 95% of the samples. When we add a third type and let $K = 3$, the classification accuracy worsens, but not by much; $T_0 = 20$ attains perfect classification, with or without the smoothness restriction, in more than 88% of the samples. For all simulation specifications, $T_0 = 30$ attains perfect classification in more than 99% of the samples.

7 Application

To show how the type-specific diff-in-diff estimator applies to a real dataset, we revisit Lutz (2011). Since the Supreme Court ruling on Brown v. Board of Education of Topeka in 1954 that found state laws in US 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 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

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

$$b_j: \ \# \text{ of black students in school } j, \quad w_j: \ \# \text{ of white students in school } j$$

$$J_i: \ \text{the set of school in school district } i,$$

$$B_i = \sum_{j\in J_i} b_j, \quad W_i = \sum_{j\in J_i} w_j,$$

The dissimilarity index ranges from 0 to 100, with 100 being perfectly segregated schools and 0 being perfectly representative schools.[5]

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. We use the following linear model for

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

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

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

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 the purpose of comparison, here we present the main empirical specification of Lutz (2011):

$$Y_{it} - Y_{it-1} = \delta_{jt} + \sum_{r=-4}^{10} \beta_r \mathbf{1}\{t = E_i + r\} + X_i{}^{\intercal}\theta_t + U_{it} \tag{7}$$

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: $X_i = X_{i,-T_0-1}$.[6] 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: $\delta_t(k) = \delta(k)$.[7] Then, we applied the K-means clustering classifier with $K = 2$.[8] The first-step classification assigns 8 treated units

[6] 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.

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

[8] 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 $K = 3$. For more discussion, see

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

Given the first-step classification result, we conducted a within-type balancedness test; Table 3 contains within-type balancedenss test using control covariates from $t = 1988$. Within the two types, the control covariates are well-balanced across treatment status: treated v. never-treated. Thus, we apply the unweighted type-specific diff-in-diff estimator from Section 3. Figure 1 contains the type-specific diff-in-diff 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 pooled regression with control covariates from Lutz (2011) estimated the dynamic treatment effect to be around 4-5 at $r = 4$, depending on specifications, whereas averaging the type-specific diff-in-diff estimates across types gives us estimate 4.00; the census region fixed-effects is successful in estimating the average effect, though it does not explore the treatment effect heterogeneity. 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 estimates for $\beta_r(k)$ such that $r < 0$; none of the pretreatment trend was found to be away from zero at 0.05 significance level.

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 4 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 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

Supplementary Appendix.

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: for example, 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 a 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 diff-in-diff 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, "Semiparametric difference-in-differences estimators," *The review of economic studies*, 2005, *72* (1), 1–19.

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

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

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

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

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

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

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

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," *arXiv preprint arXiv:2108.12419*, 2021.

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

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

Chernozhukov, Victor, Christian Hansen, Yuan Liao, and Yinchu Zhu, "Inference for Heterogeneous Effects using Low-Rank Estimation of Factor Slopes," 2019.

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

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

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

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

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár, "Contamination bias in linear regressions," Technical Report, National Bureau of Economic Research 2022.

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

Hsiao, Cheng, H Steve Ching, and Shui Ki Wan, "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*, 2012, *27* (5), 705–740.

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

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

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," Technical Report, Working Paper 2022.

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

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

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

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

Xu, Yiqing, "Generalized synthetic control method: Causal inference with interactive fixed effects models," *Political Analysis*, 2017, *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 of the outcome variable, they may be motivated 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 a unconfoundedness assumption would be free of this commitment to a functional form. When

$$\left\{Y_{it}(e)\right\}_{t,e} \perp\!\!\!\perp E_i|k_i, \tag{8}$$

a parallel trend assumption with any functional form would hold.[9] This comes at a cost of assuming distributional independence, which is stronger than the mean independence used in 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

[9] This statement is only true for the unconfoundedness assumption as given in (8). 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 $Y_{i2}(\infty) \perp\!\!\!\perp D_{i2}|\big(k_i, Y_{i1}\big)$

timing or the unconfoundedness assumption, we have some theoretical guidance on this choice. Roth and Sant'Anna (2023a) assume random treatment timing and find an efficient estimator among diff-in-diff type estimators that uses different weights across different temporal differences. Given the same classification result from Theorem 1, the unconfoundedness assumption (8) can be used to find an efficient type-specific diff-in-diff estimator following the procedure of Roth and Sant'Anna (2023a), using the same argument from the proof for Corollary 2: the classification error is faster than $1/\sqrt{n}$.

When a researcher does choose to follow 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 (2023b); Ghanem et al. (2022) provide insights to this question. Roth and Sant'Anna (2023b) 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 (8) is indeed strictly stronger than the type-specific parallel trend assumption holding for every monotone 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 diff-in-diff estimator compares to other treatment effect estimators that assign weights over control units to construct a counterfactual outcome. In specific, we consider the conventional diff-in-diff estimator and the synthetic control estimator.

Find that the classification result from (4) 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. In light of this, we can compare the type-specific diff-in-diff estimator with existing methods in terms of the weights that it considers. Consider a simple case where there is only one post-treatment period and only one treated unit: $T_1 = 1$ and $N_1 = 1$. $E_i = \infty$ for every $i \leq N_0$ and $E_n = 0$. Consider a treatment effect estimator $\hat{\beta}$ which can be written as a weighted sum of Y_{it}: $\hat{\beta} = \sum_{i,t} w_{it} Y_{it}$. In a simple diff-in-diff estimator using $t \in \{-1, 0\}$, 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 diff-in-diff,

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

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 diff-in-diff estimator, the type-specific diff-in-diff estimator admits more flexible cross-sectional weights by possibly using only a subset of the never-treated units. Compared to the synthetic control estimator, the type-specific diff-in-diff estimator is less flexible in terms of the cross-sectional weights since it is dichotomous cross-sectionally; a never-treated unit gets a uniform weight if and only if it shares the same type with the treated unit and gets zero weight otherwise. However, the synthetic control estimator assigns nonzero weights only to contemporaneous outcomes while the type-specific diff-in-diff estimator takes temporal difference. Lastly, though the weights are not as straightforward as with other methods discussed here, the synthetic diff-in-diff 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 Asymptotic results with control covariate X_{it}

C.1 Type estimation with pretreatment outcomes

To have the classification result under the linear outcome model with control covariates (5), 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 $t < 0$, $\mathbf{E}\big[U_{it} | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\big] = 0$ and $\mathbf{E}\big[{U_{it}}^4 | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\big] \leq M$.

  For every $t, s < 0$, $\mathbf{E}\big[{X_{it}}^\intercal X_{is}\big] \leq M$. 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 $\alpha[t] \leq \exp(-a t^{d_1})$. Also, with* *some positive constant d_2 and b, U_{it} satisfies the following tail probability: for any $u > 0$,*

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

uniformly over i and $t < 0$.

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 (5). 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 (5) 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 that there is 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 an adjusted version of Assumption 7-h by replacing X_{it} with $X_{it} - X_{it-1}$ give us the same classification result as in Theorem 2.

C.2 Outcome model approach

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 (6) can be helpful when we are interested in treatment effect heterogeneity and we would like to impose some restrictions on the heterogeneity due to 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 $t \geq 0$,

$$Y_{it} - Y_{it-1} = m(X_{it}, k_i; \xi) + U_{it}.^{10}$$

In the example (6), $\xi = \Big(\{\delta_t(k)\}_{t\geq 0,k}, \{\beta_r(k)\}_{r\geq 0,k}, \theta\Big)$. Note that the dimension of ξ is fixed; the dimension is $2T_1K + p$ and T_1 and K are fixed. Let $\tilde{\xi}$ be the infeasible least-square estimator for ξ and $\hat{\xi}$ be the plug-in least-square estimator for ξ:

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

Assumption 8. *Ξ, the parameter space for ξ, is bounded: with some $M > 0$,*

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

[10] Though the first-differenced outcome variables are used in the post-treatment outcome model for internal consistency with (5), we can also consider models with outcome variable in level. In that case, one could use unit fixed-effects or treatment-timing-by-type fixed-effects to address unit-level heterogeneity in level.

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

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

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

Corollary 2. *Let Assumptions 3 and 7-8 hold. 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 $n \to \infty$.

Proof. The result is direct from finding 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 $CATT_t(k,e)$ or $\beta_r(k)$ in Section 3. In the example of (6), 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.[11] Thus, we

[11] 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. (2021); Sun and Abraham (2021) among others) and potentially distort hypothesis testing (see Baker et al. (2022)). Also, Goldsmith-Pinkham et al. (2022) show that even when the treatment timing is random, treatment effect estimators still suffer from contamination bias when dynamic treatment effect specification is used.

consider an alternative approach in the next subsection.

C.3 Assignment model approach

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 the control covariates 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 the control covariate X_i.

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 the control covariate X_i and the latent type k_i:[12]

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

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

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

[12] 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.

An example is an ordered logistic model:

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

In this example, $\xi = (\theta, \delta_e(k))_{k,e}$ and its dimension is fixed: $T_1 K + p$. Using $\hat{\xi}$, the type-specific diff-in-diff estimators can be constructed as follows:

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

where

$$\begin{aligned}
\widehat{CATT}_t(k,e) &= \frac{\sum_{i=1}^{n} \left(Y_{it} - Y_{i,e-1}\right) \mathbf{1}\{\hat{k}_i = k, E_i = e\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = e\}} \\
&\quad - \frac{\sum_{i=1}^{n} \left(Y_{it} - Y_{i,e-1}\right) \mathbf{1}\{\hat{k}_i = k, E_i = \infty\} \pi_e(X_i, k, \hat{\xi}) / \pi_{\infty}(X_i, k, \hat{\xi})}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = \infty\} \pi_e(X_i, k, \hat{\xi}) / \pi_{\infty}(X_i, k, \hat{\xi})} \\
\hat{\mu}(k,e) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = e\}.
\end{aligned}$$

To discuss asymptotic properties of the type-specific diff-in-diff estimator, we adopt the following assumption:

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,*

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

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

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

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 with regard to $\|\cdot\|_2$ on B_ξ.*

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

Corollary 3. *Let Assumptions 3 and 6-9 hold by replacing X_{it} with X_i. 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, 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 $\{1, \cdots, K\}$,

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

with some $\sigma^2 > 0$, as $n \to \infty$.

Proof. See Supplementary Appendix. □

D Tables and figures

Table 1: Simulation Results, $K = 2$

Bias

(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(50, 10)$	-0.474	-0.245	-0.285	-0.220	-0.073
$(50, 20)$	-0.489	-0.108	-0.125	-0.023	-0.020
$(50, 30)$	-0.453	-0.090	-0.087	-0.030	-0.030
$(100, 10)$	-0.458	-0.194	-0.262	-0.134	-0.088
$(100, 20)$	-0.446	-0.059	-0.092	-0.003	-0.003
$(100, 30)$	-0.440	-0.031	-0.043	-0.007	-0.007
Constant slope	-	-	-	NO	YES

MSE

(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(50, 10)$	0.605	0.692	0.484	0.469	0.370
$(50, 20)$	0.589	0.662	0.367	0.325	0.322
$(50, 30)$	0.603	0.701	0.410	0.385	0.385
$(100, 10)$	0.417	0.421	0.272	0.212	0.199
$(100, 20)$	0.380	0.297	0.174	0.156	0.156
$(100, 30)$	0.411	0.314	0.189	0.185	0.185
Constant slope	-	-	-	NO	YES

Classification success probability

(n, T_0)	$\leq 5\%$ misclass.		No misclass.	
$(50, 10)$	0.462	0.926	0.108	0.384
$(50, 20)$	1.000	1.000	0.952	1.000
$(50, 30)$	1.000	1.000	1.000	1.000
$(100, 10)$	0.798	0.988	0.072	0.152
$(100, 20)$	1.000	1.000	0.988	0.998
$(100, 30)$	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
$(50, 10)$	0.372	0.160	0.145	0.163	0.068
$(50, 20)$	0.362	0.026	0.039	-0.006	-0.012
$(50, 30)$	0.346	0.044	0.025	0.016	0.017
$(100, 10)$	0.308	0.110	0.065	0.069	0.015
$(100, 20)$	0.310	0.014	0.012	-0.022	-0.022
$(100, 30)$	0.368	0.017	0.011	-0.005	-0.005
Constant slope	-	-	-	NO	YES

MSE					
(n, T_0)	DiD	SC	synthetic DiD	type-specific DiD	type-specific DiD
$(50, 10)$	0.682	0.753	0.457	0.506	0.447
$(50, 20)$	0.678	0.658	0.412	0.435	0.435
$(50, 30)$	0.740	0.660	0.387	0.419	0.417
$(100, 10)$	0.371	0.413	0.213	0.243	0.226
$(100, 20)$	0.373	0.329	0.190	0.207	0.205
$(100, 30)$	0.409	0.270	0.177	0.195	0.195
Constant slope	-	-	-	NO	YES

Classification success probability				
(n, T_0)	$\leq 5\%$ misclass.		No misclass.	
$(50, 10)$	0.118	0.872	0.026	0.308
$(50, 20)$	0.976	1.000	0.884	1.000
$(50, 30)$	0.996	1.000	0.990	1.000
$(100, 10)$	0.342	0.975	0.022	0.078
$(100, 20)$	1.000	1.000	0.964	1.000
$(100, 30)$	1.000	1.000	1.000	1.000
Constant slope	NO	YES	NO	YES

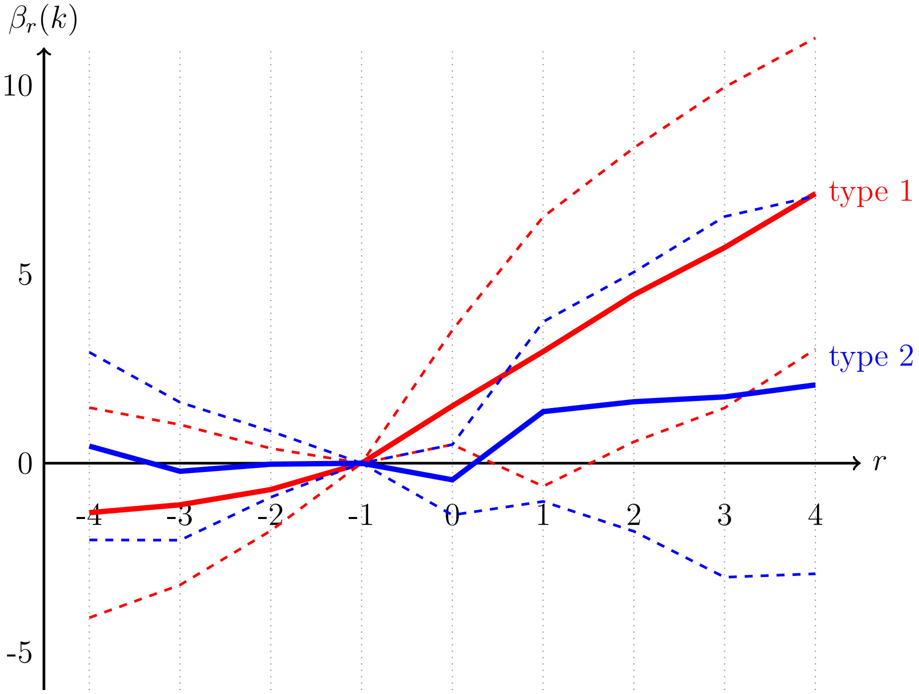
TABLE 3: WITHIN-TYPE BALANCEDNESS TEST, $t = 1988$

Type 1	treated	never-treated	Diff
$\mathbf{1}$\{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
$\mathbf{1}$\{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 treated units and never-treated units are all zeros.

Figure 1: Type-specific CATT



The graph reports the type-specific diff-in-diff estimates for the 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.

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 dashed lines denote the confidence intervals are at 0.05 significance level and are computed with asymptotic standard errors.

Table 4: 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)
$\mathbf{1}\{\text{central city}\}$	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.