NBER WORKING PAPER SERIES

A FLEXIBLE, HETEROGENEOUS TREATMENT EFFECTS DIFFERENCE-IN-DIFFERENCES ESTIMATOR FOR REPEATED CROSS-SECTIONS

Partha Deb
Edward C. Norton
Jeffrey M. Wooldridge
Jeffrey E. Zabel

Working Paper 33026
http://www.nber.org/papers/w33026

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2024

The authors thank conference participants at the New York Camp Econometrics XVIII, at the American Society of Health Economists Annual Meeting and 31st European Workshop on Econometrics and Health Economics, and seminar participants at Oregon Health Sciences University. We also thank Anjelica Gangaram and Govert Bijwaard for very helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Flexible, Heterogeneous Treatment Effects Difference-in-Differences Estimator for Repeated Cross-Sections

Partha Deb, Edward C. Norton, Jeffrey M. Wooldridge, and Jeffrey E. Zabel

NBER Working Paper No. 33026
October 2024
JEL No. C01, C21, C50

ABSTRACT

This paper proposes a method to estimate treatment effects in difference-in-differences designs in which the treatment start is staggered over time and treatment effects are heterogeneous by group, time, and covariates, and when the data are repeated cross-sections. We show that a linear-in-parameters regression specification with a sufficiently flexible functional form consisting of group-by-time treatment effects, two-way fixed effects, and interaction terms yields consistent estimates of heterogeneous treatment effects under general conditions. The estimates are efficient and aggregation of treatment effects and inference are straightforward. We call it FLEX, because it is a flexible linear model estimated by OLS with covariates (X). We illustrate the use of FLEX with two empirical examples and provide comparisons to other recently derived estimators.

Partha Deb
Hunter College
Department of Economics
695 Park Avenue
Room 1524 West
New York, NY 10065
and NBER
partha.deb@hunter.cuny.edu

Edward C. Norton
Department of Health Management and Policy
Department of Economics
University of Michigan
School of Public Health
1415 Washington Heights, M3108 SPHII
Ann Arbor, MI 48109-2029
and NBER
ecnorton@umich.edu

Jeffrey M. Wooldridge
Department of Economics
Michigan State University
East Lansing, MI 48824
wooldri1@msu.edu

Jeffrey E. Zabel
Department of Economics
Tufts University
Medford, MA 02155
jeff.zabel@tufts.edu

1 Introduction

The difference-in-differences (DID) study design is an important tool for causal inference in economics. In recent years, there has been an extraordinary number of new theoretical papers on how to obtain DID estimates in regression-based implementations of these study designs. In particular, the discovery that the two-way (group and time) fixed effects estimator of a model with a constant treatment effect can produce biased estimates of treatment effects when there is staggered treatment timing and heterogeneous treatment effects (e.g., de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021), has led to new approaches for dealing with both staggered timing and heterogeneous treatment effects (e.g., Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).

What remains unknown is how the recent proposed methods to address heterogeneous treatment effects and staggered treatment timing work with repeated cross-sectional data that are common in applied research, as opposed to balanced panel data. The theoretical results showing consistent estimates of heterogeneous treatment effects with staggered treatment timing have been proven with panel data (Wooldridge, 2021; Borusyak et al., 2024), but not yet with repeated cross-sections. Proofs with repeated cross-sections are more complicated because covariates that are inherently time invariant are instead time varying when averaged at the group level (also true at the individual level) because of the changing composition of individuals over time. Furthermore, it is not possible to include only pre-treatment values of potentially endogenous time-varying variables, because these data are not observed for individuals who only appear once in the cross-sectional sample.

This paper addresses theoretical and empirical issues with repeated cross-sections with heterogeneous treatment effects and staggered treatment timing. We extend the methods proposed in Wooldridge (2021) to derive an estimator for repeated cross-sections under clearly stated assumptions that allow control variables to appear in a flexible way. We propose a linear-in-parameters specification with sufficient generality to be valid under general circumstances. Given a sample of repeated cross-sections, this specification can be estimated using ordinary least squares (OLS), which is consistent and efficient under the parallel trends and no anticipation assumptions. We show that this approach is closely

related to the imputation method in Borusyak et al. (2024).

Our proposed DID method estimates a regression with heterogeneous treatment effects using OLS. It is simple and transparent. It is easy to know how identification is achieved, which treatment observations are compared to which control observations, and how many parameters are estimated. Because our method uses OLS regression, it is efficient in the class of linear estimators. Our method is flexible in how to incorporate covariates, which can be included additively or in a way that allows the treatment effects to vary with covariates. Our method has precise estimates, even in flexible model specifications with numerous treatment parameters. In fact, the standard errors in our applications are smaller than those found using Callaway and Sant'Anna or stacked regressions (Callaway and Sant'Anna, 2021; Cengiz et al., 2019). We call it FLEX, because it is a flexible linear model estimated by OLS with covariates (X).

In addition to the theoretical results, we demonstrate the use of FLEX with two empirical examples. Both of these policy-relevant examples use individual-level data from repeated cross-sections, have policies that were implemented by cohorts of states staggered over a number of years, and are unlikely to have constant treatment effects at the group-time (state-time) level. We compare the estimated average treatment effects on the treated (ATET) and their standard errors with those obtained using other popular methods (Cengiz et al., 2019; Callaway and Sant'Anna, 2021). The examples demonstrate the features of our proposed FLEX estimator for repeated cross-sections with heterogeneous treatment effects, including being easy to estimate, transparent about the number of parameters estimated, allowing flexible controls for individual-level covariates, and having efficient standard errors. We also show how the results from our regressions can be displayed in a variety of commonly used graphical forms. The Stata code is available upon request.

2 Literature Review

2.1 Theoretical literature

Our paper is related to the extensive recent developments in econometrics about the estimation of treatment effects in difference-in-differences designs. De Chaisemartin and

D'Haultfœuille (2020) and Goodman-Bacon (2021) showed that difference-in-differences regressions that control for cohort and year fixed effects can identify treatment effects only when the treatment effect is constant over time. This common method, often called two-way fixed effects, uses observations in already treated groups as controls for other observations that start treatment later, in addition to using never-treated observations as controls. The overall average treatment effect is then a weighted average of comparisons that include earlier treated units to later treated units, instead of only treated to never-treated units plus not yet treated units.

Several other authors have proposed methods to avoid such unwanted comparisons. Callaway and Sant'Anna (2021) compare treated units in pre- and post-treatment periods to the period just prior to treatment to the same comparison for controls that can include the never- and not-yet treated units. Sun and Abraham (2021) use an event-study approach by exploiting both leads and lags. Borusyak et al. (2024) take a different approach and use what they call *imputation* to sweep out the effects of covariates and then use the residuals to estimate the treatment effects. Cengiz et al. (2019) propose a procedure that creates samples (stacks) of treated cohort observations pooled with never-treated observations, then combines those samples, which resolves the issue of unwanted comparisons. Cohort-level treatment effects are then estimated using a linear regression commonly referred to as a *stacked regression*. Wooldridge (2021) proposes an extended two-way fixed effects regression approach that includes interactions between treated cohorts and time and covariates, allowing for estimation of heterogeneous treatment effects by cohort, time and covariates. All these methods compare treated observations to not-yet-treated and never-treated observations, but never compare them to previously treated observations. There are recent summary papers by de Chaisemartin and D'Haultfœuille (2023), Roth et al. (2023), and Freedman et al. (2023).

2.2 Contribution

Our paper has several theoretical contributions. We propose FLEX—flexible linear model estimated by OLS with covariates (X)—for heterogeneous treatment effects in a difference-in-differences study design with repeated cross-sectional data and staggered treatment starting

times. The linear-in-parameters specification can be estimated in a single linear regression. FLEX delivers consistent parameter estimates, allows for a flexible functional form with respect to covariates, and provides access to the OLS toolbox for inference and specification testing.

We prove that the FLEX treatment effect parameter estimates are asymptotically unbiased estimates of the group-time heterogeneous treatment effects that can be obtained in the repeated cross-section setting by an imputation method. The FLEX estimates are identical to those from the Borusyak et al. (2024) imputation estimator extended to the case that allows for treatment-effect heterogeneity by group and time (not simply cohort and time). Therefore, this imputation estimator extends both Borusyak et al. (2024) and Wooldridge (2021) to the group-time level for repeated cross-sectional data.

Aggregated treatment effects, such as the average treatment effect on the treated (ATET), can be derived easily from the estimated heterogeneous treatment effects.

3 Theory

3.1 The Population Setting, Definitions, and Assumptions

Our goal is to set out a framework that focuses on the analysis of repeated cross sections where new units, i, belonging to one and only one group g, where $g = 1, 2, ..., G$, are (randomly) sampled from the population in every time period $t = 1, 2, ..., T$. Our framework also applies to panel data settings, most transparently when units i in groups g are followed over time t. In that case we are studying a stable population. When the data available are repeated cross sections, we assume the existence of a stable population from which units are randomly sampled in each time period. In defining parameters and stating assumptions, we assume that the populations are the same across t. But, in practice, there may be changes in the population across t, a problem that can, at least partly, be overcome by controlling for observable characteristics.

We now turn to the policy setting. We assume that the intervention occurs at the group level indexed by $g = 1, 2, ..., G$. Each group, among those that receive treatment in the period of observation $t = 1, 2, ..., T$, receives the intervention for the first time in a time period from

$t = 2, ..., T$ implying that there are no "always treated" groups. One or more groups ($< G$) receive the intervention for the first time in a particular time period. Let q denote the first time period in which treatment is received. We follow the existing literature by referring to the collection of groups associated with q as a cohort, $c = q$. Other subsets of untreated groups receive the intervention in subsequent time periods. Let $\bar{g}$ groups receive treatment by time period $Q \leq T$. Without loss of generality, let groups $g = 1, 2, ..., \bar{g}$ index the treated groups ordered by their cohort membership. Consequently, the never-treated groups can be indexed by $g = \bar{g} + 1, ..., G$. In what follows, it is convenient to refer to never-treated groups as being "treated" at ∞, i.e., $c = \infty$. Note that new entrant groups are not needed for each and every period. Also, note that groups, g, must map one-to-one to their cohort, i.e., the first time period in which the treatment is received.

Define potential outcomes by treatment group g in time period t,

$$Y_t(g)\,,\; g \in \{1, ..., G, \}\,,\; t \in \{1, 2, ..., T\}\,.$$

The assumption of a stable population implies that each unit in the population has the full set of potential outcomes in each time period t. For now, treatment assignment is absorbing, i.e., it is not reversible.

Let $R_g \in R_1, ..., R_{\bar{g}}, R_{\bar{g}+1}, ..., R_G$ denote a binary indicator for group membership. Let $\{R_g\}^c$ denote the subset of groups in cohort c. The ATETs commonly of interest in staggered DID settings are the mean differences in the potential outcomes using $Y_t(\infty)$ as the reference outcome in a treated period t, i.e., the outcomes in the groups associated with the never-treated groups:

$$\tau_{gt} = E\left[Y_t(g) - Y_t(\infty)\,|R_g = 1\right],\; t = q, ..., T;\;\; g = 1, ..., \bar{g}. \tag{3.1}$$

For each (eventually) treated group g, τ_{gt}, $t = q, ..., T$ are the ATETs in all time periods including the first period in which treatment is received, q, and all following ones through the end of the observation period, T.

We allow potential outcomes to be conditioned on a set of observed covariates that can be indexed by time: $\{\mathbf{X}_t = 1, ..., T\}$, where $\mathbf{X}_t$ is a $1 \times K$ vector. Much of the literature assumes the controls are dated prior to the first intervention date and do not change over time. This

restriction helps ensure that one is not including 'bad controls' in the analysis—that is, elements in $\mathbf{X}_t$ that might be affected in the current or future periods by the intervention. We do not index the $\mathbf{X}_t$ using potential outcomes notation [such as $\mathbf{X}_t(q)$], and so we are maintaining that the covariates do not change with the treatment assignment. (This is different from saying that the treatment assignment cannot depend on $\mathbf{X}_t$—which, of course, we allow.) Allowing $\mathbf{X}_t$ to have time variation means that we can include predictors of the outcome whose paths are not influenced by treatment. For example, the outcome may be affected by local weather conditions, which are time-varying but not influenced by the treatment. More commonly, in the repeated cross-section case, the controls will necessarily be time varying—due to sampling variation across time periods.

We present two key assumptions, required for difference-in-differences analyses, conditional on covariates, with a special case being when $\mathbf{X}_t$ is null. We also allow for the parallel trends assumption to be stated conditional on group indicators that are less coarse than the cohort indicators. Consequently, because cohorts typically include more than one group and because the groups within a cohort may be heterogeneous, by conditioning on group indicators the assumptions are more general. Note that it is always possible to treat cohorts as "groups" in our framework.

The first assumption rules out anticipatory changes in the potential outcomes prior to the intervention occuring for each eventually treated group. We adapt the formulation from Wooldridge (2021) (also in Wooldridge, 2023) which is stated for the panel data case.

Assumption 3.1 (Conditional No Anticipation, CNA)**.** For groups $g \in 1, 2, ..., \bar{g}$ and $t \in \{1, ..., c-1\}$,

$$E\left[Y_t(g)|R_1, ..., R_G, \mathbf{X}_t\right] = E\left[Y_t\left(\infty\right)|R_1, ..., R_G, \mathbf{X}_t\right].$$

This simply means that, in any time period before the intervention occurs for any of the eventually treated groups, the potential outcomes are the same as the potential outcomes in the never treated state. This assumption can be violated if units within groups that are eventually treated anticipate the intervention and change their behavior prior to the first period of intervention.

Let $P_t \in P_1, P_2, ..., P_T$ denote binary indicators for observations in time periods $t =$

$1, 2, ..., T$. Then, in the staggered intervention case without exit, the time-varying treatment indicator is

$$W_t = \{R_g\}^q \cdot (P_q + ... + P_T) + \{R_g\}^{q+1} \cdot (P_{q+1} + ... + P_T) + ... + \{R_g\}^Q \cdot (P_Q + ... + P_T).$$

The observed outcome in every period is

$$Y_t = \{R_g\}^q \cdot Y_t(q) + \{R_g\}^{q+1} \cdot Y_t(q+1) + ...\{R_g\}^Q \cdot Y_t(Q) + \{R_g\}^\infty \cdot Y_t(\infty)$$

We state the parallel trends assumption assuming linearity of the conditional expectations and conditional on covariates and groups:

Assumption 3.2 (Conditional Parallel Trends, CPT)**.** For $t = 1, 2, ..., T$,

$$E[Y_t(\infty) | R_1, ..., R_G, \mathbf{X}_t] = \sum_{g=1}^{G} \beta_g R_g + \sum_{g=1}^{G} (R_g \cdot \mathbf{X}_t) \gamma_g + \mathbf{X}_t \pi_t + \eta_t.$$

Note that the Conditional Parallel Trends assumption in 3.2 also relaxes the analogous assumption in Wooldridge (2021) (also in Wooldridge, 2023) and other specifications where conditioning on cohort identity is replaced by conditioning on group identity, g.

Technically, we need not condition on the entire history of the covariates, $\{\mathbf{X}_t, t = 1, ..., T\}$ in assumption 3.2, and so the covariates need not satisfy a strict exogeneity assumption (see Wooldridge, 2010, Chapter 10). Nevertheless, if we think the treatment assignment influences the covariates in the future, Assumption CPT would generally fail. For a recent discussion of 'bad controls' in the DID setting, see Wooldridge (2024). Also, we require a sufficient number of observations per stratum in order to get precise estimates of the β_g and η_g in assumption 3.2.

Even if the covariates do not change over time, the terms $\mathbf{X}_t \pi_t$ and $(R_g \cdot \mathbf{X}_t) \gamma_g$ allow relaxation of the usual parallel trends assumption. Their inclusion plays the same role as in Callaway and Sant'Anna (2021), who apply standard treatment effects estimators when covariates are available; see also Wooldridge (2021, 2023). The presence of $(R_g \cdot \mathbf{X}_t) \gamma_g$ allows for substantial heterogeneity in how the average potential outcome changes with the groups (and therefore with the treatment groups).

Under assumptions 3.1 and 3.2, the parameters in 3.2 are identified using the untreated observations. In the subpopulation of untreated units at time t, we can write

$$E\left(Y_t|R_1,...,R_G,\mathbf{X}_t,W_t=0\right)=\sum_{g=1}^{G}\beta_g R_g+\sum_{g=1}^{G}\left(R_g\cdot\mathbf{X}_t\right)\gamma_g + \sum_{t=2}^{T}\eta_t P_t+\sum_{t=2}^{T}\left(P_t\cdot\mathbf{X}_t\right)\pi_t+\mathbf{X}_t\zeta_t. \tag{3.2}$$

Equation 3.2 shows that all of the parameters are identified using the untreated observations, provided we have some units in every group.

The identification argument is easier to see in the simple 2×2 case, i.e., let $T=2$ and $G=2$. Let W denote the (only) treatment indicator and, to frame it like a typical 2×2 DID specification, let α denote the intercept for the value of the outcome for $G=1$ and $T=1$. We also remove group and time subscripts from all coefficients for simplicity. Then, equation 3.2 can be written as

$$E\left[Y_1\left(\infty\right)|W,\mathbf{X}_1\right]=\alpha+\beta W+\left(W\cdot\mathbf{X}_1\right)\gamma+\mathbf{X}_1\zeta \tag{3.3}$$

$$E\left[Y_2\left(\infty\right)|W,\mathbf{X}_2\right]=\alpha+\beta W+\left(W\cdot\mathbf{X}_2\right)\gamma+\eta_2+\mathbf{X}_2\pi+\mathbf{X}_2\zeta \tag{3.4}$$

Under CNA, the expectation in equation 3.3 is the same when we replace $Y_1\left(\infty\right)$ with $Y_1\left(2\right)$. Because $W=0$ implies $Y_1=Y_1\left(\infty\right)$ and $W=1$ implies $Y_1=Y_1\left(2\right)$,

$$E\left(Y_1|W,\mathbf{X}_1\right)=\alpha+\beta W+\left(W\cdot\mathbf{X}_1\right)\gamma+\mathbf{X}_1\zeta,$$

which shows that, provided there are some treated and control units and $\mathbf{X}_1$ does not have perfectly collinear elements, the parameters α, β, γ, and ζ are identified using the control and (eventually) treated units in $t=1$. By equation 3.4,

$$E\left(Y_2|W=0,\mathbf{X}_2\right)=\alpha+\eta_2+\mathbf{X}_2\pi+\mathbf{X}_2\zeta=\left(\alpha+\eta_2\right)++\mathbf{X}_2(\pi+\zeta)$$

Again, ruling out perfect collinearity in $\mathbf{X}_2$, $(\alpha+\eta_2)$ and $(\pi+\zeta)$ are identified by the second period control units. Because α and ζ are identified, so are η_2 and π. Returning to equation 3.4, we have

$$E\left[Y_2\left(\infty\right)|W=1,\mathbf{X}_2\right]=\left(\alpha+\beta+\eta_2\right)+\mathbf{X}_2\left(\gamma+\pi+\zeta\right)$$

and so

$$E\left[Y_2\left(\infty\right)|W=1\right]=\left(\alpha+\beta+\eta_2\right)+E\left(\mathbf{X}_2|W=1\right)\left(\gamma+\pi+\zeta\right)$$

Identification of $E\left(\mathbf{X}_2|W=1\right)$ follows immediately because we observe the second-period covariates for the all units, including the treated units. Because the other parameters are identified by assumptions 3.1 and 3.2, $E\left[Y_2\left(\infty\right)|W=1\right]$ is identified. Then

$$\tau_2=E\left[Y_2\left(2\right)-Y_2\left(\infty\right)|W=1\right]=E\left(Y_2|W=1\right)-E\left[Y_2\left(\infty\right)|W=1\right]$$

is identified. The argument in the general staggered case is similar, with $E\left[Y_t\left(g\right)|W_g=1\right]=E\left(Y_t|W_g=1\right)$ always identified and $E\left[Y_t\left(\infty\right)|W_g=1\right]$ identified under assumptions 3.1 and 3.2.

3.2 Estimation by Imputation and Ordinary Least Squares

The identification argument in subsection 3.1 immediately suggests an imputation approach to estimation. We assume the availability of random samples from the population at each time period t. The draws, indicated by i, are independent, but not generally identically distributed because the population distribution of both the outcome and control variables may change across t. (This same kind of heterogeneity is allowed in panel data settings. Even if the population is stable across time, we allow for changing distributions across t.) For unit i, $t\left(i\right)$ represents the time period. The observed data for unit i is $Y_{i,t(i)}$, $\mathbf{X}_{i(t)}$, and the group indicators, $R_{i,t(i),j}$, $j=1,...,J$. As discussed earlier, the treatment assignment is determined by the group. Typically, the group for a unit does not change over time—an individual lives in the same state, say, over the time periods in question—but even if that is true, the repeated cross sections setting means that the draws of group indicators over time are from different samples of units.

Estimation of the parameters in equation 3.2 can proceed by estimating an OLS regression of the outcome on the untreated observations in the repeated cross-sectional dataset. Then, one mimics iterated expectations in the sample by using an out-of-sample prediction for $Y_t\left(\infty\right)$ (for the treated observations). Equivalently, the out-of-sample residuals are averaged over the appropriate group-time period pair to produce group-time specific ATT estimates. The following procedure extends Wooldridge (2021) to allow repeated cross sections and

time-varying covariates. It is also related to the imputation method in Borusyak et al. (2024), who mention applying imputation in the repeated cross sections case. Here, we have derived an imputation estimator for repeated cross sections under clearly stated assumptions that allow random control variables to appear in a flexible way.

Procedure 3.1. [Imputation Estimation]

1. Use the control observations to estimate the parameters

$$(\beta_1, ..., \beta_G, \eta_1, ..., \eta_G, \gamma_2, ..., \gamma_T, \pi_2, ..., \pi_T) \tag{3.5}$$

by OLS:

$$\begin{aligned} Y_{i,t(i)} \text{ on } & R_{i,t(i),1}, ..., R_{i,t(i),G},\ P_{2,t(i)}, ..., P_{T,t(i)},\ \mathbf{X}_{i,t(i)} \\ & R_{i,t(i),1} \cdot \mathbf{X}_{i,t(i)}, ..., R_{i,t(i),G} \cdot \mathbf{X}_{i,t(i)},\ P_{2,t(i)} \cdot \mathbf{X}_{i,t(i)}, ..., P_{T,t(i)} \cdot \mathbf{X}_{i,t(i)} \end{aligned} \tag{3.6}$$

2. For unit i, impute $Y_{i,t(i)}(\infty)$ as

$$\begin{aligned} \hat{Y}_{i,t(i)}(\infty) = & \sum_{g=1}^{G} \hat{\beta}_g R_{i,t(i),g} + \sum_{g=1}^{G} \left(R_{i,t(i),g} \cdot \mathbf{X}_{i,t(i)}\right) \hat{\gamma}_g \\ & + \sum_{t=2}^{T} \hat{\eta}_t P_{t(i)} + \sum_{t=2}^{T} \left(P_{t(i)} \cdot \mathbf{X}_{i,t(i)}\right) \hat{\pi}_t + \mathbf{X}_{i,t(i)} \hat{\zeta} \end{aligned} \tag{3.7}$$

3. For treatment group g, in period t, obtain

$$\begin{aligned} \hat{\tau}_{gt} &= N_{gt}^{-1} \sum_{i=1}^{N} R_{i,t(i),g} \cdot 1[t(i) = t] \cdot \left[Y_{i,t(i)}(g) - \hat{Y}_{i,t(i)}(\infty)\right] \\ &\equiv N_{gt}^{-1} \sum_{i=1}^{N} R_{i,t(i),g} \cdot 1[t(i) = t] \cdot \widehat{TE}_{i,t(i)} \\ &= \bar{Y}_{gt} - N_{gt}^{-1} \sum_{i=1}^{N} R_{i,t(i),c} \cdot 1[t(i) = t] \cdot \hat{Y}_{i,t(i)}(\infty) \end{aligned} \tag{3.8}$$

where

$$N_{gt}^{-1} = \sum_{i=1}^{N} Q_{i,t(i),c} \cdot 1[t(i) = t]$$

is the number of units in treatment group g in time period t, $\widehat{TE}_{i,t(i)} = Y_{i,t(i)} - \hat{Y}_{i,t(i)}(\infty)$ is the unit-specific estimated treatment effect, and

$$\bar{Y}_{gt} = N_{gt}^{-1} \sum_{i=1}^{N} Q_{i,t(i),g} \cdot 1[t(i) = t] \cdot Y_{i,t(i)} \tag{3.9}$$

is the average of the observed outcomes for units in treatment group g in period t. □

With a sufficient number of observations in each (g,t) cell, $\hat{\tau}_{gt}$, will have good statistical properties by the law of large numbers and central limit theorem. Nevertheless, the multi-step nature of the estimation makes inference challenging. A similar issue arises in the panel data setting in Borusyak et al. (2024), where unit-specific fixed effects are included in the first imputation step. Fortunately, there is an algebraic equivalence between imputation and a longer regression that uses all of the data. To describe the longer regression, let $\bar{\mathbf{X}}_{gt}$ be the average value of the covariates for treatment group g and time period t. Specifically, to the list of regressors in equation 3.6 we add treatment indicators and treatment indicators interacted with demeaned covariates:

$$R_{i,t(i),g} \cdot P_{t(i)},\ R_{i,t(i),g} \cdot P_{t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right),\ t = q, ..., T;\ \ g = 1, 2, ..., \bar{g} \tag{3.10}$$

If $R_{i,t(i),c} \cdot P_{t(i)} = 1$ then unit i is in receiving treatment in time period t. The interactions $R_{i,t(i),c} \cdot P_{t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right)$ allow for heterogeneity in the treatment effects as a function of the observed covariates.

Proposition 3.1. [Equivalence of Imputation and OLS regression]

Using all of the data, consider the regression that includes all regressors in equations 3.6 and 3.10:

$$\begin{aligned}
Y_{i,t(i)} \text{ on } & \mathbf{X}_{i,t(i)}, \\
& R_{i,t(i),1},\ ...,\ R_{i,t(i),G},\ R_{i,t(i),1} \cdot \mathbf{X}_{i,t(i)},\ ...,\ R_{i,t(i),G} \cdot \mathbf{X}_{i,t(i)}, \\
& P_{2,t(i)},\ ..., P_{T,t(i)},\ P_{2,t(i)} \cdot \mathbf{X}_{i,t(i)},\ ...,\ P_{T,t(i)} \cdot \mathbf{X}_{i,t(i)} \\
& R_{i,t(i),q} \cdot P_{q,t(i)},\ ...,\ R_{i,t(i),q} \cdot P_{Q,t(i)},\ ...,R_{i,t(i),Q} \cdot P_{Q,t(i)}, \\
& R_{i,t(i),q} \cdot P_{q,t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right),\ ...,\ R_{i,t(i),Q} \cdot P_{Q,t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right)
\end{aligned}$$

Let the coefficients be $\left\{\tilde{\beta}_g : j = 1, ..., G\right\}$, $\{\tilde{\eta}_g : j = 1, ..., G\}$, $\{\tilde{\gamma}_t : t = 2, ..., T\}$, $\{\tilde{\pi}_r : t = 2, ..., T\}$, $\{\tilde{\tau}_{gt} : t = c, ..., T; g = 1, ..., \bar{g}\}$, and $\left\{\tilde{\delta}_{gt} : t = c, ..., T; g = 1, ..., \bar{g}\right\}$. Then

(i) For all g and t, $\tilde{\beta}_g = \hat{\beta}_g$, $\tilde{\eta}_g = \hat{\eta}_g$, $\tilde{\gamma}_t = \hat{\gamma}_t$, and $\tilde{\pi}_t = \hat{\pi}_t$.

(ii) For all $g \in \{1, 2, ..., \bar{g}\}$ and $t \in \{c, ..., T\}$,

$$\tilde{\tau}_{gt} = \hat{\tau}_{gt} \qquad \square$$

The equivalences in Proposition 3.1 are practically useful. Standard errors for the treatment effects $\tilde{\tau}_{gt} = \hat{\tau}_{gt}$ from the regression in equation 3.1 are readily available, and issues of clustering can be resolved in a standard pooled OLS setting. In addition to providing the $\hat{\tau}_{gt}$ and their standard errors, the $\tilde{\delta}_{gt}$ can be studied to determine if there are heterogeneous treatment effects.

As a special case of the regression in equation 3.1, one might have a single binary indicator, X_{it} (probably not time-varying), separating units into one of two groups. The scalar coefficients $\tilde{\delta}_{gt}$ would be the difference-in-difference-in-differences (DIDID) estimator for the group represented by $X_{it} = 1$. Again, inference is straightforward.

4 Regression Specifications

As the literature on methods for estimation of models for data with staggered entry into treatment has grown rapidly, the nomenclature used to describe various techniques has also proliferated. In naming methods, we think it is a) important to distinguish between the estimands and the estimation methods, and b) to name the various estimators recently developed for estimating models in ways that indicate functional distinctions. In terms of the characteristics of the treatment effects parameters, some estimators assume homogeneous effects across groups and time. Others assume heterogeneity in one or the other dimension but not both. Yet others allow for heterogeneous effects across both cohorts and time but not across groups and time.

On the dimension of the parallel trends assumption, some estimators impose the parallel trends assumption for all periods prior to treatment, i.e., they assume that the regression specifications have only lagged treatment parameters. Other estimators impose the parallel trends assumption only for one baseline period prior to treatment (typically the period just preceding treatment) and the regression specifications include lag and lead treatment parameters with the lag parameters being the coefficients of interest for the ATET. Such models are often referred to as *event-study* models but that terminology corrupts definitions of event-study models from other areas of the econometrics literature.

We clarify our own use of language and notation by formally specifying the regression specifications we estimate using OLS to implement equation 3.1. In equation 4.1 below,

we present a FLEX specification that explicitly displays all the regression parameters (and associated variables). The coefficients in the line denoted *lags* are the treatment effects at the group-time level in post-treatment periods. The coefficients in the line denoted *leads* are the pre-treatment differences between treated groups and the never-treated groups, except in one "baseline" period (chosen, without loss of generality, to be the period prior to treatment initiation, $q-1$). When the parameters in *leads* are estimated, we refer to these as *lags and leads* specifications. This specification is commonly referred to as an *event study* in the recent literature, (e.g. Roth, 2024), but we prefer the term *lags and leads* to disambiguate from an older, distinct use of the term "event study" in the econometrics literature (e.g. MacKinlay, 1997). In *lags only* specifications, all *lead* coefficients are set equal to zero. In other words, *lags only* specifications assume that all pre-treatment differences between treated cohorts and the never-treated cohort are identically equal to zero. These specifications include group and year fixed effects. These specifications also allow covariates to enter additively and interacted with each lag, lead, and fixed effect indicators.

$$
\begin{aligned}
&E\left(Y_{it}|\{R_{ig}\},\{P_{it}\},\boldsymbol{X}_{it}\right) = \\
&\quad\sum_{g=1}^{\bar{g}}\sum_{t=q}^{T}\tau_{gt}R_{ig}P_{it} + \sum_{g=1}^{\bar{g}}\sum_{t=q}^{T}R_{ig}P_{it}\cdot(\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_{gt})\boldsymbol{\kappa}_{gt} && \textit{lags} \\
&+\sum_{g=1}^{\bar{g}}\sum_{t=1}^{q-2}\tau_{gt}R_{ig}P_{it} + \sum_{g=1}^{\bar{g}}\sum_{t=1}^{q-2}R_{ig}P_{it}\cdot(\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_{gt})\boldsymbol{\kappa}_{gt} && \textit{leads} \\
&+\sum_{g=1}^{G}\beta_g R_{ig} + \sum_{t=2}^{T}\eta_t P_{it} + \sum_{g=1}^{G}R_{ig}\cdot\boldsymbol{X}_{it}\boldsymbol{\gamma}_g + \sum_{t=1}^{T}P_{it}\cdot\boldsymbol{X}_{it}\boldsymbol{\pi}_t + \boldsymbol{X}_{it}\boldsymbol{\zeta} && (4.1)
\end{aligned}
$$

Note that when covariates are interacted with treatment indicators, they are specified as deviations from group means. However, when they are interacted with group and time indicators, the raw values are used.

Generally, the *lags and leads* specifications can be expected to be less efficient than *lags only*, because the latter uses all implications of the conditional parallel trends assumption. Moreover, *lags and leads* specifications may have more resiliency against some violations of parallel trends, but they will be more sensitive to other violations, e.g., they will be especially sensitive to violations of parallel trends that occur just before the intervention.

4.1 Aggregating the Treatment Effects

Rather than report a full set of treatment effect estimates for each treatment group and year, it is common to aggregate the effects to the cohort level, to the time level, to the exposure-time level (for staggered start), or most commonly to the aggregate level. One could do this by simple averaging. For example, the immediate effects, $\hat{\tau}_{gt}$, can be averaged over time periods $t = q$, ..., T. The one-period dynamic effects, $\hat{\tau}_{t,t+1}$, can be averaged over $t = q$, ..., $T-1$; and so on for each of the exposure lengths. To be precise, the aggregated average treatment effect on the treated (ATET) corresponding to equation 4.1 is given by:

$$\hat{\tau} = \frac{1}{N} \sum_{g=1}^{\bar{g}} \sum_{t=q}^{T} \sum_{i=1}^{n_{gt}} \left[\hat{\tau}_{gt} \left(R_{ig} P_{it} \right) + \left(R_{ig} P_{it} \cdot \left(\boldsymbol{X}_{it} - \bar{\boldsymbol{X}}_{gt} \right) \right) \hat{\boldsymbol{\kappa}}_{gt} \right] \tag{4.2}$$

where n_{gt} is the number of observations in the g^{th} treated group in the t^{th} time period and $N = \sum_{g=1}^{\bar{g}} \sum_{t=q}^{T} n_{gt}$. This simplifies to

$$\hat{\tau} = \frac{1}{N} \sum_{g=1}^{\bar{g}} \sum_{t=q}^{T} \left[n_{gt} \hat{\tau}_{gt} \right] \tag{4.3}$$

because $\bar{x}_{gt}$ is de-meaned so that part of the equation sums to zero.

Note that $\hat{\tau}$ is simply the average of the treatment effects over all treated observations in all treated time periods. In specifications with no covariates, or covariates entered only additively, or if the covariates are de-meaned (using the means of covariates for the treated sample), then $\hat{\tau}$ is also a weighted average of the group-time effects $\hat{\tau}_{gt}$, where each weight is the sample size associated with each group and year, n_{gt}. Standard errors for the unweighted or weight averages are easy to obtain using standard software packages.

4.2 Other Approaches

As we mentioned previously, the two-step imputation method of Borusyak et al. (2024) using the group dummies and the OLS regression on the group dummies using all of the data produce numerically identical results. Moreover, if we include the pre-treatment indicators then, without controls, we obtain a repeated cross-sections version of the Sun and Abraham (2021) leads and lags event-study estimator that allows full heterogeneity by group and calendar time.

There are some other approaches that have become popular in empirical research. Callaway and Sant'Anna (2021) propose what are effectively leads and lag estimators because their methods reduce to estimating many 2×2 DiDs using the never treated group as the control group and the period just prior to the intervention as the control period. Without covariates, the Callaway and Sant'Anna (2021) approach is identical to Sun and Abraham (2021). With covariates, Callaway and Sant'Anna (2021) implement a regression adjustment approach to estimation of the treatment effects.

A novel aspect of Callaway and Sant'Anna (2021) is that, with covariates, they allow estimation methods other than linear regression adjustment. One estimator with nice robustness properties combines propensity score weighting with linear regression adjustment (a version of IPWRA). Without the propensity score and without covariates, the Callaway and Sant'Anna (2021) approach is the same as FLEX without covariates. When covariates are introduced, Callaway and Sant'Anna (2021) approach is equivalent to a fully saturated FLEX specification. Note that the IPWRA estimator can have some resiliency to the assumed linear functional form.

Another popular approach uses a stacked difference-in-differences procedure (Stacked DID) described in Cengiz et al. (2019) and Wing et al. (2024). In each, the observations for each treated cohort are pooled with the observations from the never-treated cohort to form a data set that are then appended to form one stack. Note that this includes separate time and cohort indicators for each cohort. By doing so, a constant treatment effect difference-in-differences regression specification with group and time fixed effects delivers unbiased estimates of the cohort-level treatment effect. The issues reported by Goodman-Bacon (2021) and others do not apply. Then each of these stacks is pooled and a regression that interacts each term in the standard regression specification with stack indicators produces all cohort-by-time treatment effects in one application of a regression procedure. The same never-treated observations are used multiple times (as many times as there are treated cohorts). If there are many untreated groups, then the sample size for Stacked DID can be very large which can substantially increase the time to estimate the parameters.

4.3 Heterogeneous Time Trends

Like the estimates obtained from procedure 3.1, the event-study estimates require a kind of parallel trends assumption for consistency (Roth, 2022; Dette and Schumann, 2024). One approach to accounting for violations of parallel trends that occur even after including controls is to assume relatively simple heterogeneous trends in the absence of the intervention. In particular, with at least two pre-intervention periods per treated group, one can include in equation 3.1 interactions between the group indicators, $R_{i,t(i),j}$ and a linear time trend, t. The coefficients on the trend terms can be used to test for pre-trends (Dette and Schumann, 2024). As shown in Wooldridge (2021) in the panel data case, such tests do not suffer from 'contamination bias', provided the covariates are included flexibly, as in equation 3.1. That is because the imputation and OLS approaches continue to be identical.

Including group-specific trends can be costly in terms of precision because their inclusion creates collinearity with the treatment indicators. Of course, using a pre-test to decide whether to drop these terms can be problematic—just as when using the event-study approach.

4.4 Practical Considerations

In this section we discuss several practical considerations when estimating difference-in-differences models for data with repeated cross-sections, heterogeneous treatment effects, and staggered timing. Estimating treatment effects in a difference-in-differences model is usually a two-step process. The first step is to estimate the treatment parameters of a model and the second step is to aggregate some of those treatment parameters into estimated average treatment effects. For example, estimating the average treatment effect on the treated, or treatment effects by group or by time since the start of treatment. Some DID software packages seamlessly combine these two steps into one, showing only a final average treatment effect on the treated but not the intermediate step of estimating coefficients from a regression model. Thinking about DID as a two-step process is useful because it clarifies that there are two steps to decision making by the researcher, each with its own set of questions.

The set of questions for the first step revolves around identification of treatment effects and model specification. How are the treatment effects identified? What level of hetero-

geneity is allowed, specifically, heterogeneity at the group or the cohort level? Should the model assume that parallel trends holds in the pre period or allow for heterogeneous effects across groups in each time period before treatment starts? To what extent should the model specification control for covariates?

The set of questions for the second step includes answering the research question, presenting the results, and describing the extent of the heterogeneity in the treatment effects. After estimating the regression model, how should the estimated treatment coefficients be combined to form an overall ATET? Should the treatment effects be aggregated to other levels, for example, to show an event-study graph?

4.4.1 Estimation step

Our proposed FLEX approach also provides a useful modeling framework for all difference-in-differences models. Our general approach to estimation is that OLS can be used to estimate any flexible model, whether it is a flexible model with heterogeneous treatment effects or a parsimonious model with homogeneous treatment effects. Treatment effects can be heterogeneous with respect to time, group (or cohort), or both, and there can additionally be heterogeneous treatment effects by covariates. Most existing methods can be estimated with an OLS regression in the FLEX framework.

In our empirical examples, we explain when FLEX estimates the same treatment coefficients as other estimators.

4.4.2 Basic modeling decisions

There are three main modeling decisions, which can be seen as whether to allow more or less flexibility in the estimation step. They can also be seen as variations on equations 4.1. The first decision is whether to allow heterogeneity at the group level or the cohort level, when there are multiple groups per cohort. In our empirical examples, there are multiple states (groups) that start the policy treatment in some years and so are in the same cohort. In general we think that it is restrictive to impose homogeneity across all states that happen to start their policy in the same year. Therefore, our model specifications are at the group level. However, if the number of groups is large (e.g., $3,000$ counties), then one needs to

think harder about the tradeoffs between model flexibility and possible over-fitting.

A second decision is whether to model lags only or both lags and leads. We classify treatment effects as being either lags (after the start of treatment) or lags and leads (including allowing for separate effects of treated groups prior to treatment). The event-study approach, which allows for different effects for treatment observations in the pre-treatment periods than the control observations, is most flexible because it does not impose the conditional parallel trends assumptions in the pre-treatment periods. Alternatively, one can impose those assumptions and focus the modeling of heterogeneity to the time periods after the start of treatment. A lags-without-leads model specification would assume that all coefficients in the second rows of equation 4.1 are equal to zero.

The third is about how to include covariates. Covariates could be included separately as additional controls or interacted with treatment effects. Interacting covariates with treatment effects allows for further heterogeneity. This is a decision about whether to include the third and fourth rows of equation 4.1 or to assume that some of those coefficients are zero. The potential downside of this flexibility includes over-fitting. Alternatively, one can lean on economic theory and knowledge of institutions to focus on just one or a few covariates to interact with the treatment effects instead of all covariates. In our experience, including interactions between treatment effects and covariates does not lead to larger standard errors.

There are two main reasons not to control for covariates. One is if those covariates are endogenous. For example, if the treatment is a job training program and a covariate measures job skills, then the covariate is endogenous to the treatment. The other is if the covariates are time invariant and at the group level. Such covariates would be swept up in group fixed effects mechanically.

In the flexible specifications of the regressions, group-by-year treatment effects vary by individual- and group-specific covariates. This is accomplished by interacting the group-by-year indicator variables with the covariates. To interpret the coefficients on the group-by-year indicators as the treatment effects in such models, it is necessary to demean the covariates in the interaction terms. Demeaning requires subtracting the group-by-year means of each covariate from its individual-level values. In these flexible specifications, the group and year fixed effects are also interacted with the covariates. But for these indicators,

the interactions are with the raw covariates, not the demeaned ones. In addition, the raw covariates themselves are entered into the regression specification additively. We want to emphasize that it is not necessary to de-mean the covariates to estimate the ATET. As long as one is careful in combining the weighted average of the coefficients for the de-meaned variables, the final ATET is identical.

4.4.3 Aggregation and graphing

After estimating treatment effects at the group-time level, researchers have several options for how to present the results. One common approach is to aggregate to the time-since-treatment level to create an event study plot. This is useful when estimating a lags and leads model because one can plot the estimated difference between ever-treated groups and control groups in each pre period. This provides a visual test of the parallel trends and no anticipation assumptions. If one suspects or wants to test for heterogeneity by calendar time or by group, one could aggregate treatment effects to those levels and plot the results.

Researchers often want an overall ATET. This is easy to do in the FLEX approach. The ATET is the weighted average of treatment effects, where the weights are the number of treated observations in each group-time set in the post periods. We provide Stata code to show how to do this after estimating the single regression model, including standard errors.

4.4.4 Collapsing the data to the group level

When there are many individual observations for each group (or cohort), it is worth considering whether to collapse the data to the group (or cohort) level. The advantage is that collapsed data run faster in weighted OLS while returning identical parameter estimates and standard errors. If the data set is large, collapsed data could run considerably faster than the original data set. However, if there are covariates, then collapsing the data may not be feasible. Another case where one might consider collapsing the data is when there are a large number of groups (e.g., all U.S. counties) that require a very large number of parameters to estimate.

The identity (identical covariates from WOLS with collapsed data to covariates from OLS with non-collapsed data) only holds if you collapse not to the group-time level, but

for every possible combination of covariates for each group-time combination. Therefore, the covariates must be discrete. For example, if there are 50 groups and 10 years of data, then there are 500 group-time combinations. Now suppose there are three dummy variables, which can be combined eight ways, that means that the data would be collapsed to 4000 possible group-time-covariate combinations. That could still be significantly smaller than the original data set and faster to run, but the potential benefits of collapsing are small when there are many covariates. When there are continuous covariates, one can still collapse the data by specifying discrete versions of these variables.

5 Empirical examples

We show two empirical examples using two different data sets that have features typical of repeated cross-section data with a difference-in-differences study design. One measures the effect of punitive substance use policies for expecting mothers on their mental health outcomes. The other tests the effect of right-to-work laws on hourly earnings. For both examples, the data are collected at the individual level, with individuals grouped within states. The treatments are at the state level and the timing of the start of treatments is staggered across several years. Although there are multiple years of data, these are not panel data sets but instead are repeated cross sections. All covariates (other than state and year fixed effects) are at the individual level. Therefore, both examples are typical of many difference-in-differences study designs and are appropriate to illustrate our theoretical results.

5.1 Prior empirical literature

There is a large and rapidly expanding empirical literature that uses difference-in-differences study design to assess the effects of policy changes. Here we briefly mention just a few that are most relevant to our two empirical examples. These are all repeated cross sections with staggered treatment timing, where the data and outcomes are at the individual level and the treatment is at the state level.

Meinhofer et al. (2022) study whether punitive or priority prenatal substance use policies affect neonatal drug withdrawal syndrome, other birth outcomes, and the use of prenatal

care. They use data from the Healthcare Cost and Utilization Project for 46 states from 2008–2018. The repeated cross-section data are at the state-year level, with weights equal to the number of births in the state-year. Some states adopted either punitive or priority prenatal substance use policies during the study period. They estimated several different difference-in-differences models, including basic two-way fixed effects, Sun and Abraham (2021) event study, and Callaway and Sant'Anna (2021). Meinhofer et al. (2022) argue that punitive prenatal substance use policies may exacerbate health problems, while priority policies may alleviate them. That is exactly what they find, with babies in punitive states having significantly worse neonatal drug withdrawal syndrome and babies in priority states having slight reductions in the probability of low gestational age and low birth weight.

A second set of papers examine whether right-to-work laws affect wages, unionization, and individual well-being. The National Labor Relations Act of 1935 allowed private-sector workers to unionize and collectively bargain with employers (Makridis, 2019; Fortin et al., 2023; Wexler, 2022). Furthermore, the National Labor Relations Act required that every worker covered by the contract must pay dues to the union. The Taft-Hartley Act of 1947 allowed this federal National Labor Relations Act to be replaced by state laws. States could therefore pass right-to-work Laws, which changed one important feature of unions. Workers covered by the contract no longer had to pay union dues. By the end of the 1940s, twelve states had adopted right-to-work Laws. There were seven more in the 1950s and 1960s, with six more since 2000. The effect of right-to-work Laws has been shown to diminish paid union membership over time, but the effects on employment and wages are mixed (Makridis, 2019; Fortin et al., 2023; Wexler, 2022).

Fortin et al. (2023) primarily use Current Population Survey data on individuals from 2003–2019 to compare unionization rates and wages. Their identification comes from changes in right-to-work laws in six states over their study period. They estimate event study and two-way fixed effects models, arguing that in their case the problems from TWFE should be small. They find that right-to-work laws reduce both wages and unionization rates. Makridis (2019) uses Gallop poll data and a similar study design to show that individual well-being improved slightly in states that adopted right-to-work laws.

5.2 Prenatal substance use policies and BRFSS data

For an illustrative example of the difference-in-differences methods to measure the effect of prenatal substance use policies, we use the Behavioral Risk Factor Surveillance System (BRFSS) data from 2005–2018. The BRFSS is an annual survey by the Centers for Disease Control and Prevention of about 400,000 adults in the United States about their risk behaviors, chronic health conditions, and use of preventive services. The representative sample of women of child-bearing age (18–44) with at least one child from 34 states has 440,446 observations.

There is considerable variation in the timing of the start of the policy at the state level. Between 2007 and 2018, 13 states enacted prenatal substance use policies. Idaho was the first in 2007, then South Carolina followed in 2008 and Arizona in 2009. The remaining states started their policies in 2012 (UT), 2013 (AL), 2014 (GA, MS, RI, TN), 2015 (NM), 2017 (CT, OH), 2018 (KY). There are 21 states in the control group. The full list of treated and untreated states and the year that prenatal substance use policies began are shown in Table 1.

The outcome of interest is the number of days that the mother has spent in good mental health in the last month. It is plausible that the punitive prenatal substance use policies could adversely affect a person's mental health. This outcome variable has a mean of about 25.5, meaning that most people are in good mental health most days. For illustrative purposes, we control for a limited set of individual demographic characteristics. These include age, race and ethnicity, educational attainment, and income. The mean age of the mothers in the sample is about 34 years old. We limited the sample to those aged 18–44 to focus on mothers of child-bearing age. For the full table of descriptive statistics, see Table 2.

5.3 Right-to-work laws and CPS data

We use the Current Population Survey (CPS) to show an example of how to use difference-in-differences methods to estimate the effect of state right-to-work laws on hourly earnings. The U.S. Bureau of Labor Statistics fields the CPS to gather information about labor force participation, wages, and demographics. Although the CPS collects data monthly, we use the annual survey from 2008–2019. The representative sample of adults who have education

beyond high school (an associates degree, a bachelor's degree, or an advanced degree) from 29 states has 973,578 observations.

Out of the 29 states in our sample, five states changed the law to become a right-to-work state and 24 did not. The states that had passed right-to-work laws prior to 2002 are excluded from our sample because they are treated before the start of our data collection. In 2012, Indiana and Michigan became right-to-work states. Later, Wisconsin, West Virginia, and Kentucky passed right-to-work laws in 2015, 2016, and 2017, respectively. Although the amount of variation in treatment is not enormous, it is more than sufficient to illustrate our points and estimate treatment effects. The list of treated states, the year the right-to-work law began, and the list of untreated states are shown in Table 3.

We estimate the effect of a change in the right-to-work laws on hourly earnings for those with positive earnings, which ranges from nearly \$0 to over \$2,076 per hour. It is plausible that right-to-work laws would directly affect earnings if it weakens unions by reducing their revenue from dues. Among this sample of persons with non-zero reported earnings, the median is \$20 and the mean is \$25. We have a limited set of covariates, including age, gender, race and ethnicity, education, and marital status. The average age of persons in this adult non-elderly sample is 44 years old. About 50 percent are women and 61 percent are married. For the full table of descriptive statistics, see Table 4.

5.4 Model specifications

We estimate several alternative FLEX specifications based on the general model specification in equation 4.1 for each of the two samples of repeated cross-sectional observations. Our preferred FLEX specifications, which are equivalent to imputation estimators, estimate heterogeneous treatment effects over either groups and event time or over cohorts and event time. In these examples, groups are defined by states because the policy-relevant laws are at the state level and the treatment effects are plausibly heterogeneous by state. However, some states changed their laws in the same calendar year as other states. When multiple states begin treatment at the same time, those states (groups) are in the same cohort. The first example (prenatal substance use policies) has four states in the 2014 cohort, two in the 2017 cohort, and one each in seven other cohort years. Therefore, in this example, it likely

matters whether to allow heterogeneous treatment effects at the group level or only at the cohort level. In the right-to-work example, all but one cohort have a single state, perhaps making that distinction less important empirically.

The *lags-only* models estimate treatment effects only in the periods after the start of the policies, i.e., these specifications use indicators for treatment lags only. This means that there are no separate coefficients for treatment and control groups in each pre-treatment period. In *lags and leads* models, we allow all periods—except one reference baseline period, which we specify as the period preceding initiation of treatment—to have treatment indicators for each ever-treated group or cohort. There are cohort-by-time or group-by-time coefficients in the pre-treatment periods as well as the post-treatment periods. In the *lags and leads* models, the treatment effects are compared to the year prior to the start of treatment; in the *lags-only* models, the treatment effects are compared to the average of all the years prior to the start of treatment.

We made three different choices about how to include covariates for individual-level characteristics. As a reminder, because the data are repeated cross-sections, although the sample populations are reasonably stable, the exact mean of individual-level variables for any state will change slightly from year to year. We estimate specifications that exclude covariates or include covariates in a flexible way by interacting all covariates with each lead and lag treatment coefficient, each group (or cohort) fixed effect, and each time fixed effect.

We show the results for both empirical examples in tabular (Tables 5 and 6) and graphical form (Figures 1 and 2). The tables list the average treatment effect on the treated (ATET) for a variety of model specifications. The top panel of each table shows results for *lags-only* models, while the lower panel of each table shows results for *lags and leads* models. We compare our FLEX model specifications to several other model specifications, which differ in the estimand, estimator, sample, model specification, and whether and how covariates are included. The estimand for treatment effects varies widely across the models. The simplest assumes that the treatment effects are constant (homogeneous), both across groups and over time. For the *lags-only* models, this specification is commonly referred to as the homogeneous two-way fixed effects (TWFE) regression. It assumes a one-time homogeneous shift in the outcome due to treatment. Note that the use of TWFE to refer to the homogeneous effects

specification is misleading because it does not distinguish among a variety of estimands, each of which involve regression specifications that include fixed effects along two dimensions, all of which are more general than the homogeneous effect specification. For the *lags and leads* models, the simplest version assumes a separate effect for each point in time since the event, as described in Sun and Abraham (2021), but is homogeneous across groups (see the rows labeled *Event time ES* for the treatment heterogeneity in the lags and leads models).

We also estimate treatment effects using two popular alternative techniques, each of which has a way to resolve the issues arising from staggered entry into treatment. One approach uses stacked samples of data, where the observations for each cohort are first paired with all never-treated control observations, and then the cohort-specific datasets are pooled (Cengiz et al., 2019; Wing et al., 2024). These *stacked data* regressions can be used to estimate models with *lags-only* or *lags and leads* specifications of treatment coefficients. Another *lags and leads* model uses the method of Callaway and Sant'Anna (2021) which implements a flexible regression adjustment estimator.

6 Results

6.1 Prenatal substance use policies

The outcome variable in the analysis of punitive prenatal substance use policies is mental health status defined by the number of good mental health days in a month. The results in Table 5 show that the ATET estimates are always negative, indicating that punitive prenatal substance use policies lead to worse mental health for women. The FLEX estimates are that punitive prenatal substance use policies decreases the number of good mental health days in a month by about $0.15 - 0.20$ days. Given that the average number of bad mental health days among women in never-treated states is 4.3 (the mean number of good days is 25.7), the estimate implies at least a 3.5% increase in the number of bad mental health days due to the change in policy.

The estimates are negative but not statistically significant when the homogenous (constant) specification (TWFE) is used. The estimates are also not statistically significant in a lags and leads specification that allows for heterogeneous estimates in event-time but homo-

geneous across groups. Also, in both lags only and lags and leads specifications, the stacked data regressions, which allow for heterogeneous effects across event-time, produce estimates that are not statistically significant. These results do not qualitatively differ in specifications with and without covariates.

In more general regressions specifications in which treatment varies by cohort and year or by group and year, the ATET estimates are negative and statistically significant at all conventional levels. The point estimates are a bit larger in the cohort-by-year specifications as compared to those in the group-by-year specifications. But the standard errors in the group-by-year specification are consistently smaller than those in the cohort-by-year specifications. These results can be seen in lags only and lags and leads specifications. It appears that the additional generality implied in the group-by-year heterogeneous specifications produces treatment estimates with greater precision because there is substantial within-cohort (by year) heterogeneity in outcomes.

The Callaway and Sant'Anna estimates, which allow for general heterogeneity at the cohort-by-year level, are similar to those obtained using OLS. These estimates are also statistically significant but the standard errors are more than double the size of those obtained in the group-by-year linear regression specifications estimated using OLS.

Another interesting pattern is that the ATET shrinks towards zero when covariates are interacted with the heterogeneous treatment effects. While this is not a universal finding, we have noticed in numerous empirical examples that including covariates often changes the magnitude of the ATET, while the standard errors either stay about the same or shrink.

We use our preferred FLEX specification, a linear-in-parameters model with heterogeneous treatment coefficients at the group-by-year level and interacted with covariates, to calculate estimates of ATET at disaggregate levels of interest to researchers. In panels (a) and (b) of Figure 1, we show the event study plots, from *lags only* models in panel (a) and from *lags and leads* models in panel (b). The latter is what is often referred to as the *event study* figure. An eyeball check shows that the parallel trends assumption is good for at least four pre-periods. The negative treatment effects by exposure year in the treatment periods are quite similar in magnitude across the *lags only* specification and the *lags and leads* specification. The ATET estimates in each calendar year in a selection of treated periods are

shown in panels (c) and (d). These are similar across the *lags only* specification (c) and *lags and leads* specification (d). The ATET by cohorts are shown in panels (e) and (f). They show substantial heterogeneity, with the early cohorts generally having negative effects and the later cohorts having more positive effects. This heterogeneity would be worth exploring further.

6.2 Right-to-work laws

The ATET of right-to-work laws on hourly earnings is negative across all our specifications and methods (Table 6). Picking a typical value of the ATET of $-.40$ (forty cents), this means that hourly earnings fell by about 1.6% after the start of right-to-work laws were passed, relative to the average earnings in never treated states of about $25.

The results show that the estimates of ATET from the homogeneous effects specification are quite close to those obtained from specifications that allow for heterogeneous effects at the event-time level and those that allow for cohort-by-year or group-by-year heterogeneity. But the standard errors in the group-by-year specification are half the size of those in the other specifications. These results can be seen in lags only and lags and leads specifications. It appears that the additional generality implied in the group-by-year heterogeneous specifications produces treatment estimates with greater precision because there is substantial within-cohort (by year) heterogeneity in outcomes.

The Callaway and Sant'Anna estimates, which allow for quite general heterogeneity at the cohort-by-year level, are also similar to those obtained using regression specifications that allow for cohort-by-year heterogeneity. These estimates are also statistically significant but the standard errors are about twice as large as their OLS counterparts, and more than double the size of those obtained in the group-by-year linear regression specifications estimated using OLS.

Once again use our preferred FLEX specification, a linear-in-parameters model with heterogeneous treatment coefficients at the group-by-year level and interacted with covariates, to calculate estimates of ATET at disaggregate levels of interest to researchers. On the left side of Figure 2, we show estimates from *lags only* specifications while on the right side we show estimates from *lags and leads* models. As in the case using data from the

BRFSS, the ATET estimates at the cohort-level are different across specifications. Notably, the event study plots shown in panels (a) and (b) show that the effects in the treatment periods are quite similar across the *lags only* specification and the *lags and leads* specification even though there appears to be evidence of pre-program effects. The ATET estimates in each calendar year in a selection of treated periods are remarkably similar across the *lags only* specification and *lags and leads* specification (see panels (c) and (d)). Yet, there is considerable heterogeneity across cohorts, from close to -1 to insignificantly different than zero.

6.3 Transparency

One advantage of our FLEX approach is that it is transparent. The model specification is clear; there are no hidden estimated parameters. This transparency is useful for comparing different possible FLEX model specifications and for comparing FLEX with other estimators. To demonstrate this transparency, we created tables showing the number of parameters estimated across different model specifications and estimators in the first stage of DID, before any aggregation of the treatment effects. The number of parameters estimated for the punitive prenatal substance use policies example are shown in Table 7, and the number of parameters for the for the right-to-work example are in Table 8. The numbers of parameters are shown separately for the main treatment effects (lags and leads), main fixed effects, and interactions of the main effects with covariates. All model specifications include main fixed effects for the group and time periods (what is often called two-way fixed effects). The total number of parameters also includes an intercept and parameters for each of the covariates (7 covariates for the first example and 9 for the second example).

There are several important patterns apparent in Tables 7 and 8. By definition, there are no lead effects estimated in the lags-only models. The lags-only models assume that the parallel trends assumption is correct in the pre-periods; those models do not allow separate parameters for treatment and control groups each pre period. This assumption, if correct, is more efficient. However, models with both lags and leads allow one to plot event study graphs and test whether the parallel trends and no anticipation assumptions are correct.

Another difference between models is whether the model uses groups or cohorts. Because

cohorts are collections of at least one group, there are at least as many parameters in the group models as in the cohort models. The interactions in the FLEX models are between covariates and treatment effects. Despite adding several hundred additional covariates, in our experience this additional treatment effect heterogeneity does not generally increase the standard errors of the estimated average treatment effect on the treated. Also, the researcher has the option to only interact a small number of the covariates with the treatment effects, based on what economic theory predicts.

Comparing the parameters helps to show the relationship between the different models. One interesting comparison is that the estimated coefficients for two of the models are identical. The lags-and-leads FLEX model with cohort and time effect heterogeneity and no covariates interacted is identical to the Callaway and Sant'Anna model with cohort and time effect heterogeneity and no covariates. Although the estimated coefficients are identical (not shown), the two estimated average treatment effect on the treated are slightly different because of different weights used to average those identical estimated treatment effects (see Tables 5 and 6).

6.4 Computational time

Because we ran many different related models on the same data set, we kept track of the computational time to generate the estimates. A comparison of computational time, in relative terms, provides some insights into how long it would take to estimate these models in cases where the data sets and model specifications are much larger. Clearly, the specific amount of time will vary by computer memory, software, sample size, and model complexity. Yet we found some patterns that are illuminating.

Using the *lags only*, heterogeneous group-by-year regression specification without covariates as the benchmark for computational time, the specification that allows for covariates to be fully interacted takes almost 17 times as much computational time. Incorporating *leads* parameters into each specification (without covariates and with covariates fully interacted) adds only a small amount of computational time. The stacked-data regression specifications take about 6 times as much time as the benchmark regression and, once covariates are introduced, computational time increases to about 9 times relative to the benchmark. Overall,

the stacked-data specifications require about half the computational time compared to our preferred heterogeneous at the group-by-year with flexibly entered covariates specification, but, as we have shown above, the flexibility and allowance for additional heterogeneity are of considerable empirical value. Not surprisingly, the Callaway and Sant'Anna regression adjustment estimator takes the most computational effort: 30–50 times as much time as the benchmark when the specification has no covariates and 30–60 times as much time when covariates are introduced. Relative to the *lags and leads* specification with fully-interacted covariates estimated using OLS, the regression adjustment estimator takes about twice the computational time.

7 Conclusions

Our paper makes several theoretical and practical contributions to the difference-in-differences literature for the analysis of cross-sectional data. On the theoretical side, we prove that a linear regression with a sufficiently flexible functional form consisting of group-by-time treatment effects, two-way fixed effects, and interaction terms yields consistent estimates of heterogeneous treatment effects. The estimates are efficient and aggregation of treatment effects and inference are straightforward. The result holds when both the parallel trends and the no anticipation assumptions are true. We prove that an event-study model with leads and lags and appropriate interaction terms, estimated by ordinary least squares, returns numerically identical results as the imputation method by Borusyak et al. (2024). The theoretical result about repeated cross-sectional data is of importance to many applied researchers, because data are often not balanced panel data.

On the empirical side, we demonstrated our FLEX methods with two publicly available data sets to answer two research questions. Both empirical examples used individual-level cross-sectional data with staggered treatment at the state level. In both examples, our FLEX method and the imputation method obtained the same result. Our FLEX method generally has smaller standard errors than other popular estimators. In summary, our FLEX method has the advantage of being easy to implement, fast, and best among linear unbiased estimators.

References

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies*, page ˙https://doi.org/10.1093/restud/rdae007.

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

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The Effect of Minimum Wages on Low-Wage Jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454.

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

de Chaisemartin, C. and D'Haultfœuille, X. (2023). Two-way fixed effects and difference-indifferences with heterogeneous treatment effects: a survey. *Econometrics Journal*, 26:C1–C30.

Dette, H. and Schumann, M. (2024). Testing for Equivalence of Pre-Trends in Difference-in-Differences Estimation. *Journal of Business & Economic Statistics*, 1–13.

Fortin, N., Lemieux, T., and Lloyd, N. (2023). Right-to-Work Laws, Unionization, and Wage Setting. Celebratory Volume Research in Labor Economics.

Freedman, S. M., Hollingsworth, A., Simon, K. I., Wing, C., and Yozwiak, M. (2023). Designing Difference in Difference Studies With Staggered Treatment Adoption: Key Concepts and Practical Guidelines. NBER Working Paper 31842. https://www.nber.org/papers/w31842.

Goodman-Bacon, A. (2021). Difference-in-differences with Variation in Treatment Timing. *Journal of Econometrics*, 225(2):254–277.

MacKinlay, A. C. (1997). Event Studies in Economics and Finance. *Journal of Economic Literature*, 35(1):13–39.

Makridis, C. A. (2019). Do Right-to-Work Laws Work? Evidence on Individuals' Well-Being and Economic Sentiment. *The Journal of Law and Economics*, 62(4):713–745.

Meinhofer, A., Witman, A., Maclean, J. C., and Bao, Y. (2022). Prenatal substance use policies and newborn health. *Health Economics*, 31(7):1452–1467.

Roth, J. (2022). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*, 4(3):305–322.

Roth, J. (2024). Interpreting Event-Studies from Recent Difference-in-Differences Methods.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2023). What's Trending in Difference-in-differences? A Synthesis of the Recent Econometrics Literature. *Journal of Econometrics*, 235(2):2218–2244.

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

Wexler, N. (2022). Wage and Empoloyment Effects of Right-to-Work Laws in the 2010s.

Wing, C., Freedman, S. M., and Hollingsworth, A. (2024). Stacked Difference-in-Differences. https://www.nber.org/papers/w32054.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. SSRN Working Paper 3906345. https://papers.ssrn.com/abstract=3906345.

Wooldridge, J. M. (2023). Simple Approaches to Nonlinear Difference-in-differences with Panel data. *The Econometrics Journal*, 26(3):C31–C66.

Table 1: States with punitive prenatal substance use policies by cohort

Cohort	States	Number of Observations
2007	ID	9,732
2008	SC	15,297
2009	AZ	10,476
2012	UT	21,864
2013	AL	10,857
2014	GA, MO, RI, TN	36,096
2015	NM	12,102
2017	CT OH	26,191
2018	KY	13,566
Never treated	AK CA DE HI IA KS ME MI MS MT NE NH NJ NY NC OR PA VT WA WV WY	284,265
Total		440,446

Notes: The repeated cross-sectional data are from the Behavioral Risk Factor Surveillance System (BRFSS) data for 34 states and for 14 years from 2005–2018. A cohort comprises states that implemented the policies in particular year. There are 10 cohorts (including a never treated cohort) labeled by the year in which they are first treated. See section 5.2 for more details.

Table 2: Sample means by punitive prenatal substance use policies treatment status

	Treated	Never treated
Number of good mental health days	25.306	25.655
State has punitive prenatal substance use policies	0.390	0.000
Age	34.151	34.308
Black race	0.130	0.083
Hispanic ethnicity	0.128	0.130
Education level:		
High school diploma or GED	0.260	0.245
Some college but no degree	0.315	0.299
Bachelors degree or higher	0.339	0.369
Annual household income ($'000s)	50.309	52.066
Observations	156181	284265

Notes: The repeated cross-sectional data are from the Behavioral Risk Factor Surveillance System (BRFSS) data for 34 states in 10 cohorts (including a never treated cohort) and for 14 years from 2005–2018. See section 5.2 for more details.

Table 3: States with right-to-work laws by cohort

Cohort	States	Number of Observations
2012	IN MI	61,129
2015	WI	29,564
2016	WV	22,038
2017	KY	20,789
Never treated	AK CA CO CT DE DC HI IL ME MD MA MN MO MT NH NJ NM NY OH OR PA RI VT WA	840,058
Total		973,578

Notes: The repeated cross-sectional data are from the Current Population Survey (CPS) data for 29 states and for 12 years from 2008–2019. A cohort comprises states that implemented the policies in particular year. There are 5 cohorts (including a never treated cohort) labeled by the year in which they are first treated. See section 5.3 for more details.

Table 4: Sample means by right-to-work treatment status

	Treated	Never treated
Hourly earnings ($)	21.881	24.957
State has Right to Work law	0.485	0.000
Female	0.494	0.498
Age	43.946	43.467
Black race	0.066	0.084
Hispanic ethnicity	0.037	0.128
Education level:		
Some college but no degree	0.173	0.156
Associates degree	0.129	0.109
Bachelors degree	0.215	0.256
Graduate or Professional degree	0.116	0.158
Married	0.643	0.605
Observations	133520	840058

Notes: The repeated cross-sectional data are from the Current Population Survey (CPS) data for 29 states in 5 cohorts (including a never treated cohort) and for 12 years from 2008–2019. See section 5.3 for more details.

Table 5: ATET of punitive prenatal substance use policies on mental health

Model	Effect heterogeneity	Covariates	ATET	Std. err.	p-value
LAGS ONLY MODELS					
FLEX	Group & Time	None	-0.1899	0.0511	0.0007
		Interacted	-0.1445	0.0451	0.0030
	Cohort & Time	None	-0.2463	0.0768	0.0030
		Interacted	-0.2010	0.0677	0.0056
Stacked DID	Time-since-event	None	-0.1394	0.1062	0.1891
		Additive	-0.1229	0.1114	0.2698
TWFE	Homogeneous	None	-0.1520	0.1028	0.1487
		Additive	-0.1204	0.1080	0.2734
LAGS AND LEADS MODELS					
FLEX	Group & Time	None	-0.1644	0.0424	0.0005
		Interacted	-0.1453	0.0385	0.0006
	Cohort & Time	None	-0.2150	0.0588	0.0009
		Interacted	-0.1835	0.0525	0.0014
Stacked DID	Time-since-event	None	-0.1566	0.0983	0.1112
		Additive	-0.1416	0.1062	0.1824
CSRA	Cohort & Time	None	-0.2177	0.1007	0.0306
		Flexible	-0.2114	0.1018	0.0377
Event Study	Time-since-event	None	-0.1410	0.0928	0.1384
		Additive	-0.1312	0.1079	0.2323

Notes: For models with heterogeneous effects, ATET is a weighted average of the estimand. Standard errors of ATET are based on cluster (group) robust standard errors of the coefficients in the estimand. As summarized in Table 2, 7 covariates are used. Some regression specifications have no covariates (None), in some covariates enter only additively (Additive), while in others covariates enter additively and interacted with the estimand coefficients and with group and year indicators (Interacted). FLEX refers to the model developed in this paper. TWFE refers to the homogeneous effect two-way fixed effects estimator. Event Study refers to the heterogeneous over event-time but constant across groups two-way fixed effects estimator. Stacked DID refers to regressions on samples of data in which each cohort is first associated with never-treated controls and then the samples associated with each cohort are pooled (Cengiz et al., 2019). CSRA refers to the Callaway and Sant'Anna (2021) regression-adjustment estimator that uses a flexible regression specification to estimate the parameters of the conditional mean model and an influence function approach to estimate the standard errors of the estimates.

Table 6: ATET of right-to-work laws on hourly earnings

Model	Effect heterogeneity	Covariates	ATET	Std. err.	p-value
LAGS ONLY MODELS					
FLEX	Group & Time	None	-0.6687	0.1100	0.0000
		Interacted	-0.3656	0.0757	0.0000
	Cohort & Time	None	-0.6516	0.2031	0.0033
		Interacted	-0.3391	0.1301	0.0145
Stacked DID	Time-since-event	None	-0.6837	0.2053	0.0009
		Additive	-0.4565	0.1806	0.0115
TWFE	Homogeneous	None	-0.6457	0.2015	0.0034
		Additive	-0.4232	0.1767	0.0236
LAGS AND LEADS MODELS					
FLEX	Group & Time	None	-0.5837	0.1161	0.0000
		Interacted	-0.3171	0.0872	0.0011
	Cohort & Time	None	-0.5670	0.1809	0.0040
		Interacted	-0.3031	0.1286	0.0256
Stacked DID	Time-since-event	None	-0.6377	0.1739	0.0002
		Additive	-0.4201	0.1461	0.0040
CSRA	Cohort & Time	None	-0.5757	0.2063	0.0053
		Flexible	-0.2910	0.1377	0.0347
Event Study	Time-since-event	None	-0.6194	0.1640	0.0008
		Additive	-0.4052	0.1386	0.0068

Notes: For models with heterogeneous effects, ATET is a weighted average of the estimand. Standard errors of ATET are based on cluster (group) robust standard errors of the coefficients in the estimand. As summarized in Table 4, 9 covariates are used. Some regression specifications have no covariates (None), in some covariates enter only additively (Additive), while in others covariates enter additively and interacted with the estimand coefficients and with group and year indicators (Interacted). FLEX refers to the model developed in this paper. TWFE refers to the homogeneous effect two-way fixed effects estimator. Event Study refers to the heterogeneous over event-time but constant across groups two-way fixed effects estimator. Stacked DID refers to regressions on samples of data in which each cohort is first associated with never-treated controls and then the samples associated with each cohort are pooled (Cengiz et al., 2019). CSRA refers to the Callaway and Sant'Anna (2021) regression-adjustment estimator that uses a flexible regression specification to estimate the parameters of the conditional mean model and an influence function approach to estimate the standard errors of the estimates.

Table 7: Numbers of parameters estimated in the punitive prenatal substance use policies on mental health application

	Effect		Main effects			Interactions			
Model	heterogeneity	Covariates	Lags	Leads	FE	Lags	Leads	FE	Total
Lags only models									
FLEX	Group & Time	0	75	0	46	0	0	0	122
		7	75	0	46	525	0	322	976
	Cohort & Time	0	58	0	22	0	0	0	81
		7	58	0	22	406	0	154	648
Stacked DID	Time-since-event	0	12	0	414	0	0	0	427
		7	12	0	414	0	0	0	434
TWFE	Homogeneous	0	1	0	46	0	0	0	48
		7	1	0	46	0	0	0	55
Lags and leads models									
FLEX	Group & Time	0	75	59	46	0	0	0	181
		7	75	94	46	525	658	322	1728
	Cohort & Time	0	58	59	22	0	0	0	140
		7	58	59	22	406	413	154	1120
Stacked DID	Time-since-event	0	12	12	414	0	0	0	439
		7	12	12	414	0	0	0	446
CSRA	Cohort & Time	0	58	59	234	0	0	0	468
		7	58	59	234	0	0	0	1287
Event Study	Time-since-event	0	12	12	46	0	0	0	71
		7	12	12	46	0	0	0	78

Notes: The repeated cross-sectional data are from the Behavioral Risk Factor Surveillance System (BRFSS) data for 34 states in 10 cohorts (including a never treated cohort) over 14 years. Some regression specifications have no covariates, in some covariates enter only additively, while in others covariates enter additively and interacted with the estimand coefficients and with group and year indicators. FLEX refers to the model developed in this paper. TWFE refers to the homogeneous effect two-way fixed effects estimator. Event Study refers to the heterogeneous over event-time but constant across groups two-way fixed effects estimator. Stacked DID refers to regressions on samples of data in which each cohort is first associated with never-treated controls and then the samples associated with each cohort are pooled (Cengiz et al., 2019). For each of these regression specifications, the column labeled "Total" also includes an intercept. CSRA refers to the Callaway and Sant'Anna (2021) regression-adjustment estimator that uses a flexible regression specification to estimate the parameters of the conditional mean model and an influence function approach to estimate the standard errors of the estimates.

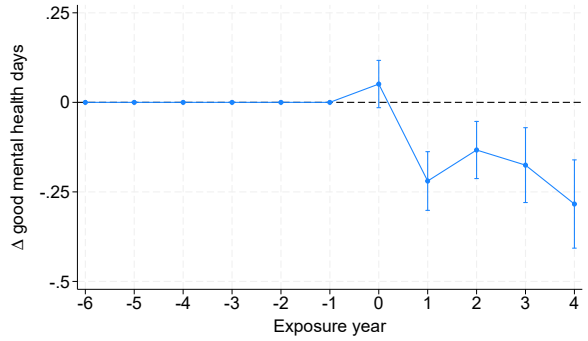
Table 8: Numbers of parameters estimated in the right-to-work laws on hourly earnings application

	Effect	Main effects				Interactions			
Model	heterogeneity	Covariates	Lags	Leads	FE	Lags	Leads	FE	Total
Lags only models									
FLEX	Group & Time	0	28	0	39	0	0	0	68
		9	28	0	39	28	0	351	456
	Cohort & Time	0	20	0	15	0	0	0	36
		9	20	0	15	20	0	135	200
Stacked DID	Time-since-event	0	12	0	156	0	0	0	169
		9	12	0	156	0	0	0	178
TWFE	Homogeneous	0	1	0	39	0	0	0	41
		9	1	0	39	0	0	0	50
Lags and leads models									
FLEX	Group & Time	0	28	24	39	0	0	0	92
		9	28	27	39	28	27	351	510
	Cohort & Time	0	20	24	15	0	0	0	60
		9	20	24	15	20	24	135	248
Stacked DID	Time-since-event	0	12	12	156	0	0	0	181
		9	12	12	156	0	0	0	190
CSRA	Cohort & Time	0	20	24	88	0	0	0	176
		9	20	24	88	0	0	0	572
Event Study	Time-since-event	0	12	12	39	0	0	0	64
		9	12	12	39	0	0	0	73

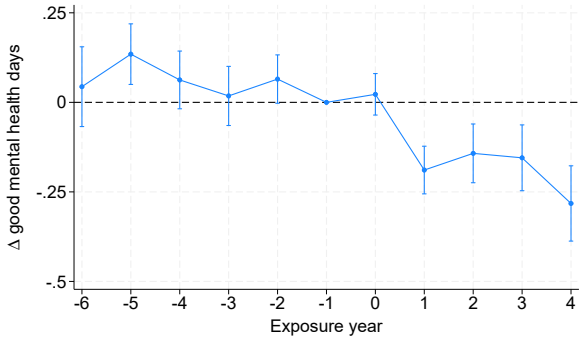
Notes: The repeated cross-sectional data are from the Current Population Survey (CPS) data for 29 states in 5 cohorts (including a never treated cohort) over 12 years. Some regression specifications have no covariates, in some covariates enter only additively, while in others covariates enter additively and interacted with the estimand coefficients and with group and year indicators. FLEX refers to the model developed in this paper. TWFE refers to the homogeneous effect two-way fixed effects estimator. Event Study refers to the heterogeneous over event-time but constant across groups two-way fixed effects estimator. Stacked DID refers to regressions on samples of data in which each cohort is first associated with never-treated controls and then the samples associated with each cohort are pooled (Cengiz et al., 2019). For each of these regression specifications, the column labeled "Total" also includes an intercept. CSRA refers to the Callaway and Sant'Anna (2021) regression-adjustment estimator that uses a flexible regression specification to estimate the parameters of the conditional mean model and an influence function approach to estimate the standard errors of the estimates.

Figure 1: Heterogeneous ATET of punitive prenatal substance use policies on mental health

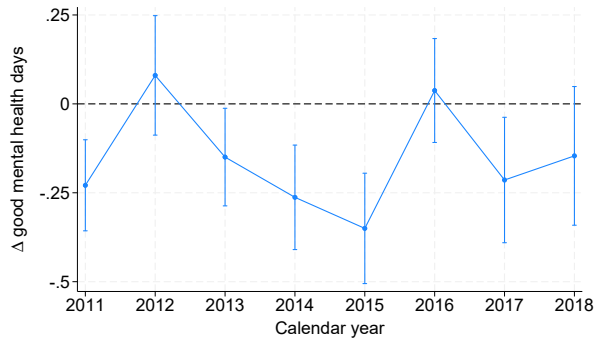
(a) ATET by exposure year: *lags only*



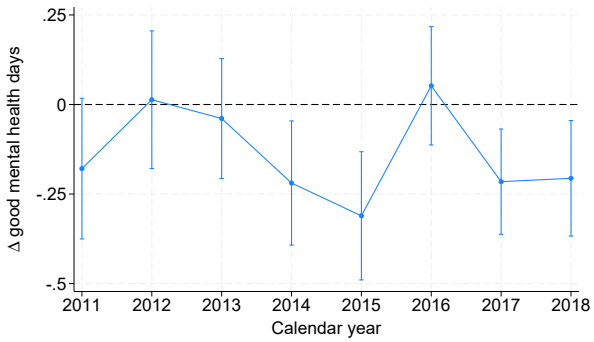
(b) ATET by exposure year: *lags and leads*



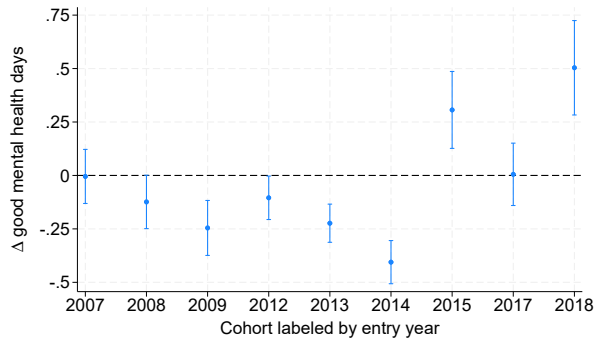
(c) ATET by calendar year: *lags only*



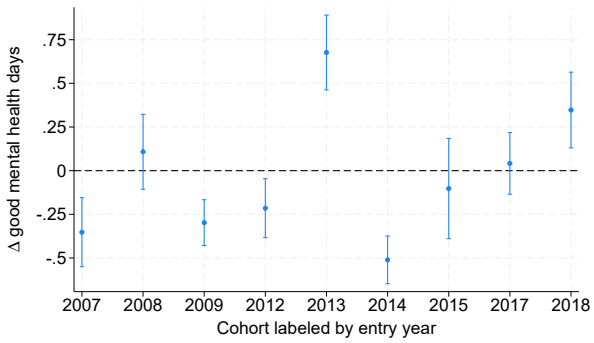
(d) ATET by calendar year: *lags and leads*



(e) ATET by cohort: *lags only*



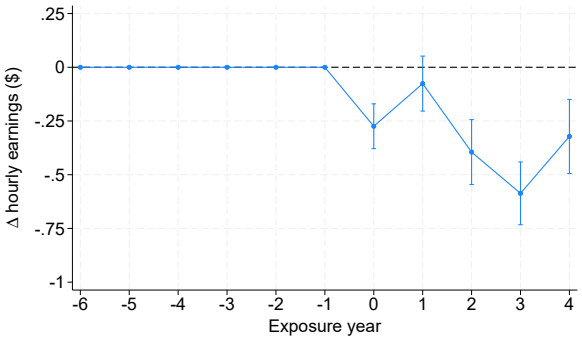
(f) ATET by cohort: *lags and leads*



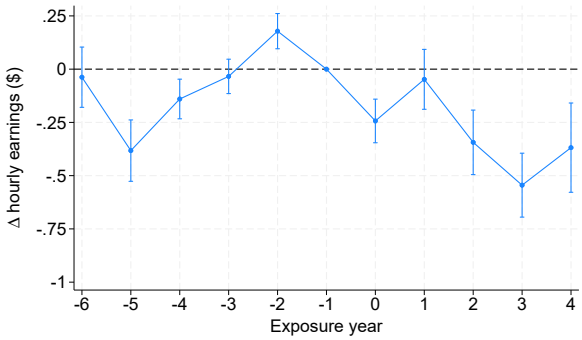
Notes: Regression models estimated with a fully interacted specification with estimands specified at the group by time level.

Figure 2: Heterogeneous ATET of Right-to-work laws on hourly earnings

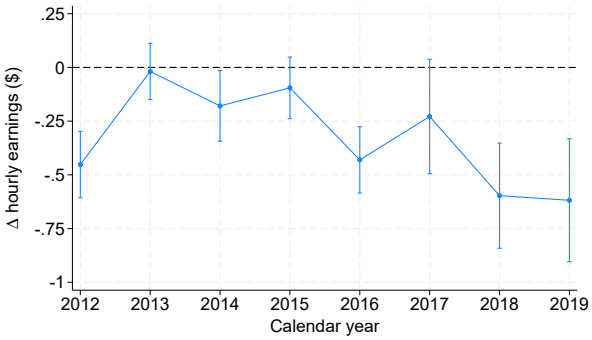
(a) ATET by exposure year: *lags only*



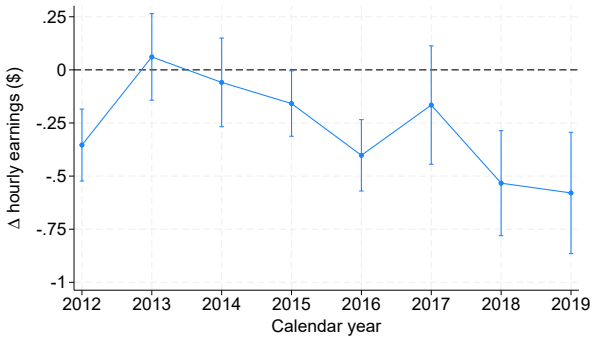
(b) ATET by exposure year: *lags and leads*



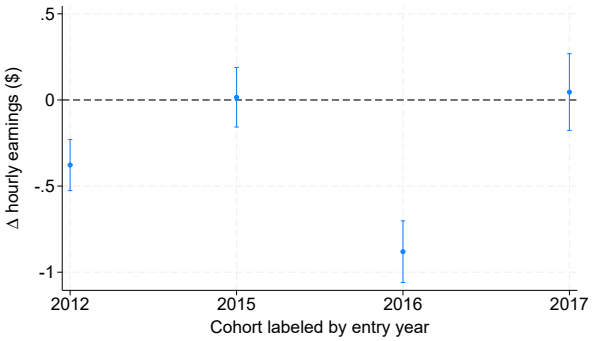
(c) ATET by calendar year: *lags only*



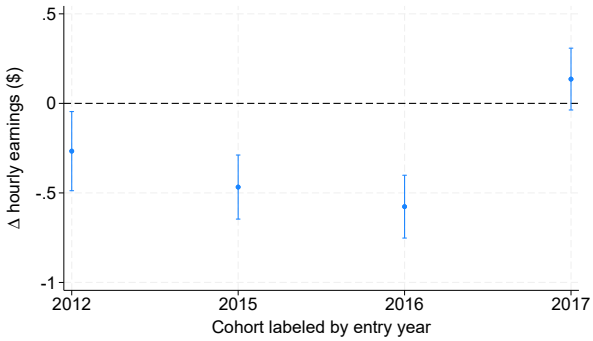
(d) ATET by calendar year: *lags and leads*



(e) ATET by cohort: *lags only*



(f) ATET by cohort: *lags and leads*



Notes: Regression models estimated with a fully interacted specification with estimands specified at the group by time level.