Two-stage differences in differences

John Gardner*

This version: September 2020

(Please see the most recent version)

Abstract

A recent literature has shown that when adoption of a treatment is staggered and average treatment effects vary across groups and over time, difference-in-differences regression does not identify an easily interpretable measure of the typical effect of the treatment. In this paper, I extend this literature in two ways. First, I provide some simple underlying intuition for why difference-in-differences regression does not identify a group×period average treatment effect. Second, I propose an alternative two-stage estimation framework, motivated by this intuition. In this framework, group and period effects are identified in a first stage from the sample of untreated observations, and average treatment effects are identified in a second stage by comparing treated and untreated outcomes, after removing these group and period effects. The two-stage approach is robust to treatment-effect heterogeneity under staggered adoption, and can be used to identify a host of different average treatment effect measures. It is also simple, intuitive, and easy to implement. I establish the theoretical properties of the two-stage approach and demonstrate its effectiveness and applicability using Monte-Carlo evidence and an example from the literature.

Keywords: Differences in differences, treatment effects, program evaluation, heterogenous treatment effects, misspecification.

*Department of Economics, University of Mississippi, University, MS. jrgardne@olemiss.edu.

JEL codes: C01, C10, C21, C22, C23.

1 Introduction

The difference-in-differences methodology has become an essential part of the applied empiricist's toolkit for evaluating treatment effects. Recently, however, an insightful literature has shown that, when the adoption of the treatment by different groups is staggered over time, and the average effects of the treatment vary over groups and across time, the usual difference-in-differences regression specification does not identify a readily interpretable measure of the typical effect of the treatment (see Borusyak and Jaravel, 2017; Athey and Imbens, 2018; Goodman-Bacon, 2018; de Chaisemartin and D'Haultfœuille, 2020; Imai and Kim, 2020; Sun and Abraham, 2020). Given the popularity, and the utility, of differences in differences, this is disconcerting.

In this paper, I extend the literature on difference-in-differences with staggered adoption and heterogeneous treatment effects in two ways. First, I present some clarifying intuition for why differences-in-differences may not identify the average effect of the treatment on the treated. Motivated by this intuition, I then develop a simple two-stage alternative to difference-in-differences regression that is robust to treatment-effect heterogeneity when adoption is staggered.

Presumably, part of why difference-in-differences regression is ubiquitous in settings with multiple groups and time periods is because it seems natural that it should identify the average effect of the treatment on the treated.[1] Since it *does* identify the average of heterogeneous treatment effects as long as those effects are distributed identically across treatment groups and periods (a condition that holds automatically in the classical two-group, two-period setting), this is an understandable misconception. When those distributions are not identical, however, conditional mean outcomes are not linear in group, period, and treat-

[1] As Borusyak and Jaravel (2017) write, "...there is a perception that [differences in differences] should estimate average treatment effects with some reasonable weights."

ment status, so the standard differences-in-differences regression model is misspecified, and therefore does not identify the average effect of the treatment on the treated.

The helps explain why differences-in-differences may not identify the average effect of the treatment on the treated, but says little about what it does identify. Several papers have provided alternative representations of the difference-in-difference regression estimand. Borusyak and Jaravel (2017) show that regression difference-in-differences identifies a regression-weighted mean of the average effect of the treatment in each post-treatment period, and de Chaisemartin and D'Haultfœuille (2020) show that all two-way fixed-effects regression estimates (which include difference-in-differences regressions as a special case) identify weighted averages of group- and period-specific average treatment effects. Since the weights in both of these representations can be negative, the difference-in-differences estimand may be difficult to interpret. Goodman-Bacon (2018) further shows that the regression difference-in-differences estimate represents a weighted average of all two-group, two-period differences in differences, which under parallel trends identifies a combination of weighted averages of group×period-specific average treatment effects and changes over time in those effects. As I discuss below, these decomposition results can be interpreted as describing how misspecified difference-in-differences regression models project heterogeneous treatment effects onto group and period fixed effects, rather than treatment status itself.

There are several alternatives to the difference-in-differences regression approach that are robust to heterogeneity across groups and periods when treatment adoption is staggered. Following Gibbons et al. (2017), Borusyak and Jaravel (2017), and Sun and Abraham (2020), one alternative is to estimate separate average treatment effects for each group and period, which can then be aggregated to form measures of the overall effect of the treatment (Gibbons et al., 2017, suggest an approach like this for fixed effects models, Borusyak and Jaravel, 2017, suggest such a solution for difference-in-differences models in which the duration-specific effects of the treatment are identical across groups, and Sun and Abraham, 2020, suggest such a solution for event-study regressions where duration-specific average treatment effects

vary across groups). Another is the "stacked" difference-in-differences approach (see, e.g., Gormley and Matsa, 2011; Deshpande and Li, 2019; Cengiz et al., 2019), which attempts to transform the staggered adoption setting to a two-group, two-period design (in which difference in differences identifies the average effect of the treatment on the treated) by stacking separate datasets containing observations on treated and control units for each treatment group.[2] Callaway and Sant'Anna (2018, cf. Abadie, 2005) develop a propensity-score reweighting approach that can be used to identify a suite of average treatment-effect measures.[3]

I develop an alternative, two-stage regression approach to identification that is robust to treatment-effect heterogeneity when adoption of the treatment is staggered. In its simplest form, the first stage of the procedure consists of a regression of outcomes on group and period fixed effects, estimated using the subsample of untreated observations. In the second stage, the estimated group and period effects are subtracted from observed outcomes, and these adjusted outcomes are regressed on treatment status. Under the usual parallel trends assumption, this procedure identifies the average effect of the treatment on the treated, even when average treatment effects are heterogeneous over groups and periods.

The two-stage approach can be adapted to recover a variety of treatment effect measures, and extended to event-study analyses of pre-trends and duration-specific average treatment effects. It can be computed easily, along with valid asymptotic standard errors, using standard statistical software, with little programming beyond that required to estimate a regression. It is also simple, and preserves the intuition behind identification in the two-group, two-period case: it recovers the average difference in outcomes between treated and untreated units, after removing group and period effects.

I motivate the problem with the difference-in-differences regression approach, and discuss

[2] In Appendix A, I show that this estimator identifies an average of group-specific average treatment effects, weighted by the relative sizes of the group-specific datasets and the variance of treatment status within those datasets.

[3] A major advantage of this approach is that it is able to flexibly handle covariates, which I abstract away from in this paper.

the difference-in-differences regression estimand, in Section 2. I introduce the two-stage approach and its properties in Section 3. I illustrate the performance of the two-stage approach using Monte Carlo evidence in Section 4 and its application in Section 5. I conclude in Section 6.

2 Motivation

2.1 The problem with difference-in-differences regression

Difference-in-differences research designs attempt to identify the causal effects of treatments under the parallel or common trends assumption, which asserts that, absent the treatment, treated units would experience the same change in outcomes as untreated units. Mathematically, this amounts to the assumption that average untreated potential outcomes decompose into additive group and period effects. More formally, let i index units (e.g., states or, with microdata, individuals) and t index calendar time (often years). Further divide individuals and time into treatment groups $g \in \{0, 1, \ldots, G\}$ and periods $p \in \{0, 1, \ldots, P\}$ defined by the adoption of the treatment among successive groups, so that members of group 0 are untreated in all periods, only members of group 1 are treated in period 1, members of groups 1 and 2 are treated in period 2, and so on. Let Y_{gpit}, Y_{1gpit} and Y_{0gpit} denote the observed, treated, and untreated potential outcomes for the ith member of group g during time t of period p, D_{gp} be an indicator for whether members of group g are treated in period p, and $\beta_{gp} = E(Y_{1gpit} - Y_{0gpit}|g, p)$ denote the average causal effect of the treatment for members of g in p. Under parallel trends, mean outcomes satisfy

$$E(Y_{gpit}|g, p, D_{gp}) = \lambda_g + \gamma_p + \beta_{gp} D_{gp}. \tag{1}$$

The idea behind differences in differences is to eliminate the permanent group effects λ_g and secular period effects γ_p, in order to identify the average effect of the treatment. In

the classic setup, there are only two periods (pre and post) and two groups (treatment and control). In this setting, within-group differences over time eliminate the group effects and within-period differences between groups eliminate the period effects. Hence the between-group difference in post-pre differences (i.e., the difference in differences) identifies the average effect of the treatment for members of the treatment group during the post-treatment period.

The two-period, two-group difference-in-differences estimate can be obtained manually, by calculating each of the four group×period averages and taking differences, or via a regression of outcomes on group and period fixed effects and a treatment-status indicator:

$$Y_{gpit} = \lambda_g + \gamma_p + \beta D_{gp} + \varepsilon_{gpit}. \tag{2}$$

It follows from (1) that the coefficient on D_{gp} in (2) identifies the average effect of the treatment on the treated, or $\beta_{11} = E(Y_{1gpit} - Y_{0gpit}|D_{gp} = 1)$.[4]

The regression approach suggests a natural way to extend the differences-in-differences idea to settings with multiple groups and time periods. Unfortunately, as several authors have noted (Borusyak and Jaravel, 2017; Goodman-Bacon, 2018; Athey and Imbens, 2018; de Chaisemartin and D'Haultfœuille, 2020; Imai and Kim, 2020), when the average effect of the treatment varies across groups, the coefficient on D_{gp} in specification (2) does not always identify an easily interpretable measure of the "typical" effect of the treatment. Although this result is now well established, because it is also somewhat counterintuitive, it bears further clarification.

While there are multiple ways to think about the typical effect of the treatment when

[4]There are several equivalent variations on this regression. Specification (2) is identical to a regression of outcomes on an indicator $Post_{it}$ for whether t occurs in the post-treatment period, an indicator $Treat_{it}$ for whether i belongs to the treatment group, and an interaction between the two. Often, the group and period effects λ_g and γ_p in (2) are replaced with individual and time effects λ_i and γ_t. By the Frisch-Waugh-Lovell theorem, the coefficient on D_{gp} in (2) can be obtained by regressing Y_{gpit} on the residuals from a regression of treatment status on group and period effects. Since treatment status only varies by group and period, these residuals are the same as those from a regression of treatment status on individual and time effects, so the coefficients on treatment status from both specifications are identical (despite the fact that the latter model is misspecified for $E(Y_{it}|i, t, D_{it})$).

that effect varies across groups and over time (see Section 3.1 below), an obvious candidate is the average $E(\beta_{gp}|D_{gp}=1) = E(Y_{1gpit} - Y_{0gpit}|D_{gp}=1)$ of group- and period-specific average treatment effects, taken over all units that receive the treatment and all times during which they receive it. This is what differences in differences identifies in the two period, two group case, and is probably what most people have in mind when they think about heterogenous treatment effects averaged over groups and time. Using this measure, parallel trends can be expressed as

$$E(Y_{gpit}|g,p,D_{gp}) = \lambda_g + \gamma_p + E(\beta_{gp}|D_{gp}=1)D_{gp} + [\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp}.$$

The difficulty with the regression approach arises because, except in special cases, the "error term" $[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp}$ in this expression varies at the group×period level, and is not mean zero conditional on group membership, period, and treatment status. Consequently, the conditional expectation $E(Y_{gpit}|g,p,D_{gp})$ is not, in general, a linear function of those variables (at least, not one in which the coefficient on D_{gp} is $E(\beta_{gp}|D_{gp}=1)$), and (2) is misspecified.[5] In contrast to the two-group, two-period case, the coefficient on D_{gp} from the regression differences-in-differences specification (2) does not identify the average difference $E(\beta_{gp}|D_{gp}=1)$ in outcomes between treated and untreated units after removing group and period fixed effects, unless those average effects are independent of group in time (in which case $\beta_{gp} = E(\beta_{gp}|D_{gp}=1) = \beta$, which the coefficient on D_{gp} recovers). Outside of this special case, when average treatment effects vary across groups and periods, and the adoption of the treatment by different groups is staggered over time, difference-in-differences regression does not recover a simple group×period average treatment effect.

[5] Cf. Angrist and Pischke (2009, Theorem 3.1.4): "Suppose the CEF is linear. Then the population regression function is it."

2.2 The difference-in-differences regression estimand

The preceding argument clarifies why difference-in-differences regression may not recover a readily interpretable measure of the average effect of the treatment, but says nothing about what it does identify. An easy answer to that question is that (2) identifies the linear projection of average outcomes onto group and period effects and a treatment indicator (which differs $E(Y_{gpit}|g,p,D_{gp})$ when that function is nonlinear). To provide additional insight into the difference-in-differences estimand, it can be shown that, under parallel trends, the coefficient on D_{gp} from the difference-in-differences regression specification (2) identifies

$$\beta^* = \sum_{g=1}^{G}\sum_{p=g}^{P}\omega_{gp}\beta_{gp},$$

where

$$\omega_{gp} = \frac{\{[1-P(D_{gp}=1|g)]-[P(D_{gp}=1|p)-P(D_{gp}=1)]\}P(g,p)}{\sum_{g=1}^{G}\sum_{p=g}^{P}\{[1-P(D_{gp}=1|g)]-[P(D_{gp}=1|p)-P(D_{gp}=1)]\}P(g,p)}, \tag{3}$$

$P(D_{gp}=1|p)$ is the fraction of units that are treated in period p, $P(D_{gp}=1)$ is the fraction of unit$\times$times that are treated, and $P(p,g)$ is the population share of observations that correspond to group g and period p. This representation can be obtained from Theorem 1 of de Chaisemartin and D'Haultfœuille (2020), although I present an alternative derivation based on population regression algebra in Appendix A.[6]

Appearances notwithstanding, this weighting scheme is deeply intuitive. Specification (2) assumes a conditional expectation function that is linear in group, period, and treatment status. When misspecified, it will attribute some of the heterogeneous impacts of the treatment to group and period fixed effects.[7] The longer a group's observed treatment duration

[6] An immediate implication of (3) is that the weights must sum to one. Another is that $\omega_{11}=1$ when there is only one treatment group, so the regression differences-in-differences specification (2) identifies the average effect of the treatment on the treated, as noted above.

[7] This is consistent with the intuition provided by Borusyak and Jaravel (2017), Goodman-Bacon (2018) and de Chaisemartin and D'Haultfœuille (2020) that (2) uses already-treated units as controls for newly treated ones.

(i.e., the greater $P(D_{gp} = 1|g)$ is), the more that group's treatment effects will be absorbed by group fixed effects. Likewise, the greater the probability of being treated in a particular period (i.e., the greater $P(D_{gp} = 1|p)$ is), the more treatment effects experienced during that period will be absorbed by period effects. Larger groups also receive more weight. I illustrate these phenomena by simulation in Section 4 and empirically in Section 5.

3 A two-stage approach

In the two-period, two-group case, differences-in-differences regression recovers the difference in outcomes between treated and untreated units after removing group and period effects, which under parallel trends represents the average effect of the treatment on the treated. This is not true when there are multiple groups and periods, since in this case (2) is misspecified for conditional mean outcomes. However, this observation suggests a simple two-stage average treatment effect estimator for the multiple group and period case. As long there are untreated and treated observations for each group and period, λ_g and γ_p are identified from the subpopulation of untreated groups and periods. The group×period average causal effect is then identified from a comparison of mean outcomes between treated and untreated groups, after removing the group and period effects.

Following this logic, a two-stage estimation procedure is

1. Estimate the model

$$Y_{gpit} = \lambda_g + \gamma_p + \varepsilon_{gpit}$$

on the sample of observations for which $D_{gp} = 0$, retaining the estimated group and time effects $\hat{\lambda}_g$ and $\hat{\gamma}_p$.[8]

[8]There are variations on this stage of the procedure. It is not necessary to estimate the fixed effects using only untreated observations, any correctly specified conditional mean function will do. For example, one could also use a specification that includes interactions between treatment status and period indicators, or one with group×period-specific treatment-status indicators. Since these variations utilize the entire sample to estimate the group and period effects, they might be more efficient. Because treatment status only varies at the group×period level, the group and period effects can also be replaced with fixed effects for individual units and time periods.

2. Regress adjusted outcomes $Y_{gpit} - \hat{\lambda}_g - \hat{\gamma}_p$ on D_{gp}.

Since parallel trends implies that

$$E(Y_{gpit}|g,p,D_{gp}) - \lambda_g - \gamma_p = \beta_{gp}D_{gp} = E(\beta_{gp}|D_{gp}=1)D_{gp} + [\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp},$$

where $E\{[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp}|D_{gp}\} = 0$, this procedure identifies $E(\beta_{gp}|D_{gp}=1)$, even when the adoption and average effects of the treatment are heterogenous with respect to groups and periods.[9]

3.1 Estimands

Implemented as described, the two-stage difference-in-differences estimator identifies $E(\beta_{gp}|D_{gp}=1)$, where the expectation is implicitly taken with respect to all observed units and periods. This expectation can be expressed as

$$E(\beta_{gp}|D_{gp}=1) = \sum_{g=1}^{G}\sum_{p=g}^{P} \beta_{gp} P(g,p|D_{gp}=1). \tag{4}$$

While this is a natural summary measure of group×period-specific average treatment effects, since it reflects the uneven progression of different groups through the course of the treatment, it may not be especially informative for program evaluation and policy analysis. For example, even if the effects of the treatment are identical across groups, this measure will put more weight on groups that are in early stages of the treatment.[10] Callaway and Sant'Anna (2018) provide further discussion of how heterogeneous average treatment effects can be

[9] More formally the consistency of $(\hat{\lambda}_g, \hat{\gamma}_p)$ for (λ_g, γ_p) follows from (1) and standard least-squares consistency arguments if G is fixed as the sample size grows. Otherwise, if the Y_{gpit} represent sample averages calculated using N_{gpit} observations, then $Y_{gpit} \xrightarrow{p} E(Y_{gpit}|g,p,i,t)$ as $N_{gpit} \to \infty$. Since, under parallel trends, $E[E(Y_{gpit}|g,p,i,t)|g,p] = \lambda_g + \gamma_p$, $(\hat{\lambda}_g, \hat{\gamma}_p) \xrightarrow{p} (\lambda_g, \gamma_p)$ by the continuous mapping theorem. In either case, $Y_{gpit} - \hat{\lambda}_g - \hat{\lambda}_p \xrightarrow{p} E(Y_{gpit}|g,p,i,t) - \lambda_g - \gamma_p$. Since $E[E(Y_{gpit}|g,p,i,t)|g,p] - \lambda_g - \gamma_p = \beta_{gp}D_{gp}$ is linear in D_{gp}, the second stage coefficient is consistent for $E(\beta_{gp}|D_{gp}=1)$ by the continuous mapping theorem.

[10] When the treatment effects varies by group, it is unclear whether any summary measure will be informative about how the treatment might affect future groups. External validity with this type of heterogeneity is inherently challenging.

summarized.

If there is some treatment duration $\bar{P}$ which all, or a subset of, groups has completed, then an alternative summary measure is the $\bar{P}$-period average

$$\sum_{g=1}^{G} \sum_{p=g}^{g+\bar{P}-1} \beta_{gp} P(g|D_g = 1), \tag{5}$$

where $P(g|D_g = 1)$ is the fraction of treated units that belong to group g. Because this measure averages the group-specific average effects of the treatment for a common set of completed durations, it may provide a more balanced picture of the typical effect of the treatment, although it ignores the effects of the treatment for durations longer than $\bar{P}$ periods. The two-stage procedure can be modified to identify this measure by restricting the sample used in the second step to untreated observations and treated observations with durations no greater than $\bar{P}$.

3.2 Event studies

Difference-in-differences analyses are often accompanied by event-study regressions of the form

$$Y_{gpit} = \lambda_g + \gamma_p + \sum_{r=-R}^{P} \beta_r D_{rgp} + \varepsilon_{gpit}, \tag{6}$$

where for $r \leq 0$ the $D_{rgp} \in \{D_{-Rgp}, \ldots, D_{0gp}\}$ are $(r+1)$-period leads of treatment adoption, and for $r > 0$ the $D_{rgp} \in \{D_{1gp}, \ldots, D_{Pgp}\}$ are indicators for r-period treatment durations.[11] In principle, such regressions serve a dual purpose. First, they can be used to show how the effect of the treatment evolves over the course of the treatment. Second, the coefficients on the treatment adoption leads can be used as placebo tests for the plausibility of parallel trends.

[11] In event-study regressions, it is common practice to use calendar times t in place of more coarse treatment periods p.

Sun and Abraham (2020) show that, when duration-specific average treatment effects vary across groups, event-study regressions suffer from the same problem as difference-in-differences regressions. This can be seen using an argument similar to the one presented for difference-in-differences regressions in Section 2.1. Let Y_{rgpit} denote potential outcomes after r periods of treatment, and $\beta_{rgp} = E(Y_{rgpit} - Y_{0gpit}|g, p, D_{rgp} = 1)$ be the average effect of being treated for r periods for members of group g in time period p.[12] Under parallel trends, we can write

$$E[Y_{gpit}|g, p, (D_{rgp})] = \lambda_g + \gamma_p + \sum_{r=1}^{P} E(\beta_{rgp}|D_{rgp} = 1)D_{rpg} + \sum_{r=1}^{P}[\beta_{rgp} - E(\beta_{rgp}|D_{rgp} = 1)]D_{rgp},$$

where, in general, $E\{\sum_{r=1}^{P}[\beta_{rgp} - E(\beta_{rgp}|D_{rgp} = 1)]D_{rgp}|g, p, (D_{rgp})\} \neq 0$. Hence, mean outcomes are not necessarily linear in group, period, and treatment-duration indicators, so the coefficients on the D_{rgp} from (6) do not identify the average effects of being treated for r periods. Sun and Abraham (2020) further show that the coefficients on the adoption leads and duration indicators identify weighted averages of all of the group×period-specific average treatment effects. An important consequence of this is that the coefficients on the treatment-adoption leads D_{rgp}, $r \leq 0$, may be nonzero even if trends are, in fact, parallel.

The two-stage procedure developed above can be extended to the event-study setting by amending the second stage of the procedure to:

2'. Regress $Y_{gpit} - \hat{\lambda}_g - \hat{\gamma}_p$ on $D_{-Rgp}, \ldots, D_{0gt}, \ldots, D_{Pgp}$.

Following the logic of the previous section, because $E[Y_{gpit}|g, p, (D_{rgp})] - \lambda_g - \gamma_p$ is linear in the D_{rgp}, the coefficients on the D_{rgp} identify the average effects $E(\beta_{rgp}|D_{rgp} = 1)$.[13]

[12] There is a one-to-one correspondence between duration- and period-specific treatment effects. In terms of the group×period average treatment effects β_{gp}, the duration-specific effects satisfy $\beta_{rgp} = \beta_{g,p-g+1}$. While in principle the duration-specific average treatment effects for each group might vary over time, in practice we only ever observe each treatment duration at most once for each group.

[13] This expectation is taken over all groups with durations of at least r. Since under staggered adoption the completed treatment duration varies by group, the groups over which these duration-specific effects are averaged will vary across durations. These averages are also what the interaction-weighted estimator proposed by Sun and Abraham (2020) identifies. If all groups are treated for at least $\bar{P}$ periods, an alternative is to exclude observations corresponding to treatment durations longer than $\bar{P}$ periods from the second-stage

3.3 Inference

The standard errors for the two-stage estimators need to be adjusted to account for the fact that the dependent variable $Y_{gpit} - \hat{\lambda}_g - \hat{\gamma}_p$ in the second-stage is generated using estimates obtained from the first stage of the procedure (Dumont et al., 2005). The asymptotic distribution of the second-stage estimates can obtained by interpreting the two-stage procedure as a joint GMM estimator (Hansen, 1982).

Let $W_{gpit} = [Y_{gpit}, (1(g)_{gpit}), (1(p)_{gpit}), D_{gp}]$ denote the data for observation (g, p, i, t), consisting of the outcome Y_{gpit}, the G-vector of group-membership indicators $(1(g)_{gpit})$, a $(P-1)$-vector $(1(p)_{gpit})$ of period indicators for periods $p \in \{2, \ldots, P\}$, and the treatment-status indicator D_{gp}. Let λ be the G-vector of group fixed effects, γ the $(P-1)$-vector of period fixed effects, and β the group×period average treatment effect. The two-stage difference-in-differences estimator solves the population analog of the moment condition

$$E[f(\lambda, \gamma, \beta; W_{gpit})] = E\left[\begin{array}{c}[Y_{gpit} - (1(g)_{gpit})'\lambda - (1(p)_{gpit})'\gamma](1 - D_{gp}) \\ [Y_{gpit} - (1(g)_{gpit})'\lambda - (1(p)_{gpit})'\gamma] - \beta D_{gp}\end{array}\right] = 0.$$

By Theorem 6.1 of Newey and McFadden (1994, cf. Newey, 1984), $\sqrt{N}(\hat{\beta} - \beta) \sim N(0, v)$, where v is the last element of

$$E\left[\frac{\partial f(\lambda, \gamma, \beta; W_{gpit})}{\partial(\lambda, \gamma, \beta)}\right]^{-1} E[f(\lambda, \gamma, \beta; W_{gpit}) f(\lambda, \gamma, \beta; W_{gpit})'] E\left[\frac{\partial f(\lambda, \gamma, \beta; W_{gpit})}{\partial(\lambda, \gamma, \beta)}\right]^{-1'}.$$

Asymptotics for variations on the two-stage difference-in-differences estimator (such as the event-study version) are similar.

The preceding expression can be used to manually correct the estimated second-stage variances for the use of a generated dependent variable. With modern statistical software, a simpler approach is to estimate both stages of the procedure simultaneously using a GMM

sample, in which case the two-stage approach identifies duration-specific treatment effects, averaged over all groups.

routine. In Appendix B, I provide example Stata syntax that shows how to implement the two-stage difference-in-differences approach (with valid asymptotic standard errors) via GMM.

4 Simulations

To illustrate the effectiveness of the two-stage approach, I conduct two Monte Carlo studies. For each study, I simulate 250 datasets, each consisting of observations on 50 units over 10 periods. Unit-level outcomes are determined by

$$Y_{gpit} = \lambda_i + \gamma_t + \beta_{gp} D_{gp} + \varepsilon_{gpit},$$

with $\lambda_i, \varepsilon_{it} \sim N(0, 1)$. I assume that the average effect of the treatment varies by group and period, with the treatment effects for each group stabilizing by the fourth period.[14] For the first study, three treatment groups, each consisting of five units, adopt the treatment at times 4, 5, and 6, respectively. The only difference for the second study is that the sizes of the treatment groups vary, consisting of 5, 15, and 10 units.

Focusing initially on the first simulation, Figure 1 plots the weights that the regression difference-in-differences specification (2) places on each of the group×period-specific treatment effects (here, I have aligned the weights by the duration of the treatment, rather than time periods). For the first two groups, the weight that (2) places on period-specific treatment effects decreases with each successive period, until the final group adopts the treatment and the treated share of units stabilizes. After this stabilization, treatment effects for earlier groups, who are treated for more periods, receive less weight. This is consistent with the theoretical predictions and intuition from Section 2.2.

Table 1 presents the means and standard deviations of estimates of several average treat-

[14] The vectors of average treatment effects for the first four periods are (2, 4, 6, 8), (1, 2, 3, 4), and (.5, 1, 3, 3.5) for groups one, two, and three.

ment effect measures. Results for the first simulation are presented in the first column. The first row of the top panel of the table presents the true group×period average treatment effect (i.e., (4)). The difference-in-differences regression estimate, which as (3) shows and Figure 1 illustrates, tends to put more weight on earlier treatment durations, understates the true average effect considerably.

The third row summarizes estimates of the group×period average, obtained by estimating models with separate treatment-status indicators for each group and period (as well as unit and time fixed effects), then aggregating the group×period-specific average effects according to the empirical distribution of treated groups and periods (i.e., the sample analog of (4)). Finally, the fourth row summarizes two-stage difference-in-difference regression estimates, obtained following the procedure outlined in Section 3. In this case, the two-stage estimates are indistinguishable from the aggregated estimates. Both perform well for the true group×period average treatment effect.

As I note in Section 3.1, the group×period average treatment effect might not be the best way to summarize the effect of the treatment. Since all units in this simulation are treated for at least four periods, it is possible to estimate a four-period treatment effect, averaged across all treated groups. The first row of the bottom panel of Table 1 presents the true four-period average.[15] The second row presents estimates aggregated from regressions that include group×period-specific treatment indicators (i.e., the sample analog of (5)).

The third row presents stacked difference-in-differences regression estimates of this average. To implement the stacked estimator, for each treated group I create a new dataset spanning two periods before and four periods after that group adopts the treatment, consisting of observations on the treatment group and the group of units that never receives the treatment. I then stack these group-specific datasets and regress outcomes on treatment status and dataset-specific group and period effects.[16]

[15]Because the difference-in-difference regression estimate represents a weighted average of all observed group×period treatment effects, it is not directly comparable to this four-period average.

[16]In some applications, treated units who have not yet adopted the treatment are also included as controls in each group-specific dataset. While it is possible to use the stacked approach to estimate a weighted

Finally, the fourth row presents two-stage difference-in-difference estimates, obtained by restricting the second stage to the sample of observations with treatment durations no greater than four periods. Here, the aggregated, stacked, and two-stage estimates are identical; all are centered closely on the true four-period average effect.

The sole difference between the first and second studies is that, in the second, the sizes of the treatment groups vary. The results for this study are presented in the second column of the table. Here, the relative performance of the aggregated and two-stage estimators for the group×period average treatment effect is similar to the first study. For the four-period average, the aggregated and two-stage estimates are again identical, and close to the true average. In contrast, the stacked estimator, which weights each group's average treatment effect by dataset-specific treatment variance and sample size (see Appendix A), overstates the true average.

The top panel of Figure 2 summarizes estimates from event-study regressions that include two leads of treatment adoption. The top panel of the figure plots the average point estimates (and standard deviations) from the standard event-study specification (6). Even though the data-generating process satisfies parallel trends, because the group×period average treatment effects are heterogeneous, the estimated leads of adoption exhibit a pre-treatment dip in outcomes, creating the mistaken appearance of a violation of parallel trends. This is consistent with the results of Sun and Abraham (2020) and the discussion in Section 3.2. The second panel summarizes estimates obtained by manually aggregating group×duration-specific average treatment effects across groups, and the third summarizes estimates from the two-stage event-study procedure outlined in Section 3.2. The aggregated and two-stage results are nearly identical; both present an accurate picture of pre-treatment trends, as well as the evolution of the effect of the treatment over its course (I present the means and standard deviations of the point estimates in Appendix Table 3).

group×period average treatment effect, in most applications the group-specific datasets include the same number of treatment periods for each group.

5 Empirical application

To illustrate the application of the results presented above, I revisit Autor's (2003) analysis of the effect of court rulings limiting the doctrine of employment at will on the growth of the temporary help services sector (THS). The key data are observations on the log of state-level THS employment and indicators for state-level legal exceptions to employment-at-will between 1977 and 1996. The baseline difference-in-differences specification regresses THS employment on an exception indicator and state and year effects. The estimate, reported in the top panel of Table 2, is about .11, with a standard error (clustered on state) of about .1.[17]

As the preceding results show, if the effect of the treatment is heterogeneous across treatment groups and periods, the difference-in-differences regression estimate represents a difficult-to-interpret weighted average of group×period treatment effects. The first group in this application is treated in 1976, and the treatment is initiated for additional groups in every year through 1988, with the exception of 1979, for a total of 12 groups. Since THS employment is observed between 1977 and 1996, the difference-in-differences estimate places nonzero weight on 177 group×period treatment effects. The top panel of Figure 3 plots the distribution of these weights. Some are negative, although more are positive, and the positive magnitudes tend to exceed the negative ones. The distribution is also skewed, with a handful of effects receiving relatively large weights. The bottom panel of the figure plots the weights themselves for the first five groups. The weights are consistent with the results in Section 2.2, with treatment effects for groups that are treated earlier, that occur in earlier periods for each group, and for larger groups receiving more weight.

As a point of comparison for the difference-in-differences regression estimate, Table 2 also presents an estimate of the group×period average treatment effect, obtained by estimating models with separate treatment indicators for each treated group and period,

[17] This differs slightly from the estimate reported in Autor (2003), which only uses observations between 1979 and 1995 (Autor's preferred specifications also include covariates, which I ignore for simplicity).

then averaging the coefficients on those indicators according to the empirical distribution of groups and periods among treated units. The aggregated point estimate of about .1 is only slightly smaller than the difference-in-differences regression estimate, implying that the group×period-specific average treatment effects are fairly homogeneous.[18]

Table 2 also presents a two-stage difference-in-differences estimate of this treatment effect, obtained by estimating both equations simultaneously via GMM (and clustering standard errors at the state level). The two-stage estimate is nearly identical to the aggregated estimate, illustrating the effectiveness of the two-stage approach. At the same time, the aggregated estimator was considerably more difficult to implement, requiring the estimation of 177 different group×period specific average treatment effects, estimation of the empirical distribution of groups and periods among the treated, aggregation of the group×period effects according to that distribution, and manual computation of the standard error of the aggregate. The table also presents the results from a version of the two-stage estimator in which the first-stage equation includes interactions between treatment status and period indicators, estimated using the full sample. The second-stage estimate is similar to the previous two-stage estimate (for which the first stage consists of a regression of untreated outcomes on state and time effects alone), although the standard error is slightly smaller.

As I note in Section 3.1, the treatment effects estimated above are averaged across groups with different completed treatment durations. Because the shortest observed duration is nine periods, it is possible to estimate the effect of being treated for nine periods, averaged across all treatment groups.[19] The aggregated estimate of this treatment effect is about .1. I also implement a stacked difference-in-differences estimator for this effect, in which the dataset for each treatment group consists of observations on that group and the group of never-treated observations, one period before through nine periods after the adoption of the

[18] I calculate standard error for the aggregated estimator using the delta method, taking the distribution of groups and periods among the treated as fixed, and using state-clustered standard errors for the group×period-specific regression coefficients.

[19] Because employment is not recorded for 1976 and the group that adopts the treatment in 1977 is treated in all years for which employment is available, I drop the first two treatment groups from the estimation sample.

treatment. The resulting estimate is also about .1. Finally, I estimate this effect by the two-stage difference in differences, simply by restricting the second-stage estimation sample to untreated observations and those with completed durations of nine periods or fewer. The two-stage and stacked estimates are identical.

Autor (2003) also estimates an event-study regression to test the plausibility of parallel trends. To replicate this, I estimate a standard event-study regression using a version of specification (6) that includes two leads of treatment adoption. The top panel of Figure 4 plots the treatment effects for two leads and the first nine periods of treatment (I present the full set of point estimates in Appendix Table 4). The estimates suggest that the effect of the treatment is larger in earlier periods and stabilizes to a smaller value in periods four and beyond.[20] To examine the effect of heterogeneity on the event-study estimates, I also perform an aggregated event study, based on a model that includes separate duration-specific effects for each treatment group.[21] The results, presented in the second panel of the figure, are very similar to those from the event study that ignores treatment groups, suggesting that there is little heterogeneity in the duration-specific effects with respect to the timing of adoption. Finally, the bottom panel of the figure plots the results from a two-stage event study, obtained by replacing treatment status in the second stage of the procedure with indicators for each treatment duration. The two-stage estimates are very similar to the aggregated estimates.

6 Conclusion

When adoption of a treatment is staggered across time, and the average effects of the treatment vary by group and period, the usual difference-in-differences regression specification does not identify an easily interpretable measure of the typical effect of the treatment. When

[20] This is consistent with the fact that the difference-in-differences regression estimate, which puts more weight on earlier periods for each group, overstates the group×period average treatment effect.

[21] Since the leads introduce perfect collinearity for the second and third treatment groups, I exclude those groups when calculating group-weighted average leads and duration-specific treatment effects.

the duration-specific effects are also heterogeneous, neither do the coefficients from the usual event-study specification. The ultimate source of these identification failures is that outcomes are not linear in group, period and treatment status, as difference-in-differences and event-study regression specifications assume.

The two-stage approach developed in this paper is motivated by the observation that, under parallel trends, untreated outcomes are linear in group and period effects. Those effects are therefore identified from a first-stage regression estimated using the sample of untreated observations. The average effect of the treatment on the treated is then identified from a regression of outcomes on treatment status, after removing group and period effects. This procedure is robust to the presence of heterogeneous treatment effects when treatment adoption is staggered. It is also simple and intuitive, and can be extended to identify a variety of different treatment effect measures. Monte Carlo simulations and an empirical example show that the two-stage estimators correctly identify informative average treatment effect measures, in some cases outperforming alternative estimators that are also more difficult to implement.

Appendix A: Longer proofs

Proof of (3)

From (1), we can write

$$Y_{gpit} = \lambda_g + \gamma_p + \sum_{h=1}^{G}\sum_{q=h}^{P} \beta_{hq} 1(h,q)_{gpit} + \varepsilon_{gpit}, \tag{7}$$

where $1(h,q)_{gpit}$ is an indicator for whether observation (g,p,i,t) corresponds to group h and period q, and $E[\varepsilon_{gpit}|g,p,(1(h,q)_{gpit})] = 0$.

Let $\tilde{D}_{gp}$ denote the residual from a population regression of D_{gp} on group and period fixed effects. By the Frisch-Waugh-Lovell theorem, the coefficient on D_{gp} from a population

regression of Y_{gpit} on D_{gp} and group and period effects is

$$\begin{aligned}\beta^* &= \frac{E(\tilde{D}_{gp}Y_{gpit})}{E(\tilde{D}_{gp}^2)} \\ &= \frac{E[\tilde{D}_{gp}\sum_{h=1}^{G}\sum_{q=h}^{P}\beta_{hq}1(h,q)_{gpit}]}{E(\tilde{D}_{gp}^2)} \\ &= \sum_{h=1}^{G}\sum_{q=h}^{P}\frac{E[\tilde{D}_{gp}1(h,q)_{gpit}]}{E(\tilde{D}_{gp}^2)}\beta_{hq} \\ &= \sum_{g=1}^{G}\sum_{p=g}^{P}\omega_{gp}\beta_{gp}.\end{aligned}$$

where ω_{gp} is the coefficient from a regression of $1(h,q)_{gpit}$ on D_{gp} and group and period fixed effects. The second equality uses the facts that ε_{gpit} is mean-independent of the regressors and that $\tilde{D}_{gp}$ is uncorrelated with group and period effects by construction.[22]

The weight ω_{gp} that difference in differences places on β_{gp} is the coefficient on D_{gp} from a regression of $1(g,p)_{gpit}$ on D_{gp} and group and period fixed effects. By the Frisch-Waugh-Lovell theorem, this is equivalent to the slope coefficient from a population regression of $1(g,p)_{gpit}$ on the residual from an auxiliary regression of D_{gp} on group and period effects. Using the two-way within or double-demeaned transformation, this residual can be expressed as

$$\tilde{D}_{gp} = [D_{gp} - P(D_{gp}=1|g)] - [P(D_{gp}=1|p) - P(D_{gp}=1)]. \tag{8}$$

Since $E(\tilde{D}_{gp}^2) = E(\tilde{D}_{gp}D_{gp})$, ω_{gp} can also be expressed as

$$\begin{aligned}\omega_{gp} &= \frac{E(1(g,p)_{gpit}\tilde{D}_{gp})}{Var(\tilde{D}_{gp})} \\ &= \frac{E(\tilde{D}_{gp}|1(g,p)_{gpit}=1)P(1(g,p)_{gpit}=1)}{E(\tilde{D}_{gp}|D_{gp}=1)P(D_{gp}=1)} \\ &= \frac{\{1-P(D_{gp}=1|g)-[P(D_{gp}=1|p)-P(D_{gp}=1)]\}P(g,p)}{\sum_{g=1}^{G}\sum_{p=g}^{P}\{1-P(D_{gp}=1|g)-[P(D_{gp}=1|p)-P(D_{gp}=1)]\}P(g,p)\}},\end{aligned}$$

[22] This, and the related result in Sun and Abraham (2020), can also be established by thinking of the term $\sum_{h=1}^{G}\sum_{q=h}^{P}\beta_{hq}1(h,q)_{gpit}$ in (7) as an omitted variable, and taking its projection onto the included regressors.

where the final equality uses (8).

6.1 Stacked differences in differences

In the stacked approach, a new dataset is created for each treated group, containing observations on that group $\bar{R}$ periods before, and $\bar{P}$ periods after, the treatment is adopted, as well as on units that are not yet treated during these periods. These group-specific datasets are stacked, and outcomes are regressed on treatment status and dataset-specific group and period fixed effects:

$$Y_{cgpit} = \lambda_{cg} + \lambda_{cp} + \beta D_{cgp} + \varepsilon_{cgpit},$$

where *cgpit* indexes the value of an observation in the dataset for group c for the ith member of group g during the tth time of period p.

Let D_{cgp} be an indicator for whether group g is treated during period p of the group-c dataset, and D_{rcgp} be an indicator for whether members of g have been treated for $r \in \{1, \ldots, \bar{P}\}$ periods as of period p in dataset c. Let $\tau = \bar{P}/(\bar{P} + \bar{R} + 1)$ denote the fraction of periods during which treated units in any group-specific dataset are treated, π_c denote the fraction of units in dataset c that belong to the treatment group, and ρ_c denote size of the group-c dataset relative to the stacked dataset.

The weight ω_{rg} that stacked differences in differences places on the r-period average treatment effect β_{rg} for group g is given by the slope coefficient from a population regression of D_{rcgp} on the residual $\tilde{D}_{cgp}$ from a regression of D_{cgp} on dataset×period and dataset×group effects. This residual is

$$\tilde{D}_{cgp} = D_{cgp} - P(D_{cgp} = 1|g, c) - [P(D_{cgp} = 1|p, c) - P(D_{cgp} = 1|c)],$$

where statements conditional on c are true in the population corresponding to dataset c.

Using this expression and adapting (3) to the stacked setting,

$$\begin{aligned}\omega_{rg} &= \frac{[1-\tau-(\pi_c-\tau\pi_c)]P(D_{rcgp}=1)}{\sum_{c=1}^{G}\sum_{p=1}^{\bar{P}}[1-\tau-(\pi_c-\tau\pi_c)]P(D_{rcgp}=1)} \\ &= \frac{(1-\tau)(1-\pi_c)\tau\pi_c\rho_c}{\sum_{c=1}^{G}\sum_{p=1}^{\bar{P}}(1-\tau)(1-\pi_c)\tau\pi_c\rho_c} \\ &= \frac{[1-\pi_c]\pi_c\rho_c}{\bar{P}\sum_{c=1}^{G}[1-\pi_c]\pi_c\rho_c}.\end{aligned}$$

Appendix B: Stata syntax

Suppose that `y` refers to the outcomes, `year` refers to the year, `id` refers to the group, and `d` refers to treatment status. The two-stage difference-in-differences estimator can be obtained, along with valid cluster-robust asymptotic standard errors, via GMM using the single Stata command:

```
gmm (eq1: (y-{xb: i.year}-{xg: ibn.id})*(1-d)) ///
  (eq2: y-{xb:} - {xg:} - {delta}*d), ///
  instruments(eq1: i.year ibn.id) ///
  instruments(eq2: d) winitial(identity) ///
  onestep quickderivatives vce(cluster id)
```

Variations on the two-stage estimator (such as the the two-stage event-study estimator) can be obtained using similar syntax.

References

Abadie, Alberto. 2005. "Semiparametric difference-in-differences estimators." Review of Economic Studies, 72(1): 1-19.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Athey, Susan and Guido W. Imbens. 2018. "Design based analysis in difference-in-differences settings with staggered adoption." NBER working paper no. 24963.

Autor, David. 2003. "Outsourcing at will." Journal of Labor Economics, 21(1): 142.

Borusyak, Kirill and Xavier Jaravel. 2017. "Revisiting event study designs, with an application to the estimation of the marginal propensity to consume." Working paper, Harvard University.

Callaway, Brantly, and Pedro Sant'Anna. 2018. "Difference-in-differences with multiple time periods and an application on the minimum wage and employment." Working paper.

Cengiz, Doruk, Dube, Arindrajit, Lindner, Attila and Ben Zipperer. 2019. "The effect of minimum wages on low-wage jobs." Quarterly Journal of Economics, 134(4): 1405-1454.

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." American Economic Review, forthcoming.

Deshpande, Manasi and Yue Li. 2019. "Who is screened out? Application costs and the targeting of disability programs." American Economic Journal: Economic Policy, 11(4): 213248.

Dumont, Michel, Rayp, Glenn, Thas, Oliver and Peter Willeme. 2005. "Correcting standard errors in two-stage estimation procedures with generated regressands." Oxford Bulletin of Economics and Statistics, 67(3): 03059049.

Gibbons, Charles E., Suárez Serrato, Juan Carlos, and Michael B. Urbancic. 2017. "Broken or fixed effects." Journal of Econometric Methods, 8(1): 2156-6674.

Goodman-Bacon, Andrew. 2018. "Difference-in-differences with variation in treatment timing." NBER working paper no. 25018.

Gormley, Todd A. and David A. Matsa. "Growing out of trouble? Corporate responses to liability risk." Review of Financial Studies, 24(8):27812821.

Hansen, Lars P. (1982). "Large sample properties of generalized method of moments estimators". Econometrica, 50(4): 10291054.

Imai, Kosuke and In Song Kim. 2020. "On the use of two-way fixed effects regression models for causal inference with panel data." Working paper.

Newey, Whitney. 1984. "A method of moments interpretation of sequential estimators." Economics Letters, 14: 201206.

Newey, Whitney K. and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." In R. F. Engle and D. L. McFadden (Eds.), Handbook of Econometrics, IV: 21122245. Elsevier Science.

Sun, Liyang and Sarah Abraham. 2020. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." Working paper.

Table 1: Simulation results

		Simulation 1	Simulation 2
All periods	True	4.08	3.46
	Diff-in-diff	3.51	2.71
		(1.06)	(0.24)
	Aggregated	4.12	3.48
		(1.02)	(0.23)
	Two-stage	4.12	3.48
		(0.28)	(0.23)
Four-period	True	3.17	2.75
	Aggregated	3.21	2.78
		(1.05)	(0.25)
	Stacked	3.21	2.87
		(1.05)	(0.26)
	Two-stage	3.21	2.78
		(0.32)	(0.25)

Notes: Means and standard deviations from 250 simulations.

Figure 1: Simulated DD weights

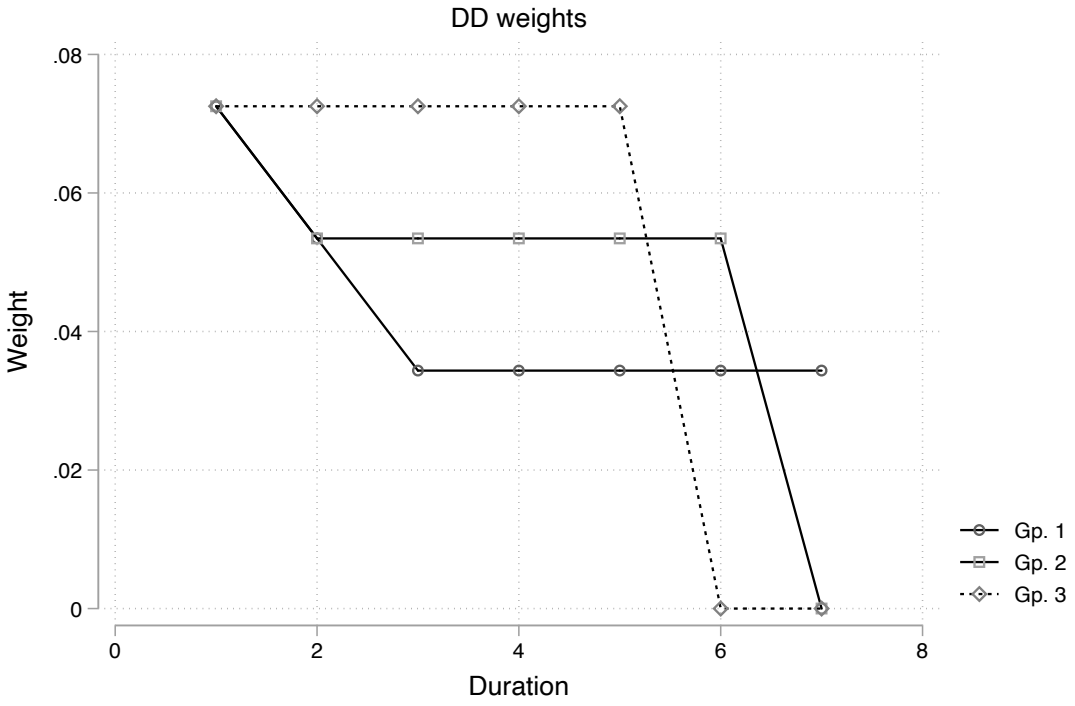
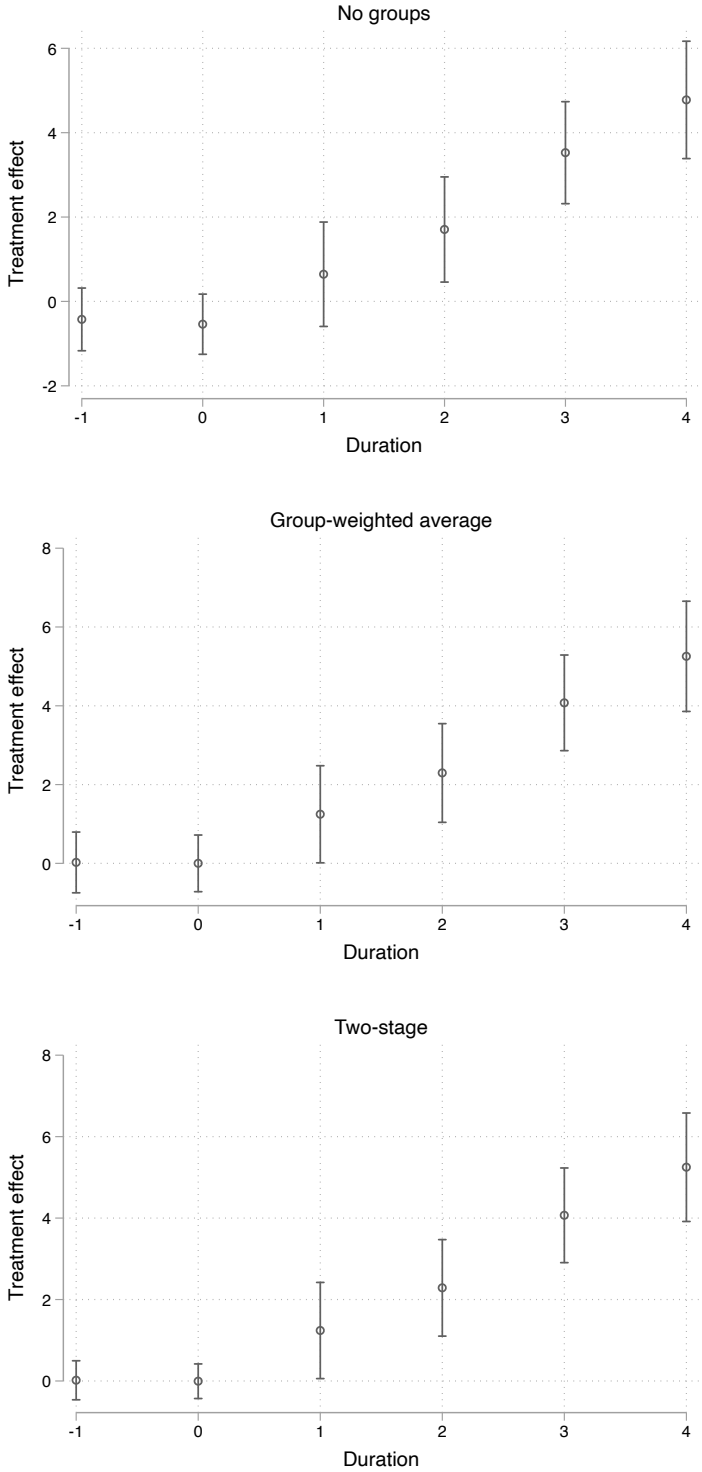


Figure 2: Simulated event studies



Notes: Means and ± 2×standard deviations from 250 simulations.

Figure 3: Application DD weights

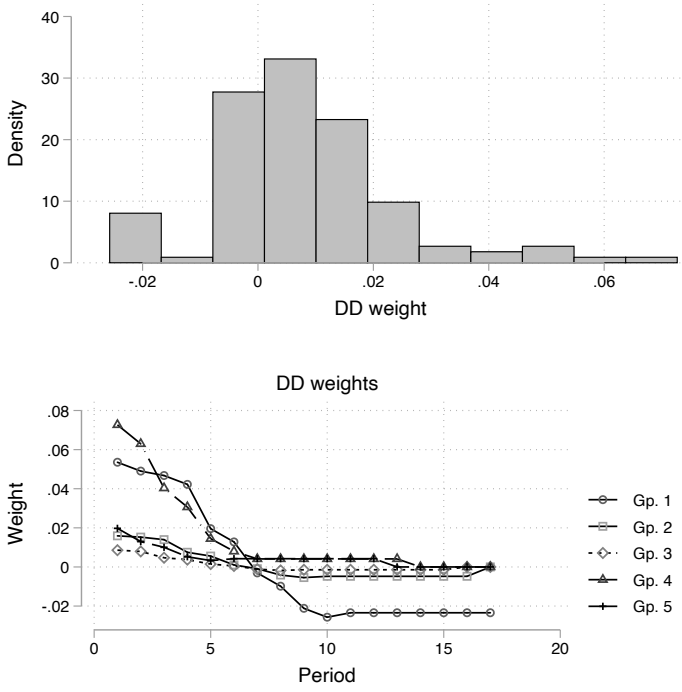
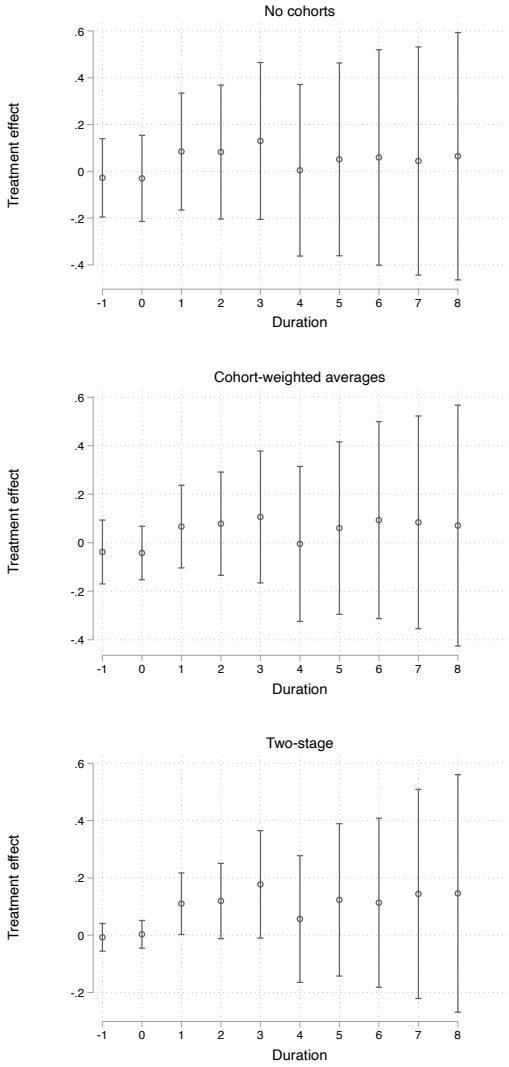


Table 2: Application estimates

All periods	Diff-in-diff	0.108
		(0.105)
	Aggregated	0.096
		(0.183)
	Two-stage	0.099
		(0.176)
	Two-stage (incl. treat×time in first stage)	0.098
		(0.169)
Nine periods	Aggregated	0.098
		(0.151)
	Stacked	0.102
		(0.144)
	Two-stage	0.102
		(0.144)

Notes: Standard errors clustered on state. Standard errors for aggregated estimators calculated using the delta method, assuming that the treated distribution of groups and periods is fixed. Two-stage estimates computed by estimating both equations simultaneously by GMM.

Figure 4: Application event studies



Notes: Means and 95% confidence intervals (based on state-clustered standard errors).

Appendix C: Event-study estimates

Table 3: Simulation event-study results

	Simulation 1			Simulation 2		
Period	Aggregated	No cohorts	Two-stage	Aggregated	No cohorts	Two-stage
-1	0.02	-0.42	0.02	0.00	-0.40	0.01
	(0.39)	(0.37)	(0.24)	(0.29)	(0.23)	(0.15)
0	0.00	-0.54	-0.01	0.00	-0.61	0.00
	(0.36)	(0.36)	(0.21)	(0.25)	(0.29)	(0.14)
1	1.24	0.64	1.24	1.02	0.31	1.01
	(0.62)	(0.62)	(0.59)	(0.47)	(0.46)	(0.44)
2	2.30	1.71	2.29	2.04	0.31	2.04
	(0.63)	(0.62)	(0.59)	(0.46)	(0.46)	(0.44)
3	4.08	3.53	4.07	3.53	2.99	3.53
	(0.61)	(0.60)	(0.58)	(0.46)	(0.48)	(0.44)
4	5.26	4.78	5.25	4.53	4.09	4.53
	(0.70)	(0.69)	(0.66)	(0.48)	(0.47)	(0.46)

Notes: Means and standard deviations from 250 simulations.

Table 4: Application event study estimates

Period	Aggregated	No cohorts	Two-stage
-1	-0.052	-0.05	-0.01
	(0.054)	(0.09)	(0.03)
0	-0.055	-0.03	-0.01
	(0.066)	(0.08)	(0.02)
1	0.048	0.05	0.09
	(0.089)	(0.12)	(0.06)
2	0.056	0.07	0.10
	(0.113)	(0.14)	(0.07)
3	0.100	0.12	0.16
	(0.145)	(0.17)	(0.10)
4	-0.002	0.03	0.04
	(0.173)	(0.18)	(0.12)
5	0.067	0.05	0.11
	(0.191)	(0.20)	(0.15)
6	0.109	0.05	0.09
	(0.218)	(0.23)	(0.16)
7	0.091	0.05	0.14
	(0.236)	(0.24)	(0.20)
8	0.092	0.07	0.14
	(0.264)	(0.26)	(0.22)
9	0.033	0.00	0.08
	(0.276)	(0.27)	(0.23)

Notes: Standard errors clustered on state. Standard errors for aggregated estimators calculated using the delta method, assuming that the treated distribution of groups and periods is fixed. Two-stage estimates computed by estimating both equations simultaneously by GMM.