ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Difference-in-differences with variation in treatment timing☆

Andrew Goodman-Bacon *

Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis, 90 Hennepin Ave, Minneapolis, MN 55401, USA
National Bureau of Economic Research, USA

ARTICLE INFO

Article history:
Received 19 January 2021
Received in revised form 19 January 2021
Accepted 17 March 2021
Available online xxxx

Keywords:
Difference-in-differences
Variation in treatment timing
Two-way fixed effects
Treatment effect heterogeneity

ABSTRACT

The canonical difference-in-differences (DD) estimator contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper shows that the two-way fixed effects estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls.

Published by Elsevier B.V.

1. Introduction

Difference-in-differences (DD) is both the most common and the oldest quasi-experimental research design, dating back to Snow's (1855) analysis of a London cholera outbreak.[1] A DD estimate is the difference between the change in outcomes before and after a treatment (difference one) in a treatment versus control group (difference two): $\left(\bar{y}^{POST}_{TREAT} - \bar{y}^{PRE}_{TREAT}\right) - \left(\bar{y}^{POST}_{CONTROL} - \bar{y}^{PRE}_{CONTROL}\right)$. That simple quantity also equals the estimated coefficient on the interaction of a treatment group dummy and a post-treatment period dummy in the following regression:

$$y_{it} = \gamma + \gamma_{i\cdot} TREAT_i + \gamma_{\cdot t} POST_t + \beta^{2x2} TREAT_i \times POST_t + u_{it}. \quad (1)$$

The elegance of DD makes it clear which comparisons generate the estimate, what leads to bias, and how to test the design. The expression in terms of sample means connects the regression to potential outcomes and shows that, under a common trends assumption, a two-group/two-period (2x2) DD identifies the average treatment effect on the treated. Almost all econometrics textbooks and survey articles describe this structure,[2] and recent methodological extensions build on it.[3]

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author at: Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis, 90 Hennepin Ave, Minneapolis, MN 55401, USA.
E-mail address: andrew.j.goodman-bacon@vanderbilt.edu.

1 A search from 2012 forward of nber.org, for example, yields 430 results for "difference-in-differences", 360 for "randomization" AND "experiment" AND "trial", and 277 for "regression discontinuity" OR "regression kink".

2 This includes Angrist and Krueger (1999), Angrist and Pischke (2009), Heckman et al. (1999), Meyer (1995), Cameron and Trivedi (2005) and Wooldridge (2010). Cunningham (2021) is an exception.

3 Inverse propensity score reweighting: Abadie (2005), double robust estimation: Sant'Anna and Zhao (2020), synthetic control: Abadie et al. (2010), changes-in-changes: Athey and Imbens (2006), quantile treatment effects: Callaway et al. (2018).

https://doi.org/10.1016/j.jeconom.2021.03.014
0304-4076/Published by Elsevier B.V.

Most DD applications diverge from this 2x2 set up though because treatments usually occur at different times.[4] Local governments change policy. Jurisdictions hand down legal rulings. Natural disasters strike across seasons. Firms lay off workers. In this case researchers estimate a regression with dummies for cross-sectional units ($\alpha_{i\cdot}$) and time periods ($\alpha_{\cdot t}$), and a treatment dummy (D_{it}):

$$y_{it} = \alpha_{i\cdot} + \alpha_{\cdot t} + \beta^{DD} D_{it} + e_{it}. \tag{2}$$

In contrast to our substantial understanding of canonical 2x2 DD, we know relatively little about the two-way fixed effects DD when treatment timing varies. We do not know precisely how it compares mean outcomes across groups.[5] We typically rely on general descriptions of the identifying assumption like "interventions must be as good as random, conditional on time and group fixed effects" (Bertrand et al., 2004, p. 250). We have limited understanding of the treatment effect parameter that regression DD identifies. Finally, we often cannot evaluate how and why alternative specifications change estimates.[6]

This paper shows that the two-way fixed effects DD estimator in (2) (TWFEDD) is a weighted average of all possible 2x2 DD estimators that compare timing groups to each other (the DD decomposition). Some use units treated at a particular time as the treatment group and untreated units as the control group. Some compare units treated at two different times, using the later-treated group as a control before its treatment begins and then the earlier-treated group as a control after its treatment begins. The weights on the 2x2 DDs are proportional to timing group sizes *and* the variance of the treatment dummy in each pair, which is highest for units treated in the middle of the panel.

I first use this DD decomposition to show that TWFEDD estimates a variance-weighted average of treatment effect parameters sometimes with "negative weights" (Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020).[7] When treatment effects do not change over time, TWFEDD yields a variance-weighted average of cross-group treatment effects and all weights are positive. Negative weights only arise when average treatment effects vary over time. The DD decomposition shows why: when already-treated units act as controls, changes in their outcomes are subtracted and these changes may include time-varying treatment effects. This does not imply a failure of the *design* in the sense of non-parallel trends in counterfactual outcomes, but it does suggest caution when using TWFE estimators to summarize treatment effects.

Next I use the DD decomposition to define "common trends" when one is interested in using TWFEDD to identify the variance-weighted treatment effect parameter. Each 2x2 DD relies on pairwise common trends in untreated potential outcomes so the overall assumption is an average of these terms using the variance-based decomposition weights. The extent to which a given timing group's differential trend biases the overall estimate equals the difference between the total weight on 2x2 DDs where it is the treatment group and the total weight on 2x2 DDs where it is the control group. Because units treated near the beginning or the end of the panel have the lowest treatment variance they can get *more* weight as controls than treatments. In designs without untreated units they always do.

Finally, I develop simple tools to describe the TWFEDD estimator and evaluate why estimates change across specifications.[8] Plotting the 2x2 DDs against their weights displays heterogeneity in the components of the weighted average and shows which terms and timing groups matter most. Summing the weights on the timing comparisons quantifies "how much" of the variation comes from timing (a common question in practice), and provides practical guidance on how well the TWFEDD estimator works compared to alternative estimators (Sun and Abraham, 2020; Borusyak and Jaravel, 2017; Callaway and Sant'Anna, 2020; Imai and Kim, 2021; Strezhnev, 2018; Ben-Michael et al., 2019). Comparing TWFEDD estimates across specifications in a Oaxaca-Blinder-Kitagawa decomposition measures how much of the change in the overall estimate comes from the 2x2 DDs (consistent with confounding or within-group heterogeneity), the weights (changing estimand), or the interaction of the two. Scattering the 2x2 DDs or the weights from different specifications show which specific terms drive these differences. I also provide the first detailed analysis of specifications with time-varying controls, which can address bias, but also changes the sources of identification to include comparisons between units with the same treatment but different covariates.

To demonstrate these methods I replicate Stevenson and Wolfers (2006), who study of the effect of unilateral divorce laws on female suicide rates. The TWFEDD estimates suggest that unilateral divorce leads to 3 fewer suicides per million

4 Half of the 93 DD papers published in 2014/2015 in 5 general interest or field journals had variation in timing.

5 Imai and Kim (2021) note "It is well known that the standard DiD estimator is numerically equivalent to the linear two-way fixed effects regression estimator if there are two time periods and the treatment is administered to some units only in the second time period. Unfortunately, this equivalence result does not generalize to the multi-period DiD design…Nevertheless, researchers often motivate the use of the two-way fixed effects estimator by referring to the DiD design (e.g., Angrist and Pischke, 2009)".

6 This often leads to sharp disagreements. See Neumark et al. (2014) on unit-specific linear trends, Lee and Solon (2011) on weighting and outcome transformations, and Shore-Sheppard (2009) on age-time fixed effects.

7 Early research in this area made specific observations about stylized specifications with no unit fixed effects (Bitler et al., 2003), or it provided simulation evidence (Meer and West, 2016). Recent research on the weighting of heterogeneous treatment effects does not provide this intuition. de Chaisemartin and D'Haultfœuille (2020, p. 2969) and Borusyak and Jaravel (2017, pp. 10–11), describe these same weights as coming from an auxiliary regression, noting that "a general characterization of [the weights] does not seem feasible". Athey and Imbens (2018) also decompose the DD estimator and develop design-based inference methods for this setting. Strezhnev (2018) expresses $\hat{\beta}^{DD}$ as an unweighted average of DD-type terms across pairs of observations and periods.

8 These methods can be implemented using the Stata command `bacondecomp` available on SSC (Goodman-Bacon et al., 2019).

women. More than a third of the identifying variation comes from treatment timing and the rest comes from comparisons to states whose reform status does not change during the sample period. Event-study estimates show that the treatment effects grow over time, though, which biases many of the timing comparisons. The TWFEDD estimate (−3.08) is therefore a misleading summary of the average post-treatment effect (about −5). Much of the sensitivity across specifications comes from changes in weights, or a small number of 2x2 DD's, and need not indicate bias.

My results show how and why the TWFEDD estimator can fail to identify interpretable treatment effect parameters and suggest that practitioners should be careful when relying on it in designs with treatment timing variation. Fortunately, recent research has developed simple flexible estimators that address the problems I describe (e.g. Callaway and Sant'Anna, 2020), enabling applied researchers to make better use of variation in treatment timing.

2. The difference-in-differences decomposition theorem

When units experience treatment at different times, one cannot estimate equation (1) because the post-period dummy is not defined for control observations. Nearly all work that exploits variation in treatment timing use the two-way fixed effects regression in Eq. (2) (Cameron and Trivedi, 2005 p. 738). Researchers clearly recognize that differences in *when* units received treatment contribute to identification, but have not been able to describe how these comparisons are made.[9] This section decomposes the TWFEDD estimator into a weighted average of simple 2x2 DD estimators.

Fig. 1 plots a simple data structure that includes treatment timing. Assume a balanced panel dataset with T periods (t) and N cross-sectional units (i) that belong to an early treatment group, k, which receives a binary treatment at $t_i = k$; a late treatment group, ℓ, which receives the binary treatment at $t_i = \ell > k$, or an untreated group, U, "treated" at $t_i = \infty$.

Throughout the paper I use "treatment group" or "timing group" to refer to collections of units either treated at the same time or not treated. I refer to units that do not receive treatment as "untreated" rather than "controls" because, while they obviously act as controls, treated units do, too. k will denote an earlier treated group and ℓ will denote a later treated group. Each timing group's sample share is $n_k \equiv \sum_i 1\{t_i = k\}/N$ and the share of time it spends treated is $\overline{D}_k \equiv \sum_t 1\{t \geq k\}/T$. I denote the sample mean of y_{it} for units treated at time t_b^* during the post period for treatment day t_a^* by: $\overline{y}_b^{POST(a)} \equiv \frac{1}{T-(a-1)} \sum_a^T \left[\frac{\sum_i y_{it} 1\{t_i=b\}}{\sum_i 1\{t_i=b\}} \right]$. ($\overline{y}_b^{PRE(a)}$ is defined similarly.)

By the Frisch–Waugh–Lovell theorem (Frisch and Waugh, 1933; Lovell, 1963), $\hat{\beta}^{DD}$ equals the univariate regression coefficient between y_{it} and the treatment dummy with unit and time means removed:

$$\frac{\hat{C}(y_{it}, \tilde{D}_{it})}{\hat{V}^D} = \frac{\frac{1}{NT}\sum_i \sum_t y_{it}\tilde{D}_{it}}{\frac{1}{NT}\sum_i \sum_t \tilde{D}_{it}^2}. \tag{3}$$

I denote grand means by $\overline{\overline{x}} = \frac{1}{NT}\sum_i \sum_t x_{it}$, and fixed-effects adjusted variables by $\tilde{x}_{it} = (x_{it} - \overline{x}_i) - (\overline{x}_t - \overline{\overline{x}})$.

One challenge in this setting has been to articulate how estimates of Eq. (2) compare the timing groups and times depicted in Fig. 1. We do, however, have clear intuition, for 2x2 estimators in which one group's treatment status changes and another's does not. (These are just 2x2 DD estimators, so without additional assumptions discussed in Section 2 they cannot be interpreted as causal estimands.) In the three-group case we could form four such designs estimable by Eq. (1) on subsamples of timing groups and time periods. Fig. 2 plots them.

Panels A and B show that if we consider only one of the two treatment groups, the TWFE estimator reduces to the canonical case comparing a treated to an untreated group:

$$\hat{\beta}_{jU}^{2x2} \equiv \left(\overline{y}_j^{POST(j)} - \overline{y}_j^{PRE(j)}\right) - \left(\overline{y}_U^{POST(j)} - \overline{y}_U^{PRE(j)}\right), \quad j = k, \ell. \tag{4}$$

Note that I use 2x2 to refer to two time windows (here *PRE* (j) and *POST* (j)) instead of only two time periods. If instead there were no untreated units, the two way fixed effects estimator would be identified only by the differential treatment timing between groups k and ℓ. For this case, panels C and D plot two clear 2x2 DDs based on sub-periods when only one timing group's treatment status changes. Before ℓ, the early units act as the treatment group because their treatment status changes, and later units act as controls during their pre-period. We compare outcomes between the window when treatment status varies, *MID*(k, ℓ), and timing group k's pre-period, *PRE* (k):

$$\hat{\beta}_{k\ell}^{2x2,k} \equiv \left(\overline{y}_k^{MID(k,\ell)} - \overline{y}_k^{PRE(k)}\right) - \left(\overline{y}_\ell^{MID(k,\ell)} - \overline{y}_\ell^{PRE(k)}\right). \tag{5}$$

The opposite situation, shown in panel D, arises after k when the later group changes treatment status but the early group does not. Later units act as the treatment group, early units act as controls, and we compare average outcomes between the periods *POST* (ℓ) and *MID* (k, ℓ):

$$\hat{\beta}_{k\ell}^{2x2,\ell} \equiv \left(\overline{y}_\ell^{POST(\ell)} - \overline{y}_\ell^{MID(k,\ell)}\right) - \left(\overline{y}_k^{POST(\ell)} - \overline{y}_k^{MID(k,\ell)}\right). \tag{6}$$

[9] Angrist and Pischke (2015), for example, lay out the canonical DD estimator in terms of means, but discuss regression DD with timing in general terms only, noting that there is "more than one…experiment" in this setting.

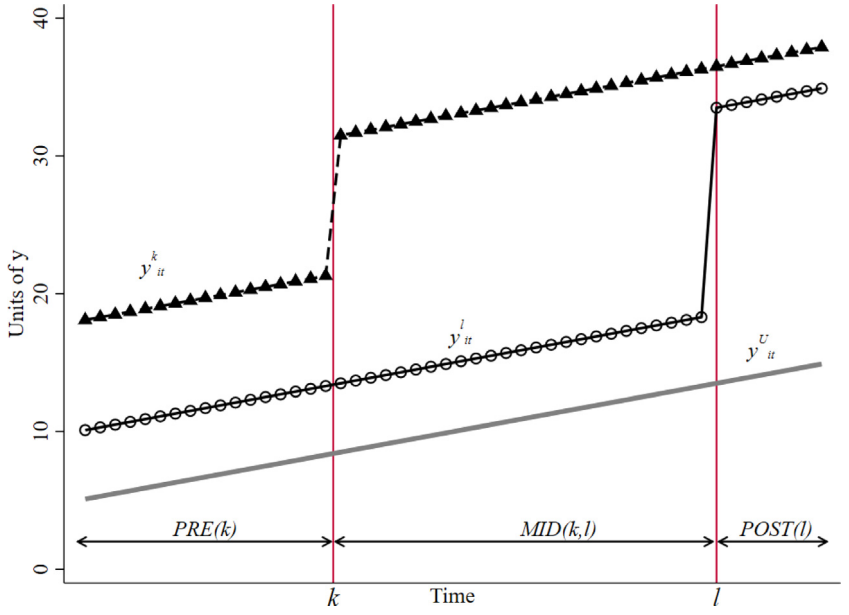

Fig. 1. Difference-in-Differences with variation in treatment Timing: Three groups. Notes: The figure plots outcomes in three timing groups: an untreated group, U; an early treatment group, k, which receives a binary treatment at $k = \frac{34}{100}T$; and a late treatment group, ℓ, which receives the binary treatment at $\ell = \frac{85}{100}T$. The x-axis notes the three sub-periods: the pre-period for timing group k, $[1, k-1]$, denoted by $PRE(k)$; the middle period when timing group k is treated and timing group ℓ is not, $[k, \ell-1]$, denoted by $MID(k,\ell)$; and the post-period for timing group ℓ, $[\ell, T]$, denoted by $POST(\ell)$. The treatment effect is 10 in timing group k and 15 in timing group ℓ.

The already-treated units in timing group k can serve as controls even though they are treated because treatment status does not change.

These simple DDs come from subsamples that relate to the full sample in two specific ways. First, each one uses a fraction of all NT observations. The treated/untreated DDs in (4) use two groups and all time periods, so their sample shares are $n_k + n_U$ and $n_\ell + n_U$. The timing DDs in (5) and (6) also use two groups but only some time periods. $\hat{\beta}^{2x2,k}_{k\ell}$ comes from timing group ℓ's pre-period so its share of all NT observations is $(n_k + n_\ell)\left(1 - \overline{D}_\ell\right)$. $\hat{\beta}^{2x2,\ell}_{k\ell}$ comes from timing group k's post-period so its share of all NT observations is $(n_k + n_\ell)\,\overline{D}_k$.

Second, each 2x2 DD is identified by how treatment varies in its subsample. The "amount" of identifying variation equals the variance of fixed-effects-adjusted D_{it} from its subsample:

$$\hat{V}^D_{jU} \equiv n_{jU}\left(1 - n_{jU}\right)\overline{D}_j\left(1 - \overline{D}_j\right), \quad j = k, \ell \tag{7}$$

$$\hat{V}^{D,k}_{k\ell} \equiv n_{k\ell}\left(1 - n_{k\ell}\right)\frac{\overline{D}_k - \overline{D}_\ell}{1 - \overline{D}_\ell}\frac{1 - \overline{D}_k}{1 - \overline{D}_\ell}, \tag{8}$$

$$\hat{V}^{D,\ell}_{k\ell} \equiv n_{k\ell}\left(1 - n_{k\ell}\right)\frac{\overline{D}_\ell}{\overline{D}_k}\frac{\overline{D}_k - \overline{D}_\ell}{\overline{D}_k}, \tag{9}$$

where $n_{ab} \equiv \frac{n_a}{n_a + n_b}$ is the relative size of timing groups in each pair. The first part of each pairwise variance measures how concentrated the timing groups are in the subsample. If n_{jU} equals zero or one, the variance goes to zero: there is either no treatment or no control group. The second part comes from *when* the treatment occurs in each subsample. The $\overline{D}$ terms equal the variance of D_{it} in each subsample's treatment group in its time window (thus the rescaling in (8) and (9)). If $\overline{D}_j$ equals zero or one the variance goes to zero: treatment does not vary over time.

My central result is that the TWFEDD estimator is an average of well-understood 2x2 DD estimators, like those plotted in Fig. 2, with weights based on subsample shares and the variances in (7)–(9):

Theorem 1 (*Difference-in-Differences Decomposition Theorem*)**.** *Assume that the data contain $k = 1, \ldots, K$ timing groups of units ordered by the time when they receive a binary treatment, $k \in (1, T]$. There may be one timing group, U, that includes units that never receive treatment. The OLS estimate, $\hat{\beta}^{DD}$, in a two-way fixed-effects regression (2) is a weighted average of*

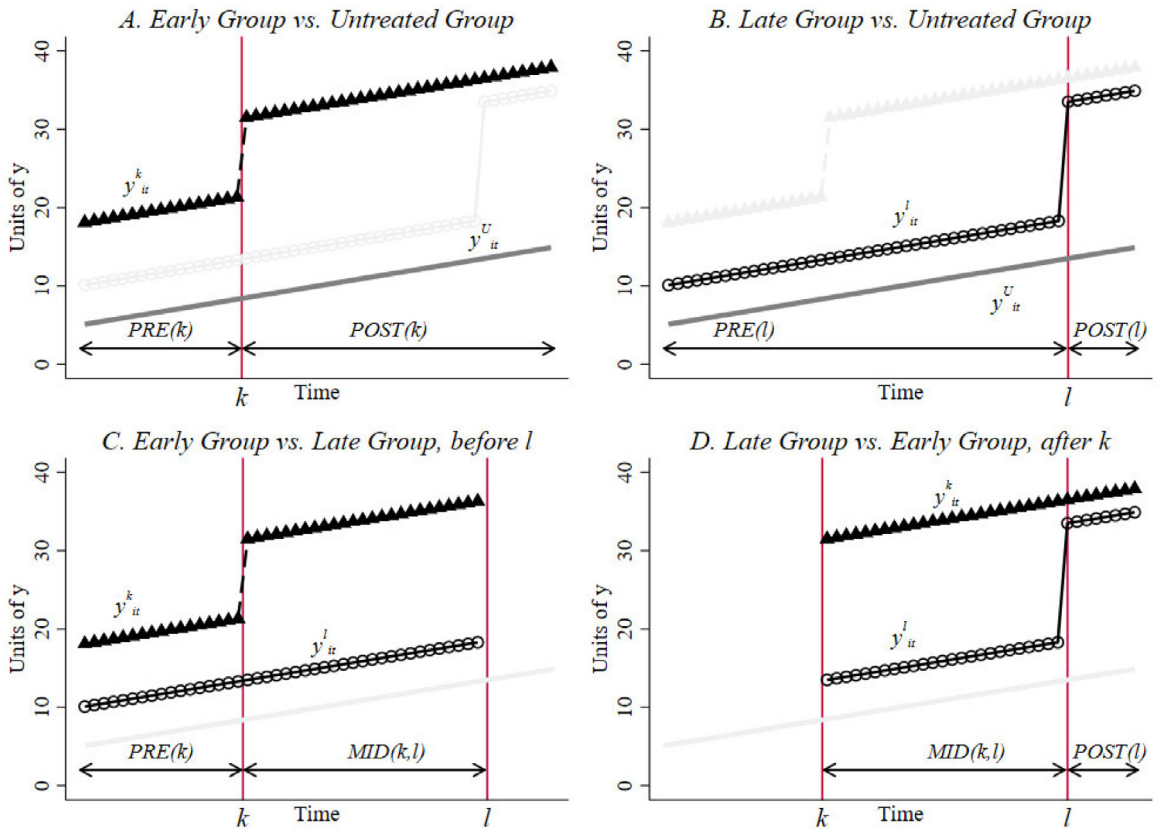

Fig. 2. The four simple (2x2) difference-in-differences estimates in the three group case. Notes: The figure plots outcomes for the subsamples that generate the four simple 2x2 difference-in-difference estimates in the three timing group case from Fig. 1. Each panel plots the data structure for one 2x2 DD. Panel A compares early treated units to untreated units ($\hat{\beta}^{DD}_{kU}$); panel B compares late treated units to untreated units ($\hat{\beta}^{DD}_{\ell U}$); panel C compares early treated units to late treated units during the late timing group's pre-period ($\hat{\beta}^{DD,k}_{k\ell}$); panel D compares late treated units to early treated units during the early timing group's post-period ($\hat{\beta}^{DD,\ell}_{k\ell}$). The treatment times mean that $\overline{D}_k = 0.67$ and $\overline{D}_\ell = 0.16$, so with equal group sizes, the decomposition weights on the 2x2 estimate from each panel are 0.365 for panel A, 0.222 for panel B, 0.278 for panel C, and 0.135 for panel D.

all possible two-by-two DD estimators.

$$\hat{\beta}^{DD} = \sum_{k \neq U} s_{kU}\, \hat{\beta}^{2x2}_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s^k_{k\ell}\, \hat{\beta}^{2x2,k}_{k\ell} + s^\ell_{k\ell}\, \hat{\beta}^{2x2,\ell}_{k\ell} \right]. \tag{10a}$$

where the 2x2 DD estimators are:

$$\hat{\beta}^{2x2}_{kU} \equiv \left(\overline{y}^{POST(k)}_k - \overline{y}^{PRE(k)}_k \right) - \left(\overline{y}^{POST(k)}_U - \overline{y}^{PRE(k)}_U \right), \tag{10b}$$

$$\hat{\beta}^{2x2,k}_{k\ell} \equiv \left(\overline{y}^{MID(k,\ell)}_k - \overline{y}^{PRE(k)}_k \right) - \left(\overline{y}^{MID(k,\ell)}_\ell - \overline{y}^{PRE(k)}_\ell \right), \tag{10c}$$

$$\hat{\beta}^{2x2,\ell}_{k\ell} \equiv \left(\overline{y}^{POST(\ell)}_\ell - \overline{y}^{MID(k,\ell)}_\ell \right) - \left(\overline{y}^{POST(\ell)}_k - \overline{y}^{MID(k,\ell)}_k \right). \tag{10d}$$

The weights are:

$$s_{kU} = \frac{(n_k + n_U)^2 \overbrace{n_{kU}(1 - n_{kU})\overline{D}_k(1 - \overline{D}_k)}^{\hat{V}^D_{kU}}}{\hat{V}^D}, \tag{10e}$$

$$s^k_{k\ell} = \frac{\left((n_k + n_\ell)\left(1 - \overline{D}_\ell\right)\right)^2 \overbrace{n_{k\ell}\left(1 - n_{k\ell}\right)\frac{\overline{D}_k - \overline{D}_\ell}{1 - \overline{D}_\ell}\frac{1 - \overline{D}_k}{1 - \overline{D}_\ell}}^{\hat{V}^{D,k}_{k\ell}}}{\hat{V}^D}, \tag{10f}$$

$$s^{\ell}_{k\ell} = \frac{\left((n_k + n_\ell)\,\overline{D}_k\right)^2 \overbrace{n_{k\ell}\,(1 - n_{k\ell})\,\frac{\overline{D}_\ell}{\overline{D}_k}\frac{\overline{D}_k - \overline{D}_\ell}{\overline{D}_k}}^{\hat{V}^{D,\ell}_{k\ell}}}{\hat{V}^{D}}. \tag{10g}$$

and $\sum_{k\neq U} s_{kU} + \sum_{k\neq U}\sum_{\ell>k}\left[s^{k}_{k\ell} + s^{\ell}_{k\ell}\right] = 1$.

Proof. See Appendix A.[10]

Theorem 1 completely describes the sources of identifying variation in a TWFEDD estimator and their importance. With K timing groups, one could form $K^2 - K$ "timing-only" estimates that either compare an earlier- to a later-treated timing group ($\hat{\beta}^{2x2,k}_{k\ell}$) or a later- to earlier-treated timing group ($\hat{\beta}^{2x2,\ell}_{k\ell}$). With an untreated group, one could form K 2x2 DDs that compare one timing group to the untreated group ($\hat{\beta}^{2x2}_{kU}$). Therefore, with K timing groups and one untreated group, the DD estimator comes from K^2 distinct 2x2 DDs.[11]

The weights on each 2x2 DD combine the absolute size of the subsample and the variance of the fixed-effects-adjusted treatment variable in the subsample.[12] The first part is the size of the subsample squared. The second part of each weight is the subsample variance from Eqs. (7)–(9), which comes from the relative size of the treatment and control groups and the timing of treatment. The variance is larger when the two timing groups are closer in size ($n_{kU} \approx 0.5$) and when treatment occurs closer to the middle of the time window ($\overline{D}_k$, $\frac{\overline{D}_k - \overline{D}_\ell}{1-\overline{D}_\ell}$, or $\frac{\overline{D}_\ell}{\overline{D}_k}$ are close to 0.5).

In Fig. 2, the 2x2 DDs with timing group k as the treatment group get the most weight. I assume equal group sizes so that the weights are completely determined by timing. I set k and ℓ so that $\overline{D}_k = 0.66$ and $\overline{D}_\ell = 0.16$. For treated/untreated DDs, $s_{kU} > s_{\ell U}$ because k is closer to the middle of the panel than ℓ is, which means: $\overline{D}_k\left(1-\overline{D}_k\right) = 0.22 > 0.13 = \overline{D}_\ell\left(1-\overline{D}_\ell\right)$. This is also true for the timing-only 2x2 DDs. Timing group k's treatment share during the $PRE(\ell)$ period is $\frac{\overline{D}_k - \overline{D}_\ell}{1-\overline{D}_\ell} = \frac{0.66-0.16}{0.84} = 0.59$, but timing group ℓ's pre-period accounts for $1 - \overline{D}_\ell = 0.84$ share of the periods. Timing group ℓ's treatment share during the $POST(k)$ period, on the other hand, is $\frac{\overline{D}_\ell}{\overline{D}_k} = \frac{0.16}{0.66} = 0.24$, and timing group k's post-period accounts for $\overline{D}_k = 0.66$ share of the periods. Therefore, $s^{k}_{k\ell} > s^{\ell}_{k\ell}$ because $\hat{\beta}^{k}_{k\ell}$ has a higher variance from treatment timing alone and it uses more data: $\left(1-\overline{D}_\ell\right)^2 \frac{\overline{D}_k - \overline{D}_\ell}{1-\overline{D}_\ell}\frac{1-\overline{D}_k}{1-\overline{D}_\ell} = 0.17 > 0.08 = \overline{D}_k{}^2 \frac{\overline{D}_\ell}{\overline{D}_k}\frac{\overline{D}_k - \overline{D}_\ell}{\overline{D}_k}$. Scaling by the overall variance of $\tilde{D}_{it}$ shows that the weights are $\left\{s_{kU}, s_{\ell U}, s^{k}_{k\ell}, s^{\ell}_{k\ell}\right\} = \{0.37, 0.22, 0.28, 0.13\}$. $\hat{\beta}^{DD}$ equals 11.75 even though the average of the treatment effects is 12.5 because TWFEDD puts more weight on the early group which in this example has a smaller effect (10 versus 15).

Theorem 1 implies that changing the number or spacing of time periods changes the weights (in addition to potentially changing the 2x2 DDs). Imagine adding T periods to the end of Fig. 2. In that case, $\overline{D}_k = 0.83$ and $\overline{D}_\ell = 0.58$ and timing group ℓ is treated closer to the middle of the panel than timing group k. The weights change to $\left\{s_{kU}, s_{\ell U}, s^{k}_{k\ell}, s^{\ell}_{k\ell}\right\} = \{0.25, 0.43, 0.07, 0.25\}$. 2x2 DDs in which timing group ℓ is the treatment group get twice as much weight in this case; 68 percent with $2T$ periods versus 35 percent with T periods. In this case $\hat{\beta}^{DD}$ equals 13.4. Therefore, panel length alone can change TWFEDD estimates substantially even if the 2x2 DDs themselves are constant.

Theorem 1 also shows how DD compares two treated groups. A two-group "timing-only" estimator is itself a weighted average of the 2x2 DDs plotted in panels C and D of Fig. 2:

$$\hat{\beta}^{2x2}_{k\ell} \equiv \overbrace{\frac{\left(1-\overline{D}_\ell\right)^2 \hat{V}^{D,k}_{k\ell}}{\left(1-\overline{D}_\ell\right)^2 \hat{V}^{D,k}_{k\ell} + \overline{D}_k{}^2 \hat{V}^{D,\ell}_{k\ell}}}^{\mu_{k\ell}} \hat{\beta}^{2x2,k}_{k\ell} + \overbrace{\frac{\overline{D}_k{}^2 \hat{V}^{D,\ell}_{k\ell}}{\left(1-\overline{D}_\ell\right)^2 \hat{V}^{D,k}_{k\ell} + \overline{D}_k{}^2 \hat{V}^{D,\ell}_{k\ell}}}^{1-\mu_{k\ell}} \hat{\beta}^{2x2,\ell}_{k\ell}. \tag{11}$$

Both timing groups serve as controls for each other during periods when their treatment status does not change, and the weight assigned to the 2x2 terms comes from how large is their subsample and how large is their treatment variance.

10 Appendices available at: http://goodman-bacon.com/pdfs/ddtiming_appendix.pdf.

11 Units that are treated before $t = 1$, "always-treated units", enter the DD decomposition just like untreated units in the sense that they only ever serve as controls in terms like $\hat{\beta}^{2x2}_{kU}$ with weights like s_{kU}. In Section 2 I discuss the theoretical issues with always-treated units.

12 Many other least-squares estimators weight heterogeneity this way. A univariate regression coefficient equals an average of coefficients in mutually exclusive (and demeaned) subsamples weighted by size and the subsample x -variance:

$$\hat{\alpha} = \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_A (y-\overline{y})(x-\overline{x}) + \sum_B (y-\overline{y})(x-\overline{x})}{\sum_i (x-\overline{x})^2} = \frac{n_A s^{A}_{xy} + n_B s^{B}_{xy}}{s^2_{xx}} = \frac{n_A s^{2,A}_{xx}}{s^2_{xx}}\hat{\alpha}_A + \frac{n_B s^{2,B}_{xx}}{s^2_{xx}}\hat{\alpha}_B$$

Similarly, two-stage least squares uses samples sizes and variances to "efficiently combine alternative Wald estimates" (Angrist, 1991). Gibbons et al. (2018) show an analogous weighting formula for one-way fixed effects. Panel data provide another well-known example: a pooled regression coefficients equals a variance-weighted average of two distinct estimators that each use less information: the between estimator for subsample means, and the within estimator for deviations from subsample means.

In (11), $\mu_{k\ell}$ simplifies to $\frac{1-\bar{D}_k}{1-(\bar{D}_k-\bar{D}_\ell)}$, which falls as $\bar{D}_k$ gets closer to one (t_k^* gets closer to the first time period). In other words, the timing group treated closest to the middle of the panel gets more weight. In the three group example $\mu_{k\ell} = 0.34/0.5 = 0.68$.[13]

Theorem 1 is not the only way to decompose the TWFEDD estimator. Strezhnev (2018), Eq. (15) decomposes $\hat{\beta}^{DD}$ into an unweighted average of comparisons between all units and all time periods so that the weights across types of comparisons (2x2 DDs) are only implicitly defined. Athey and Imbens (2018, equation 4.3) decompose $\hat{\beta}^{DD}$ into terms representing causal effects over different time-horizons. By grouping 2x2 terms according to the identifying variation (pre/post, treatment/control) that unites them, my "group-level" decomposition yields clear definitions for the weights and connects them to the features of the OLS estimation method. See Appendix D for more details on the relationship between decompositions.

The DD Decomposition theorem is different from related results that decompose the TWFEDD *estimand* into a weighted average of treatment effect parameters with potentially negative weights (Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020). Theorem 1 expresses the TWFEDD *estimator* as a weighted average of simpler estimators with strictly positive weights that sum to 1. Section 2 shows how these results are connected.

3. Theory: What parameter does DD identify and under what assumptions?

Theorem 1 relates the regression DD coefficient to sample averages, which makes it simple to analyze its statistical properties by writing $\hat{\beta}^{DD}$ in terms of potential outcomes (Holland, 1986; Rubin, 1974). Define $Y_{it}(k)$ as the outcome of unit i in period t when it is treated at $t_i = k$, and use $Y_{it}(t_i)$ to denote treated potential outcomes under unit i's actual treatment date. $Y_{it}(0)$ is the untreated potential outcome. If $t < t_i$ then $Y_{it}(t_i) = Y_{it}(0)$. The observed outcome is $y_{it} = D_{it}Y_{it}(t_i) + (1 - D_{it})Y_{it}(0)$. Following Callaway and Sant'Anna (2020) define the ATT for timing group k at time $\tau \geq k$ (the "group-time average treatment effect"): $ATT_k(\tau) \equiv E\left[Y_{i\tau}\left(t_k^*\right) - Y_{i\tau}(0) | t_i = k\right]$. Because TWFEDD averages outcomes in pre- and post-treatment windows, I define the average $ATT_k(\tau)$ in a date range W (with T_W periods):

$$ATT_k(W) \equiv \frac{1}{T_W}\sum_{t\in W} E\left[Y_{it}(k) - Y_{it}(0) | t_i = k\right]. \tag{12}$$

In practice, W will represent post-treatment windows that appear in the 2x2 components. Finally, define the difference over time in average untreated potential outcomes as:

$$\Delta Y_k^0(W_1, W_0) \equiv \frac{1}{T_{W_1}}\sum_{t\in W_1} E\left[Y_{it}(0) | t_i = k\right] - \frac{1}{T_{W_0}}\sum_{t\in W_0} E\left[Y_{it}(0) | t_i = k\right]. \tag{13}$$

Applying this notation to the 2x2 DDs in Eqs. (4)–(6), adding and subtracting average untreated outcomes for the treatment group yields the familiar result that (the probability limit of) each 2x2 DD equals an ATT plus bias from differential trends:

$$\beta_{kU}^{2x2} = ATT_k(POST(k)) + \left[\Delta Y_k^0(POST(k), PRE(k)) - \Delta Y_U^0(POST(k), PRE(k))\right] \tag{14a}$$

$$\beta_{k\ell}^{2x2,k} = ATT_k(MID(k,\ell)) + \left[\Delta Y_k^0(MID(k,\ell), PRE(k)) - \Delta Y_\ell^0(MID(k,\ell), PRE(k))\right] \tag{14b}$$

$$\begin{aligned}\beta_{k\ell}^{2x2,\ell} = {} & ATT_\ell(POST(\ell)) + \left[\Delta Y_\ell^0(POST(\ell), MID(k,\ell)) - \Delta Y_k^0(POST(\ell), MID(k,\ell))\right] \\ & - \left[ATT_k(POST(\ell)) - ATT_k(MID(k,\ell))\right].\end{aligned} \tag{14c}$$

Note that the definition of common trends in (14a) and (14b) involves only untreated potential outcomes, but in (14c) identification of $ATT_\ell(POST(\ell))$ additionally involves changes in average treatment effects in the already-treated control group.

Substituting Eqs. (14a)–(14c) into the DD decomposition theorem expresses the probability limit of the TWFEDD estimator (assuming that T is fixed and N grows) in terms of potential outcomes and separates the estimand from the identifying assumptions:

$$\underset{N\to\infty}{plim}\,\hat{\beta}^{DD} = \beta^{DD} = VWATT + VWCT - \Delta ATT. \tag{15}$$

The first term in (15) is the interpretable causal parameter that TWFEDD can estimate, which I call the "variance-weighted average treatment effect on the treated" (VWATT):

$$VWATT \equiv \sum_{k\neq U}\sigma_{kU}ATT_k(POST(k)) + \sum_{k\neq U}\sum_{\ell>k}\left[\sigma_{k\ell}^k ATT_k(MID(k,\ell)) + \sigma_{k\ell}^\ell ATT_\ell(POST(\ell))\right]. \tag{15a}$$

[13] Two recent papers use two-group timing-only estimators. Malkova (2017) studies a maternity benefit policy in the Soviet Union and Goodman (2017) studies high school math mandates. Both papers show differences between early and late groups before the reform, $PRE(k)$, during the period when treatment status differs, $MID(k,\ell)$, and in the period after both have implemented reforms, $POST(\ell)$.

The σ terms are probability limits of the weights in (10a).[14] VWATT is a positively weighted average of ATTs for the treatment groups *and* post-periods across the 2x2 DDs that make up $\hat{\beta}^{DD}$.

The second term, which I call "variance-weighted common trends" (VWCT) generalizes common trends to a setting with timing variation:

$$\begin{aligned} VWCT \equiv & \sum_{k \neq U} \sigma_{kU} \left[\Delta Y_k^0 \left(POST\left(k\right), PRE\left(k\right) \right) - \Delta Y_U^0 \left(POST\left(k\right), PRE\left(k\right) \right) \right] \\ & + \sum_{k \neq U} \sum_{\ell > k} \left[\sigma_{k\ell}^k \left\{ \Delta Y_k^0 \left(MID\left(k, \ell\right), PRE\left(k\right) \right) - \Delta Y_\ell^0 \left(MID\left(k, \ell\right), PRE\left(k\right) \right) \right\} \right. \\ & \left. + \sigma_{k\ell}^\ell \left\{ \Delta Y_\ell^0 \left(POST\left(\ell\right), MID\left(k, \ell\right) \right) - \Delta Y_k^0 \left(POST\left(\ell\right), MID\left(k, \ell\right) \right) \right\} \right]. \end{aligned} \tag{15b}$$

Like VWATT, VWCT is an average of the difference in counterfactual trends between pairs of timing groups and different time periods using the weights from the decomposition theorem. It captures the way that differential trends map to bias in (10a). Note that one timing group's counterfactual trend affects many 2x2 DDs by different amounts and in different directions depending on whether it is the treatment or control group. While the mapping from trends to bias in a given 2x2 is clear, this result for a design with timing is new.

The last term in (15) equals a weighted sum of the *change* in treatment effects within each timing group's before and after a later treatment time:

$$\Delta ATT \equiv \sum_{k \neq U} \sum_{\ell > k} \sigma_{k\ell}^\ell \left[ATT_k \left(POST\left(\ell\right) \right) - ATT_k \left(MID\left(k, \ell\right) \right) \right]. \tag{15c}$$

Because the 2x2 estimators in Eq. (14c) already-treated groups as controls, they subtract average changes in their untreated outcomes *and* their treatment effects. ΔATT equals zero if average treatment effects are constant, but when they are not, Eq. (15c) defines the resulting bias relative to $VWATT$ even when $VWCT = 0$. ΔATT is the source of the negative weights discussed in Borusyak and Jaravel (2017) and de Chaisemartin and D'Haultfœuille (2020). This does not mean that the research design is invalid. In this case specifications such as an event-study (Jacobson et al., 1993), "stacked DD" (Cengiz et al., 2019; Deshpande and Li, 2019; Fadlon and Nielsen, 2015), or reweighting estimators (Callaway and Sant'Anna, 2020) may be more appropriate.[15]

Units that are treated throughout the sample can only ever act as controls (in fact they enter into the decomposition theorem exactly like never-treated units), so if their treatment effects are changing during the sample periods they will also contribute to ΔATT. The form of the weights suggests that changes in the treatment effects for always-treated units may dominate ΔATT. 2x2 DDs in which always-treated units are the control group use all time periods (as opposed to the smaller windows used in $\hat{\beta}_{k\ell}^\ell$), so they get higher weight in (10a). If their treatment effects are changing they can substantially bias TWFEDD away from VWATT.

3.1. Interpreting the TWFEDD estimand

When the treatment effect is a constant, $ATT_k\left(W\right) = ATT$, $\Delta ATT = 0$, and $VWATT = ATT$. The rest of this section assumes that VWCT = 0 and discusses how to interpret VWATT under different forms of treatment effect heterogeneity.

3.1.1. Effects that vary across units but not over time

If treatment effects are constant over time but vary across units, then $ATT_k\left(W\right) = ATT_k$ and we still have $\Delta ATT = 0$. In this case DD identifies:

$$VWATT = \sum_{k \neq U} ATT_k \left[\overbrace{\sigma_{kU} + \sum_{j=1}^{k-1} \sigma_{jk}^k + \sum_{j=k+1}^{K} \sigma_{kj}^k}^{\equiv w_k^T} \right]. \tag{16}$$

[14] Note that a DD estimator is not consistent if T gets large because the permanently turned on treatment dummy becomes collinear with the unit fixed effects ($\frac{X'X}{T}$ does not converge to a positive definite matrix). Asymptotics with respect to T require the time dimension to grow in both directions (see Perron, 2006).

[15] Recent DD research comes to related conclusions about DD with timing, but does not describe the full estimator as in Eq. (15). Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfœuille (2020), and Sun and Abraham (2020) begin by imposing pairwise common trends ($VWCT = 0$), and then incorporating ΔATT into the DD estimand. The structure of the decomposition theorem, however, suggests that we should think of ΔATT as a source of bias because it arises from the way equation (2) forms "the" control group. This distinction, made clear in Eq. (15), separates a causal estimand ($VWATT$) from clearly defined identifying assumptions ($VWCT = 0$ and $\Delta ATT = 0$). This follows from at least two related precedents. de Chaisemartin and D'Haultfœuille (2018, p. 5), prove identification of dose–response DD models under an assumption on the treatment effects: "the average effect of going from 0 to d units of treatment among units with $D(0) = d$ is stable over time". Treatment effect homogeneity ensures an estimand with no negative weights. Similarly, the monotonicity assumption in Imbens and Angrist (1994) ensures that the local average treatment effect does not have negative weights.

VWATT weights together the group-specific ATTs not by sample shares, but by a function of sample shares and treatment variance. The weights in (16) are equal to the sum of the decomposition weights for all the terms in which timing group k acts as the treatment group, defined as w_k^T.

In general, $w_k^T \neq n_k^*$, so *VWATT* does not equal the sample *ATT*.[16] Neither are the weights proportional to the share of time each unit spends under treatment, so *VWATT* also does not equal the effect in the average treated period. The extent to which *VWATT* differs from the *ATT* depends on the relationship between treatment effect heterogeneity and treatment timing in a given sample. For example, a Roy model of selection on gains implies that treatment rolls out first to units with the largest effects. Site selection in experimental evaluations of training programs (Joseph Hotz et al., 2005) and energy conservation programs (Allcott, 2015) match this pattern. In this case, regression DD underestimates the sample-weighted *ATT* if treatment rolls out in the first half of the sample and overestimates it if treatment rolls out in the second half. The opposite conclusions follow from "reverse Roy" selection where units with the smallest effects select treatment first, which describes the take up of housing vouchers (Chyn, 2018) and charter school applications (Walters, 2018). Both the model of treatment allocation and characteristics of the sample matter for interpretation.

An easy way to gauge whether *VWATT* differs from a sample-weighted *ATT* is to scatter the weights from (16), w_k^T, against each timing group's sample share among the treated, $\frac{n_k}{1-n_U}$. These two may be close if there is little variation in treatment timing or if one timing group is very large. Conversely, weighting matters less if the ATT_k's are similar, which one can evaluate by aggregating each timing group's 2x2 DD estimates from the decomposition theorem. Finally, one could directly compare TWFEDD to estimators that target a particular parameter of interest. Several alternative estimators give differently weighted averages of *ATT*s (Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020).

3.1.2. Effects that vary over time but not across units

Time-varying treatment effects shape Eq. (15) in two ways. First, they generate heterogeneity across the 2x2 DDs that average over different post-treatment windows and up-weight short-run effects most likely to appear in the small windows between timing groups. These features typically make *VWATT* different than the sample *ATT*. Second, time-varying effects bias estimates away from *VWATT* because $\Delta ATT \neq 0$. Eqs. (14b) and (14c) show that common trends in counterfactual outcomes leaves one set of timing terms biased ($\hat{\beta}_{k\ell}^{2x2,\ell}$), while common trends between counterfactual and treated outcomes leaves the other set biased ($\hat{\beta}_{k\ell}^{2x2,k}$).

To illustrate this point, Fig. 3 plots a case where counterfactual outcomes are identical, but the treatment effect is a linear trend-break, $Y_{it}(t_i) = Y_{it}(0) + \phi \cdot (t - t_i + 1)$ (see Meer and West, 2016). $\hat{\beta}_{k\ell}^{2x2,k}$ uses timing group ℓ as a control group during its pre-period and identifies the *ATT* during the middle window in which treatment status varies: $ATT(MID\,(k,\ell)) = \phi\frac{(\ell-(k-1))}{2}$. $\hat{\beta}_{k\ell}^{2x2,\ell}$ however, is biased for $ATT(POST(\ell))$ because the control group (k) experiences a trend in outcomes due to its growing treatment effect[17]:

$$\hat{\beta}_{k\ell}^{2x2,\ell} = \overbrace{ATT_\ell\,(POST(\ell))}^{\phi\frac{(T-(\ell-1))}{2}} - \overbrace{\phi\frac{(T-(k-1))}{2}}^{\Delta ATT/(1-\mu_{k\ell})} = \phi\frac{(k-\ell)}{2} \leq 0. \tag{17}$$

This bias feeds through to $\beta_{k\ell}^{2x2}$ according to the relative weight on the 2x2 terms:

$$\hat{\beta}_{k\ell}^{2x2} = \phi\frac{\left[\left(\sigma_{k\ell}^{k} - \sigma_{k\ell}^{\ell}\right)(\ell-k)+1\right]}{2}. \tag{18}$$

The entire two-group timing estimate can be wrong signed if there is sufficiently more weight on $\hat{\beta}_{k\ell}^{2x2,\ell}$ than $\hat{\beta}_{k\ell}^{2x2,k}$ (i.e. $\sigma_{k\ell}^{\ell} > \sigma_{k\ell}^{k}$). Summarizing time-varying effects using Eq. (2) yields estimates that are too small or even wrong-signed, and should not be used to judge the meaning or plausibility of effect sizes.[18]

---

16 Sun and Abraham (2020), Borusyak and Jaravel (2017), Chernozhukov et al. (2013), de Chaisemartin and D'Haultfœuille (2020), Gibbons et al. (2018) and Wooldridge (2005) all make a similar observation. The DD decomposition theorem, provides a new solution for the relevant weights.

17 The average of the effects for timing group k during any set of positive event-times, $t - t_i$, is just ϕ times the average event-time. The $MID(k,\ell)$ period contains event-times 0 through $\ell-(k-1)$ and the $POST(\ell)$ period contains event-times $\ell-(k-1)$ through $T-(k-1)$, so we have:

$$ATT_k\,(MID\,(k,\ell)) = \phi\frac{(\ell-k)\,(\ell-(k-1))}{2\,(\ell-k)} = \phi\frac{\ell-(k-1)}{2},$$

$$ATT_k\,(POST\,(\ell)) = \phi\,(\ell-k) + \phi\frac{T-\ell+2}{2},$$

and the difference, which appears in the identifying assumption in (17) equals:

$$ATT_k\,(POST\,(\ell)) - ATT_k\,(MID\,(k,\ell)) = \phi\,(\ell-k) + \phi\frac{T-\ell+2}{2} - \phi\frac{\ell-(k-1)}{2} = \frac{\phi}{2}\,(T-(k-1)).$$

Outcomes in group ℓ actually fall on average relative to group k, which makes the DD estimate negative even when all treatment effects are positive.

18 Borusyak and Jaravel (2017) show that common, linear trends, in the post- *and* pre- periods cannot be estimated in this design. The decomposition theorem shows why: timing groups act as controls for each other, so permanent common trends difference out. This is not a meaningful limitation

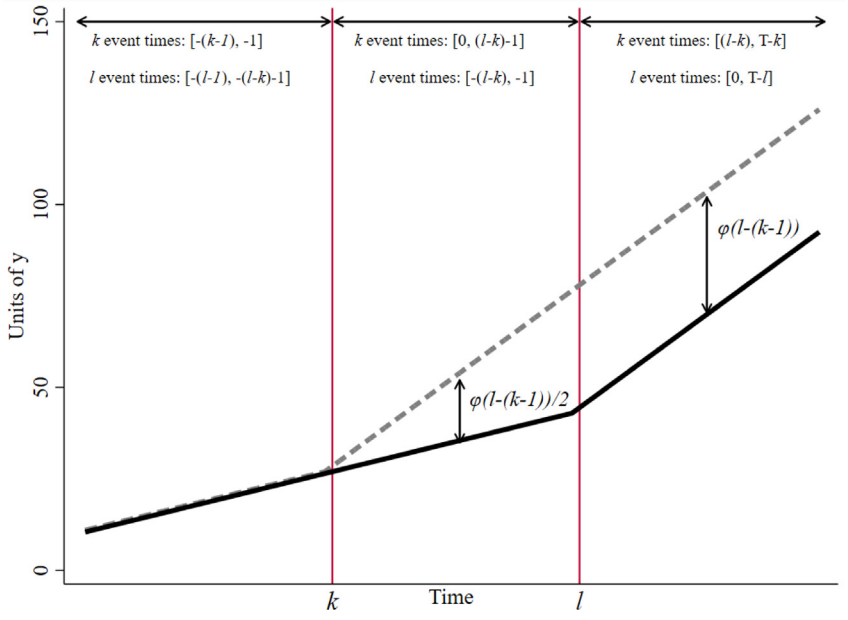

Fig. 3. Difference-in-Differences estimates with variation in timing are biased when treatment effects vary over time. Notes: The figure plots a stylized example of a timing-only DD set up with a treatment effect that is a trend-break rather than a level shift (see Meer and West, 2016). The trend-break effect equals $\phi \cdot (t - t_i + 1)$. The top of the figure notes which event-times lie in the $PRE(k)$, $MID(k, \ell)$, and $POST(\ell)$ periods for each unit. The figure also notes the average difference between timing groups in each of these periods. In the $MID(k, \ell)$ period, outcomes differ by $\frac{\phi}{2}(\ell - (k-1))$ on average. In the $POST(\ell)$ period, however, outcomes had already been growing in the early group for $\ell - k$ periods, and so they differ by $\phi(\ell - (k-1))$ on average. The 2x2 DD that compares the later-treated group to the earlier-treated group is biased and, in the linear trend-break case, weakly negative despite a positive and growing treatment effect.

Note that this bias is specific to a single-coefficient specification. More flexible event-study specifications may not suffer from this problem (although see Proposition 2 in Sun and Abraham, 2020). Fadlon and Nielsen (2015) and Deshpande and Li (2019) match treated units with controls that receive treatment a given amount of time later and gives an average of $\hat{\beta}^{2x2,k}_{k\ell}$ terms with a fixed post-period (see similar proposals in Sun and Abraham, 2020; Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020). Callaway and Sant'Anna (2020) discuss how to aggregate heterogeneous treatment effects and develop a reweighting estimator to do so.

3.2. What is the identifying assumption for VWATT?

The preceding analysis maintained the assumption of *equal* counterfactual trends across timing groups, but (15) shows that when $\Delta ATT = 0$ identification of *VWATT* only requires $VWCT = 0$. Assuming linear average untreated potential outcome trends ($\overline{Y}_{kt}(0) - \overline{Y}_{kt-1}(0) = \Delta Y^0_k$ for all t) leads to a convenient and intuitive approximation to *VWCT* as an average of each timing group's average trend in $Y(0)$ weighted by the difference between its weight as a treatment group (w^T_k from Eq. (16)) and a similar term measuring its weight as a control group[19]:

$$\begin{aligned} VWCT &\approx \sum_{k \neq U} \Delta Y^0_k \left[\left(\sigma_{kU} + \sum_{j=1}^{k-1} \sigma^k_{jk} + \sum_{j=k+1}^{K} \sigma^k_{kj} \right) - \left(\sum_{j=1}^{k-1} \sigma^j_{jk} + \sum_{j=k+1}^{K} \sigma^j_{kj} \right) \right] - \Delta Y^0_U \sum_{k \neq U} \sigma_{kU} \\ &= \sum_k \Delta Y^0_k \left[w^T_k - w^C_k \right]. \end{aligned} \tag{19}$$

Eq. (19) generalizes the definition of common trends to the timing case and shows how a given timing group's counterfactual trend biases the overall estimate. To illustrate, assume there is a positive differential trend in timing group k only: $\Delta Y^0_k > 0$. This will bias $\hat{\beta}^{2x2}_{kU}$ by ΔY^0_k which gets a weight of σ_{kU} in the full estimate. In 2x2 DDs base on timing,

for treatment effect estimation, though, because "effects" must occur after treatment. Job displacement provides a clear example Jacobson et al. (1993), Krolikowski (2017). Comparisons based on displacement timing cannot identify whether all displaced workers have a permanently different earnings trajectory than never displaced workers (the unidentified linear component), but they can identify changes in the time-path of earnings around the displacement event (the treatment effect).

19 Linearly trending unobservables lead to larger bias in 2x2 DDs that use more periods. In the linear case, differences in the magnitude of the bias cancel out across each group's "treatment" and "control" terms, and Eq. (19) holds.

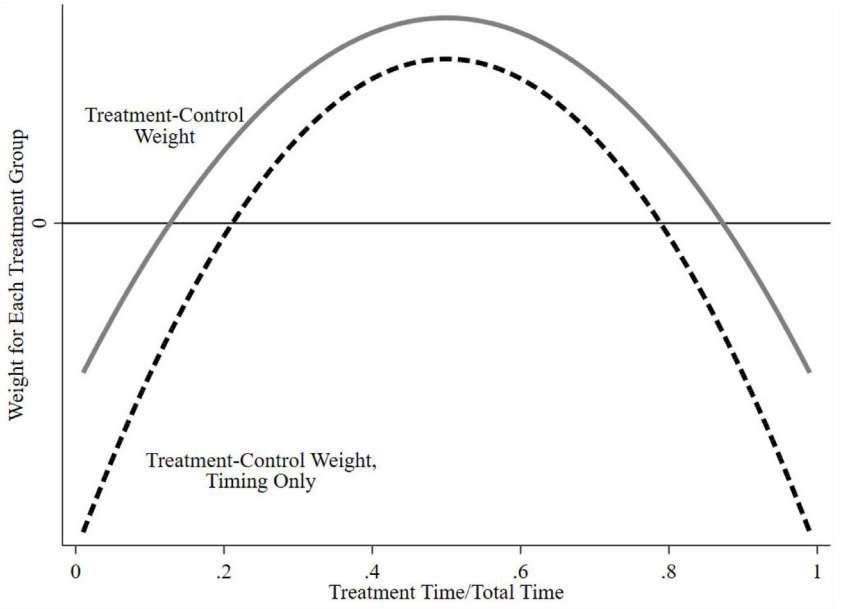

Fig. 4. Weighted common trends: The treatment/control weights as a function of the share of time spent under treatment. Notes: The figure plots the weights that determine each timing group's importance in the weighted common trends expression in Eqs. (16) and (17).

however, biases offset each other. Take the comparisons to timing group 1, for example. When timing group k is the treatment group in $\hat{\beta}^{2x2,k}_{1k}$, the bias equals ΔY^0_k and is weighted by σ^k_{1k}. When timing group k is the control group in $\hat{\beta}^{2x2,1}_{1k}$, the bias equals $-\Delta Y^0_k$ and is weighted by σ^1_{1k}. On net the bias in $\hat{\beta}^{2x2}_{1k}$ is ambiguous: $\Delta Y^0_k \left(\sigma^k_{1k} - \sigma^1_{1k}\right)$.

Similar expressions hold for the comparison of timing group k to every other group, and the total weight on each timing group's counterfactual trend equals the difference between the total weight it gets when it acts as a treatment group – w^T_k from Eq. (16) – minus the total weight it gets when it acts as a control group$-w^C_k \equiv \sum_{j=1}^{k-1} \sigma^j_{jk} + \sum_{j=k+1}^{K} \sigma^j_{kj}$. This difference is a new result that maps (linear) differential trends to bias.[20] A positive trend in timing group k induces positive bias when $w^T_k - w^C_k > 0$, negative bias when $w^T_k - w^C_k < 0$, and no bias when $w^T_k - w^C_k = 0$.[21]

Fig. 4 plots $w^T_k - w^C_k$ as a function of $\overline{D}$ assuming equal group sizes. Units treated in the middle of the panel have high treatment variance and get a lot of weight when they act as the treatment group, while units treated toward the ends of the panel get relatively more weight when they act as controls. As k approaches 1 or T, $w^T_k - w^C_k$ becomes negative which means that some timing groups effectively act as controls. This defines "the" control group in timing-only designs: all timing groups are controls in *some* terms, but the earliest and/or latest units necessarily get more weight as controls than treatments.

4. DD decomposition in practice: Unilateral divorce and female suicide

To illustrate how to use DD decomposition theorem in practice, I replicate Stevenson and Wolfers' (2006) analysis of no-fault divorce reforms and female suicide. Unilateral (or no-fault) divorce allowed either spouse to end a marriage, redistributing property rights and bargaining power relative to fault-based divorce regimes. Stevenson and Wolfers exploit "the natural variation resulting from the different timing of the adoption of unilateral divorce laws" in 37 states from 1969–1985 (see Table 1) using the "remaining fourteen states as controls" to evaluate the effect of these reforms on female suicide rates.

[20] Applications typically discuss bias in general terms, arguing that unobservables must be "uncorrelated" with timing, but have not been able to specify *how* counterfactual trends would bias a two-way fixed effects estimate. For example, Almond et al. (2011, p. 389–390) argue: "Counties with strong support for the low-income population (such as northern, urban counties with large populations of poor) may adopt FSP earlier in the period. This systematic variation in food stamp adoption could lead to spurious estimates of the program impact if those same county characteristics are associated with differential trends in the outcome variables".

[21] Clearly these results hold only under the assumption of linearity. This, however, is a common starting point, it approximates non-linear pre-trends, and it provides a simple way to increase the power of such pre-tests (see Bilinski and Hatfield, 2019). The decomposition weights could be combined with assumptions about post-treatment trend-breaks in a partial identification framework (Rambachan and Roth, 2019). Finally, when pre-treatment covariates are not measured at the same frequency as y_{it} then one must construct balance tests "by hand" since using confounders as outcomes in a fixed effects regression in a different sample will not rely on the same weights. Eq. (19) suggests a way to do so.

Table 1
The no-fault divorce rollout: Treatment times, timing group sizes, and treatment shares.

No-fault divorce year (k)	Number of states	Share of states (n_k)	Treatment share ($\overline{D}_k$)
Non-reform states	5	0.10	.
Pre-1964 reform states	8	0.16	.
1969	2	0.04	0.85
1970	2	0.04	0.82
1971	7	0.14	0.79
1972	3	0.06	0.76
1973	10	0.20	0.73
1974	3	0.06	0.70
1975	2	0.04	0.67
1976	1	0.02	0.64
1977	3	0.06	0.61
1980	1	0.02	0.52
1984	1	0.02	0.39
1985	1	0.02	0.36

Notes: The table lists the dates of no-fault divorce reforms from Stevenson and Wolfers (2006), the number and share of states that adopt in each year, and the share of periods each treatment timing group spends treated in the estimation sample from 1964 to 1996.

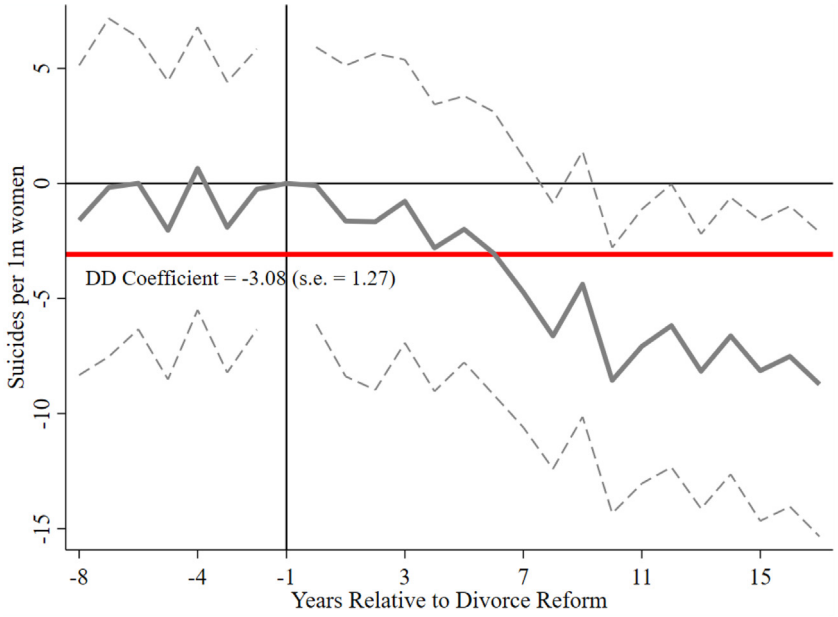

Fig. 5. Event-study and difference-in-differences estimates of the effect of no-fault divorce on female suicide: Replication of Stevenson and Wolfers (2006). Notes: The figure plots event-study estimates from the two-way fixed effects regression equation on page 276 and plotted in Fig. 1 of Stevenson and Wolfers (2006), along with the DD coefficient. The specification does not include other controls and does not weight by population. Standard errors are robust to heteroskedasticity.

Fig. 5 replicates their event-study result for female suicide using an unweighted specification with no covariates.[22] Our results match closely: suicide rates display no clear trend before the implementation of unilateral divorce laws, but begin falling soon after. They report a DD coefficient in logs of −9.7 (s.e. = 2.3). I find a DD coefficient in levels of −3.08 (s.e. = 1.13), or a proportional reduction of 6 percent.[23]

[22] Data on suicides by age, sex, state, and year come from the National Center for Health Statistics' Multiple Cause of Death files from 1964 to 1996, and population denominators come from the 1960 Census (Haines and ICPSR, 2010) and the Surveillance, Epidemiology, and End Results data (SEER, 2013). The outcome is the age-adjusted (using the national female age distribution in 1964) suicide mortality rate per million women. The average suicide rate in my data is 52 deaths per million women versus 54 in Stevenson and Wolfers (2006). My replication analysis uses levels to match their Figure, but the conclusions all follow from a log specification as well.

[23] The differences in the magnitudes likely come from three sources: age-adjustment (the original paper does not describe an age-adjusting procedure); data on population denominators; and my omission of Alaska and Hawaii.

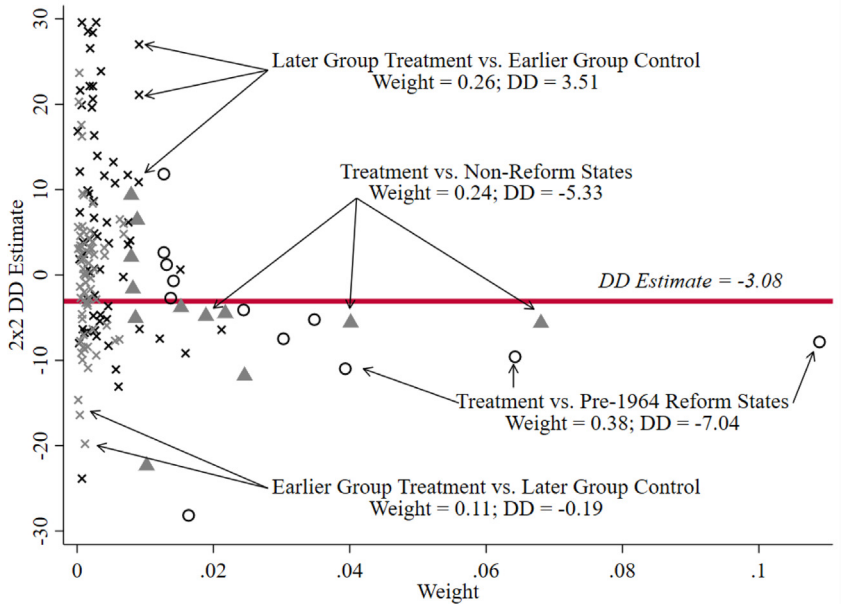

Fig. 6. Difference-in-differences decomposition for unilateral divorce and female suicide. Notes: The figure plots each 2x2 DD components from the decomposition theorem against their weight for the unilateral divorce analysis. The open circles are terms in which one timing group acts as the treatment group and the pre-1964 reform states act as the control group. The closed triangles are terms in which one timing group acts as the treatment group and the non-reform states act as the control group. The *x*'s are the timing-only terms. The figure notes the average DD estimate and total weight on each type of comparison. The two-way fixed effects estimate, −3.08, equals the average of the *y*-axis values weighted by their *x*-axis value.

4.1. Describing the design

Fig. 6 uses the DD decomposition theorem to illustrate the sources of variation. I plot each 2x2 DD against its weight and calculate the average effect and total weight for the three types of 2x2 comparisons: treated/untreated, early/late, late/early.[24] The two-way fixed effects estimate, −3.08, is an average of the *y*-axis values weighted by their *x*-axis values. Summing the weights on timing terms ($s^k_{k\ell}$ and $s^\ell_{k\ell}$) shows how much of $\hat{\beta}^{DD}$ comes from timing variation (37 percent). The large untreated group puts a lot of weight on $\hat{\beta}^{2x2}_{kU}$ terms, but more on those involving pre-1964 reform states (38.4 percent) than non-reform states (24 percent). Fig. 6 also highlights the role of a few influential 2x2 DDs. Comparisons between the 1973 states and non-reform/pre-1964 reform states account for 18 percent of the estimate, and the ten highest-weight 2x2 DDs account for over half.

The bias resulting from time-varying effects is also apparent in Fig. 6. The average of the post-treatment event-study estimates in Fig. 5 is −4.92, but the DD estimate is 60 percent as large. The difference stems from the comparisons of later- to earlier-treated groups. The average treated/untreated estimates are negative (−5.33 and −7.04) as are the comparisons of earlier- to later-treated states (although less so: −0.19).[25] The comparisons of later- to earlier-treated states, however, are *positive* on average (3.51) and account for the bias in the overall DD estimate. Using the decomposition theorem to take these terms out of the weighted average yields an effect of −5.44—close to the average of the event-study coefficients.

Fig. 7 plots the weights that each timing group gets in VWCT for the unilateral divorce analysis in solid gray circles alongside each timing group's share of the sample, n_k, in open squares. The earliest-treated states have systematically get less weight than their sample shares because their treatment variance is low. The 1970 states get almost no weight ($w^T_k - w^C_k = 0.0039$) so a differential trend in this timing group cancels out across 2x2 DDs and has little effect on $\hat{\beta}^{DD}$.[26] States that implemented unilateral divorce in 1969 get more weight as controls than they do as treatments. I also plot the common trends weights for a specification that drops always- and never-treated states in solid black triangles. In this sample states treated before 1973 get more weight as controls than treatments.

[24] There are 156 distinct DD components: 12 comparisons between timing groups and pre-reform states, 12 comparisons between timing groups and non-reform states, and $\left(12^2 - 12\right)/2 = 66$ comparisons between an earlier switcher and a later non-switcher, and 66 comparisons between a later switcher and an earlier non-switcher.

[25] This point also applies to units that are already treated at the beginning of the panel, like the pre-1964 reform states in the unilateral divorce analysis. Since their $\overline{D}_k = 1$ they can only act as a control group. If the effects for pre-1964 reform states had stabilized by 1969 they would not cause bias, but this is a particular version of the assumption that $\Delta ATT = 0$.

[26] Adding a trend of 5 × *year* to the suicide rate for the 1970 states changes the DD estimate from −3.08 to −2.75, but adding it to the 1973 states ($w^T_k - w^C_k = 0.18$) yields a very biased DD estimate of 12.28.

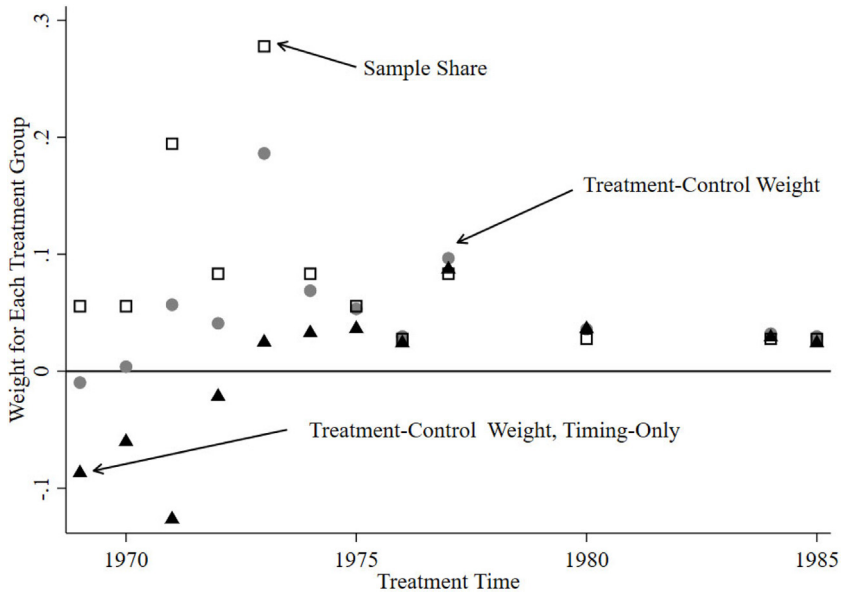

Fig. 7. Weighted common trends in the unilateral divorce analysis: The treatment/control weights on each timing group. Notes: The figure plots the weights that determine each timing group's role in the weighted common trends expression. These are show in solid triangles and equal the difference between the total weight each treatment timing group receives in terms where it is the treatment group (w_k^T) and terms where it is the control group (w_k^C): $w_k^T - w_k^C$. The solid circles show the same weights but for versions of each estimator that exclude the untreated (or already-treated) units and, therefore, are identified only by treatment timing. The open squares plot each timing group's sample share.

5. Alternative specifications

The results above refer to parsimonious regressions like (2), but researchers almost always estimate multiple specifications and use differences to evaluate internal validity (Oster, 2016) or choose projects in the first place. This section extends the DD decomposition theorem to different weighting choices and control variables, providing simple new tools for learning why estimates change across specifications.

The DD decomposition theorem suggests a simple way to understand why estimates change by writing the weighted average from Theorem 1 as a product a vector of 2x2 DDs and a vector of weights: $\hat{\beta}^{DD} = \boldsymbol{s}'\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}$. When a different estimator can also be written as a function of 2x2 DDs—$\hat{\beta}^{DD}_{alt} = \boldsymbol{s}'_{\boldsymbol{alt}}\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}_{\boldsymbol{alt}}$—the difference between the two specifications has the form of a Oaxaca–Blinder–Kitagawa decomposition (Blinder, 1973; Oaxaca, 1973; Kitagawa, 1955):

$$\hat{\beta}^{DD}_{alt} - \hat{\beta}^{DD} = \overbrace{\boldsymbol{s}'\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}_{\boldsymbol{alt}} - \hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}\right)}^{\text{Due to 2x2 DDs}} + \overbrace{\left(\boldsymbol{s}'_{\boldsymbol{alt}} - \boldsymbol{s}'\right)\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}}^{\text{Due to weights}} + \overbrace{\left(\boldsymbol{s}'_{\boldsymbol{alt}} - \boldsymbol{s}'\right)\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}_{\boldsymbol{alt}} - \hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}\right)}^{\text{Due to interaction}}. \tag{20}$$

Dividing by $\hat{\beta}^{DD}_{alt} - \hat{\beta}^{DD}$ shows the proportional contribution of changes in the 2x2 DD's, changes in the weights, and the interaction of the two.[27] It is also simple to learn which terms drive each kind of difference by plotting $\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}_{\boldsymbol{alt}}$ against $\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}$ and $\boldsymbol{s}$ against $\boldsymbol{s}_{\boldsymbol{alt}}$.

5.1. Weighting

Population weighting increases the influence of large *units* in means of y that make up each 2x2 DD (which affects $\hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}_{\boldsymbol{WLS}} - \hat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}_{\boldsymbol{OLS}}$), and it increases the influence of terms involving large *groups* by basing the decomposition weights on population rather than sample shares (which affects $\boldsymbol{s}'_{\boldsymbol{WLS}} - \boldsymbol{s}'_{\boldsymbol{OLS}}$).[28] In Table 2, population weighting changes the unilateral

[27] Grosz et al. (2018) propose a similar decomposition for family fixed effects estimates.

[28] One common robustness check is to drop untreated units, and the decomposition theorem shows that this is equivalent to setting all $s_{kU} = 0$ and rescaling the $s_{k\ell}$ to sum to one. In Table 2, this actually makes the unilateral divorce estimate positive (2.42, s.e. = 1.81). The average of the early/late and late/early DDs in Fig. 6 using the decomposition weights is: $(0.264/.374) \cdot 3.51 + (0.11/.374) \cdot -0.19 = 2.42$. The sign flip occurs because *half* of the timing terms are biased by time-varying treatment effects.

Table 2
DD estimates of the effect of unilateral divorce analysis on female suicide: Alternative specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Baseline	No untreated states	WLS	Propensity score weighting	Controls	Unit-specific trends	Group-specific pre-trends	Region-by-year fixed effects
Unilateral divorce	−3.08	2.42	−0.35	1.04	−2.52	0.59	−6.52	−1.16
	[1.27]	[1.81]	[1.97]	[1.78]	[1.09]	[1.35]	[2.98]	[1.37]
Difference from baseline specification		5.50	2.73	4.12	0.56	3.67	−3.44	1.92
Share due to:								
2x2 DDs		0	0.52	1	0.22	0.90	1	0.37
Weights		1	0.39	0	0.05	0.47	0	0.76
Interaction		0	0.09	0	<0.01	−0.36	0	−0.13
Within term		0	0	0	0.73	0	0	0

Notes: The table presents DD estimates from the alternative specifications discussed in section III. Column (1) is the two-way fixed effects estimate from Eq. (2). Column (2) drops the pre-1964 reform and non-reform states. Column (3) weights by state adult populations in 1964. Column (4) weights by the inverse propensity score estimated from a probit model that contains the sex ratio, per-capita income, the general fertility rate, and the infant mortality rate all measured in 1960. Column (5) controls for per-capita income, female homicide rates, and per-capita welfare caseloads. Column (6) includes state-specific linear time trends. Column (7) comes from a two-step procedure that first estimates group-specific trends from 1964 to 1968, subtracts them from the suicide rate, and estimates Eq. (2) on the transformed outcome variable. Column (8) includes region-by-year fixed effects. Below the standard errors I show the difference between each estimate and the baseline result, and the last three rows show the share of this difference that comes from changes in the 2x2 DD's, the weights, or their interaction as shown in Eq. (18).

divorce DD estimate from −3.08 to −0.35. Just over half of the difference comes from changes in the 2x2 DD terms, 38 percent from changes in the weights, and 9 percent from the interaction of the two.[29]

Fig. 8 scatters the weighted 2x2 DDs against the unweighted ones. Most components do not change and lie along the 45-degree line, but large differences emerge for terms involving the 1970 states: Iowa and California.[30] Weighting gives more influence to California, and makes the terms that use 1970 states as treatments more negative, while it makes terms that use them as controls more positive. This is consistent either with an ongoing downward trend in suicides in California or, as discussed above, strongly time-varying treatment effects.[31]

5.2. DD with controls

The ability to control for covariates is a common motivation for regression DD as it is thought to make a "common trends" assumption more plausible. Cameron and Trivedi (2005, p. 770), observe that "an obvious extension is to include regressors" and Angrist and Pischke (2009, p. 236) state "a further advantage of regression DD: it's easy to add additional covariates". Theoretical analyses typically focus on time-invariant $\boldsymbol{X_i}$ entered as a direct control in specifications like (1) (Sant'Anna and Zhao, 2018), or reweighting strategies that use $\boldsymbol{X_i}$ itself or pre-treatment changes in covariates or outcomes (Abadie, 2005; Ben-Michael et al., 2019). Most applications, however, include time-varying controls $\boldsymbol{X_{it}}$:

$$y_{it} = \alpha_{i\cdot} + \alpha_{\cdot t} + \boldsymbol{\Phi X_{it}} + \beta^{DD|X} D_{it} + e_{it}. \tag{21}$$

This subsection derives a decomposition result like Theorem 1 for controlled TWFEDD specifications. Appendix A discusses how covariates affect the theoretical properties of TWFEDD. Note that for covariates to aid in identification, they must be unaffected by the treatment to avoid bias from "conditioning on a post-treatment variable" (Rosenbaum, 1984).

To see how the controlled DD coefficient is identified first remove unit- and time-means and then estimate a Frisch–Waugh–Lovell regression that partials $\tilde{\boldsymbol{X}}_{\boldsymbol{it}}$ out of $\tilde{D}_{it}$:

$$\tilde{D}_{it} = \overbrace{\boldsymbol{\Gamma}\tilde{\boldsymbol{X}}_{\boldsymbol{it}}}^{\tilde{p}_{it}} + \tilde{d}_{it}. \tag{22}$$

The index of covariates, $\tilde{p}_{it} \equiv \hat{\boldsymbol{\Gamma}}\tilde{\boldsymbol{X}}_{\boldsymbol{it}}$ is a linear prediction of treatment status based on the sample-wide relationship between $\tilde{\boldsymbol{X}}_{\boldsymbol{it}}$ and $\tilde{D}_{it}$ (see Sloczynski, 2017). The covariate-adjusted treatment variable subtracts predicted treatment status

[29] Solon et al. (2015) show that differences between population-weighted (WLS) and unweighted (OLS) estimates can arise in the presence of unmodeled heterogeneity, and suggest comparing the two estimators (Deaton, 1997; Wooldridge, 2001).

[30] Lee and Solon (2011) observe that California drives the divergence between OLS and WLS estimate in analyses of no-fault divorce on divorce rates (Wolfers, 2006).

[31] Weighting by a function of the estimated propensity score is often used to impose covariate balance between treated and untreated units (Abadie, 2005). With timing variation this approach has two limitations. First, reweighting untreated observations has no effect on the timing terms. Second, reweighting untreated observations by their propensity to be in *any* timing group does not impose covariate balance for each timing group. By changing the relative weight on different untreated units but leaving their total weight the same, this strategy does not change $\boldsymbol{s}$, so all differences stem from the way reweighting affects the $\hat{\beta}^{2x2}_{kU}$ terms. Table 2 estimates reweighted specification based on a propensity score equation that contains the 1960 sex ratio and per-capita income, general fertility rate and infant mortality rate. This puts much more weight on Delaware and less weight on New York, and makes almost all $\hat{\beta}^{2x2}_{kU}$ much less negative, changing the overall DD estimate to 1.04. Callaway and Sant'Anna (2020) propose a generalized propensity score reweighted estimator to exploit timing variation.

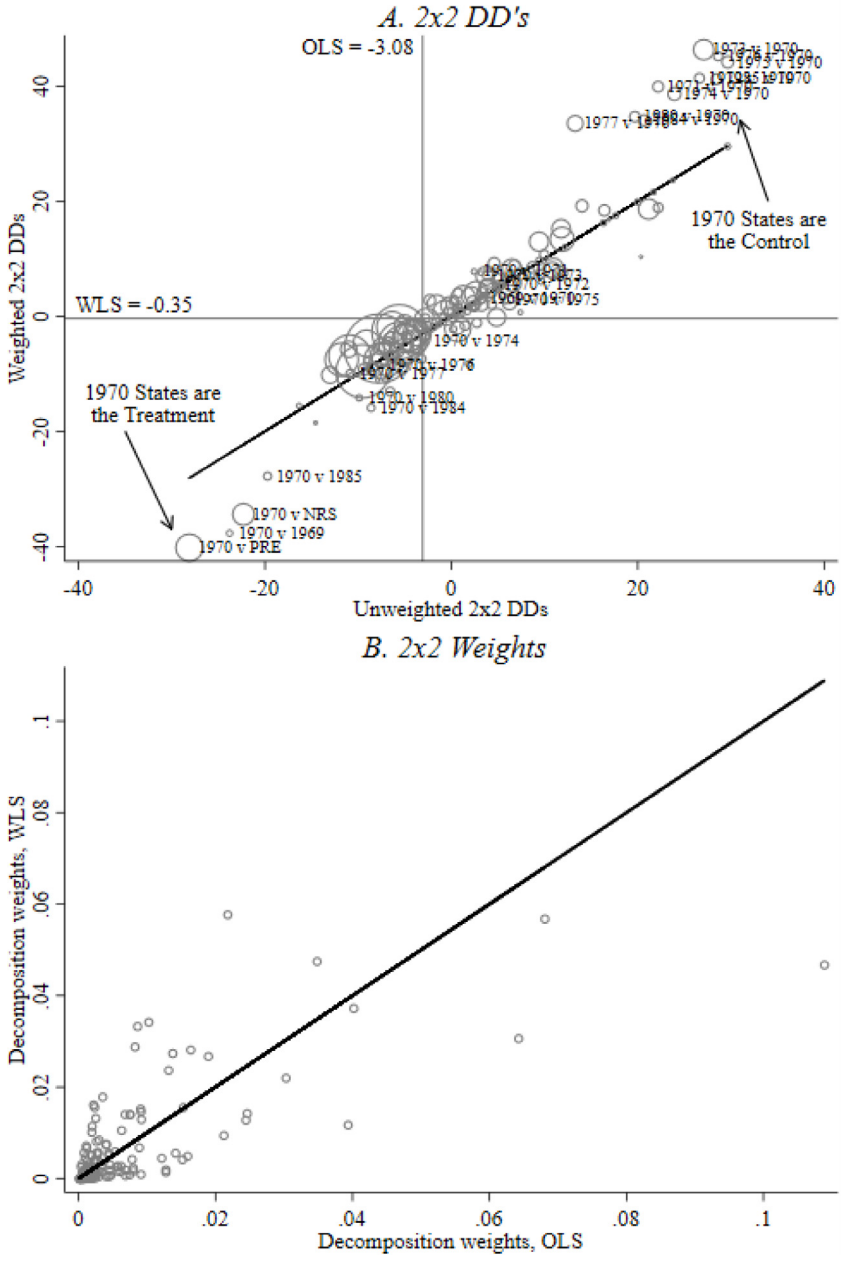

Fig. 8. Comparison of 2x2 DD components and decomposition weights with and without population weights. Notes: Panel A plots the 2x2 DD components from two-way fixed effects estimates that use population weights (*y*-axis) and do not (*x*-axis). The size of each point is proportional to its weight in an OLS version of Eq. (7). WLS estimates are much smaller than OLS estimates, and this figure shows that the source of this discrepancy is the 1970 no-fault divorce states, which include only Iowa and California. Weighting puts much more emphasis on California and, therefore, every 2x2 DD component involving the 1970 states. Dropping California changes yields an OLS estimate of −3.32 and a WLS estimate of −1.43.

from true treatment status: $\tilde{d}_{it} \equiv \left[\left(D_{it} - \overline{D}_i\right) - \left(\hat{\Gamma}\overline{\boldsymbol{X}}_{\boldsymbol{it}} - \hat{\Gamma}\overline{\boldsymbol{X}}_{\boldsymbol{i}}\right)\right] - \left[\left(\overline{D}_t - \overline{\overline{D}}\right) - \left(\hat{\Gamma}\overline{\boldsymbol{X}}_{\boldsymbol{t}} - \hat{\Gamma}\overline{\overline{\boldsymbol{X}}}\right)\right]$ so that:

$$\hat{\beta}^{DD|X} \equiv \frac{\hat{C}\left(y_{it}, \tilde{d}_{it}\right)}{\hat{V}^d} = \frac{\hat{C}\left(y_{it}, \tilde{D}_{it} - \tilde{p}_{it}\right)}{\hat{V}^d}. \tag{23}$$

$\hat{\beta}^{DD|X}$ exploits variation in both $\tilde{D}_{it}$ *and* $\tilde{p}_{it}$. $\tilde{D}_{it}$ varies by timing group and before/after treatment times, but $\tilde{p}_{it}$ (generally) varies across units, even those in the *same* treatment timing group and time period.

To derive a decomposition result for $\hat{\beta}^{DD|X}$ first split $\tilde{d}_{it}$ into a "between" timing group term and a "within" timing group term by adding and subtracting group-by-year averages $\overline{d}_{kt} - \overline{d}_k = \left(\overline{D}_{kt} - \overline{D}_k\right) - \left(\hat{\Gamma}\overline{\boldsymbol{X}}_{\boldsymbol{kt}} - \hat{\Gamma}\overline{\boldsymbol{X}}_{\boldsymbol{k}}\right)$. Let $\tilde{d}_{i(k)t}$ denote deviations

of d_{it} from unit means and timing-group-by-year averages: $\left(d_{it}-\overline{d}_i\right)-\left(\overline{d}_{kt}-\overline{d}_k\right)$. I call this the "within" component of $\tilde{d}_{it}$. The only reason that $\tilde{d}_{i(k)t}$ is non-zero is because $\tilde{p}_{it}$ is not identical for units in the same timing group. Let $\tilde{d}_{kt}$ denote a two-way fixed effects adjusted version of the mean of $\tilde{d}_{it}$ by timing group and year: $\left(\overline{d}_{kt}-\overline{d}_k\right)-\left(\overline{d}_t-\overline{\overline{d}}\right)$. $\tilde{d}_{kt}$ varies by timing-group and time period only, just like $\tilde{D}_{kt}$. I call this the "between" component of $\tilde{d}_{it}$. Rewrite $\tilde{d}_{it}$ in terms of $\tilde{d}_{i(k)t}$ and $\tilde{d}_{kt}$, and substitute into Eq. (23):

$$\tilde{d}_{it}=\overbrace{\left(d_{it}-\overline{d}_i\right)-\left(\overline{d}_{kt}-\overline{d}_k\right)}^{\tilde{d}_{i(k)t}}+\overbrace{\left(\overline{d}_{kt}-\overline{d}_k\right)-\left(\overline{d}_t-\overline{\overline{d}}\right)}^{\tilde{d}_{kt}} \tag{24}$$

$$\hat{\beta}^{DD|X}=\frac{\hat{C}\left(y_{it},\tilde{d}_{i(k)t}\right)+\hat{C}\left(y_{it},\tilde{d}_{kt}\right)}{\hat{V}^d}=\overbrace{\frac{\hat{V}^d_w}{\hat{V}^d}}^{\Omega}\hat{\beta}^p_w+\overbrace{\frac{\hat{V}^d_b}{\hat{V}^d}}^{1-\Omega}\overbrace{\left[\frac{\hat{\beta}^{DD}\hat{V}^D-\hat{\beta}^p_b\hat{V}^p_b}{\hat{V}^d_b}\right]}^{\hat{\beta}^d_b}. \tag{25}$$

I use the subscript w to denote within-timing-group terms and the subscript b to denote between-timing-group terms. $\hat{V}^d_w$ is the variance of $\tilde{d}_{i(k)t}$, the within component of the adjusted treatment variable. $\hat{V}^d_b$ and $\hat{V}^p_b$ are the variances of the between components $\tilde{d}_{kt}$ and $\tilde{p}_{kt}$. The term $\Omega\equiv\frac{\hat{V}^d_w}{\hat{V}^d}$ measures the share of the identifying variation that comes from within-timing-group comparisons.

The within coefficient, $\hat{\beta}^p_w\equiv\frac{\hat{C}\left(y_{it},\tilde{d}_{i(k)t}\right)}{\hat{V}^d_w}$, measures the relationship between y_{it} and changes over time in $\tilde{d}_{i(k)t}$ across units in the same timing group.[32] There is no variation in $\tilde{D}_{it}$ within timing groups, though, so $\tilde{d}_{i(k)t}$ only varies because of predicted treatment status. $\hat{\beta}^p_w$ compares units with the same treatment status but different predicted treatment paths. Adding controls therefore introduces a new source of identifying variation—within-group changes in $\boldsymbol{X_{it}}$—that was *not there* in the unadjusted version.

The "between" term in square brackets, $\hat{\beta}^d_b\equiv\frac{\hat{C}\left(y_{it},\tilde{d}_{kt}\right)}{\hat{V}^d_b}$, comes from timing-group-by-time-period variation, just as in Theorem 1. It contains the unadjusted DD coefficient $\hat{\beta}^{DD}$ and subtracts $\hat{\beta}^p_b$, the coefficient from a regression of y_{it} on $\tilde{p}_{kt}$ (the mean covariate index by timing group and time) and two-way fixed effects. Appendix A decomposes $\hat{\beta}^d_b$ into adjusted 2x2 DDs as in Theorem 1:

$$\hat{\beta}^d_b=\sum_k\sum_{\ell>k}\overbrace{(n_k+n_\ell)^2\frac{\hat{V}^d_{b,k\ell}}{\hat{V}^d_b}}^{s^{b|X}_{k\ell}}\overbrace{\left[\frac{\hat{V}^D_{k\ell}\hat{\beta}^{2x2}_{k\ell}-\hat{V}^p_{b,k\ell}\hat{\beta}^p_{b,k\ell}}{\hat{V}^d_{b,k\ell}}\right]}^{\hat{\beta}^d_{b,k\ell}}. \tag{26}$$

The variances and coefficients in (26) parallel those in (25) but as the subscripts indicate they come from each two-group subsample.[33] Controls change the estimate for the two reasons highlighted in Eq. (20). The weight on each 2x2 is based on $\hat{V}^d_{b,k\ell}$ rather than $\hat{V}^D_{kl}$, which shows that covariates change the way TWFEDD weights subsample estimators. The 2x2 DDs contain the unadjusted terms, $\hat{V}^D_{k\ell}\hat{\beta}^{2x2}_{k\ell}$, but also subtracts the subsample coefficient relating y_{it} and $\tilde{p}_{kt}$: $\hat{V}^p_{b,k\ell}\hat{\beta}^p_{b,k\ell}$, which shows that covariates adjust the 2x2 estimates themselves.[34]

[32] Because it comes from deviations of d_{it} from timing-group-by-time means, $\hat{\beta}^p_w$ is equivalent to regressing y_{it} on unit fixed effects, timing-group-by-time fixed effects, and d_{it}.

[33] Note that (26) decomposes $\hat{\beta}^{DD|X}_b$ by pairs of timing groups but does *not* break up the timing comparisons into terms corresponding to $\hat{\beta}^{2x2,k}_{k\ell}$ and $\hat{\beta}^{2x2,\ell}_{k\ell}$. The control term, $\hat{V}^p_{k\ell}\hat{\beta}^b_{k\ell}$, cannot be easily written as an average across overlapping subsets of time ($PRE(\ell)$ and $POST(k)$).

[34] The expression for controlled 2x2 terms in (26) does not come from estimating Eq. (21) on the subsamples. A controlled 2x2 DD – $\hat{\beta}^{2x2|X}_{b,k\ell}$ – would come from adjusting for covariates *on that subsample* using predicted treatment status $\tilde{p}^{k\ell}_{jt}\equiv\boldsymbol{\Gamma_{k\ell}X_{kt}}$. But $\hat{\beta}^d_{b,k\ell}$ adjusts by predicted treatment from the *full* sample, $\tilde{p}_{jt}$. To see how the two relate, add and subtract $\tilde{p}^{k\ell}_{jt}$ in $\hat{C}\left(y_{jt},\tilde{D}_{jt}-\tilde{p}_{jt}\right)$, the numerator of each $\hat{\beta}^d_{b,k\ell}$:

$$\begin{aligned}\hat{\beta}^d_{b,k\ell}&=\frac{\hat{C}\left(y_{jt},\tilde{D}_{jt}-\tilde{p}^{k\ell}_{jt}\right)+\hat{C}\left(y_{jt},\tilde{p}^{k\ell}_{jt}-\tilde{p}_{jt}\right)}{\hat{V}^d_{b,k\ell}},\quad j\in k,\ell\\&=\frac{\left(1-R^2_{k\ell}\right)\hat{V}^D_{k\ell}\hat{\beta}^{2x2|X}_{k\ell}+\hat{V}^{dp}_{b,k\ell}\hat{\beta}^{dp}_{b,k\ell}}{\left(1-R^2_{k\ell}\right)\hat{V}^D_{k\ell}+\hat{V}^{dp}_{b,k\ell}}\end{aligned}$$

The superscript dp refers to the difference between subsample and full sample predicted treatment, $\tilde{p}^{k\ell}_{jt}-\tilde{p}_{jt}$. $\hat{V}^{dp}_{b,k\ell}$ is its variance and $\hat{\beta}^{dp}_{b,k\ell}$ is the regression coefficient relating it to y_{jt}. $R^2_{k\ell}$ comes from the subsample Frisch–Waugh–Lovell regression. When $R^2_{k\ell}=1$, then $\tilde{p}^{k\ell}_{jt}=\tilde{D}_{jt}$ and the estimate collapses back to $\hat{\beta}^d_{b,k\ell}$ as defined in (25). In other words, adjusted 2x2 DDs still contribute even with $\boldsymbol{X_{kt}}$ and D_{kt} are perfectly collinear in the subsample. When $\boldsymbol{\Gamma_{k\ell}}\approx\boldsymbol{\Gamma}$, then $\hat{V}^{dp}_{b,k\ell}\approx 0$ and $\hat{\beta}^{2x2|d}_{b,k\ell}\approx\hat{\beta}^{2x2|X}_{k\ell}$.

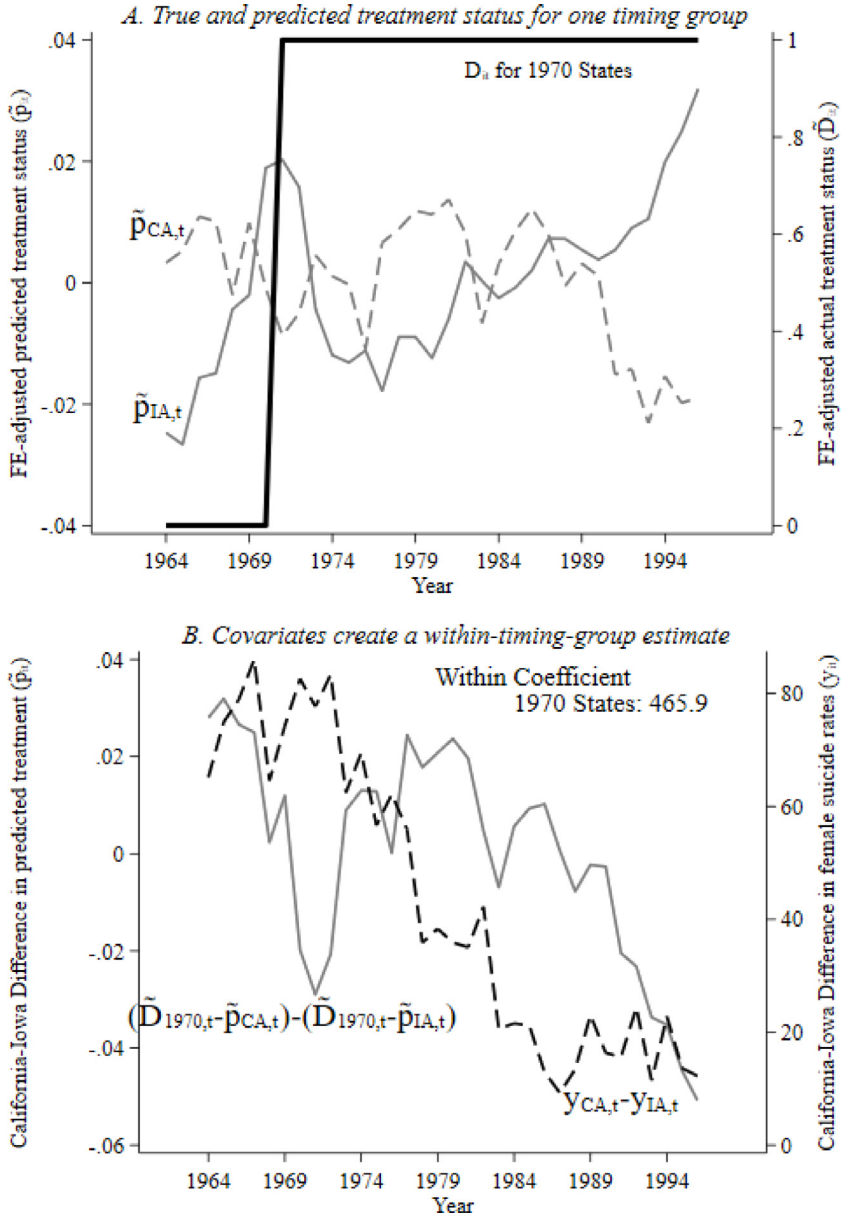

Fig. 9. Adding controls creates within-timing-group comparisons: An example with the 1970 no-fault divorce states. Notes: Panel A plots the treatment dummy and fitted values from the full-sample Frisch–Waugh–Lovell regression (predicted treatment status, $\tilde{p}_{it}$) for the 1970 states, California and Iowa. Panel B plots the difference in adjusted treatment variable, $\tilde{D}_{it} - \tilde{p}_{it}$, between California and Iowa and the same difference in female suicide rates. Both fall over time and are highly correlated. The coefficient from a regression of the difference in suicide rates on $\left(\tilde{D}_{1970,t} - \tilde{p}_{CA,t}\right) - \left(\tilde{D}_{1970,t} - \tilde{p}_{IA,t}\right)$ equals 465.9. This is the part of the within term in (22) that comes from the 1970 group.

Eqs. (25) and (26) give the full decomposition for a controlled specification:

$$\hat{\beta}^{DD|X} = \Omega\hat{\beta}^{p}_{w} + (1-\Omega)\sum_{k}\sum_{\ell>k} s^{b|X}_{k\ell}\hat{\beta}^{2x2|d}_{k\ell} \tag{27}$$

$\Omega\hat{\beta}^{p}_{w}$ is the contribution of within-timing-group variation. $(1-\Omega)$ is the weight on the covariate-adjusted between terms, $\hat{\beta}^{2x2|d}_{k\ell}$ each of which gets a weight of $s^{b|X}_{k\ell}$.

In the unilateral divorce analysis, I add three covariates: female homicide rates, per-capita income, and the welfare participation rate. Column 5 of Table 2 reports a controlled DD estimate of −2.52 (s.e. = 1.09), almost 20 percent smaller than the unadjusted coefficient. Most of the differences comes from the within term. Fig. 9 illustrates the within variation

for the two 1970 no-fault divorce states, California and Iowa. The two states have the same values of $\tilde{D}_{it}$ by definition, but panel A shows that *predicted* treatment is falling slightly in California and rising slightly in Iowa. Panel B plots the difference in treatment deviations, $\tilde{d}_{CA,t} - \tilde{d}_{IA,t} = \left(\tilde{D}_{CA,t} - \tilde{p}_{CA,t}\right) - \left(\tilde{D}_{IA,t} - \tilde{p}_{IA,t}\right) = \tilde{p}_{IA,t} - \tilde{p}_{CA,t}$, and the difference in suicide rates. The regression coefficient relating the two is large and positive (465.9). The full-sample within coefficient $\hat{\beta}_w^{DD|X}$ equals 80.01, but the within variance in predicted treatment is small ($\hat{V}_w^d = 0.005$). Within-group variation from the covariates therefore changes the DD estimate by $\Omega \times \hat{\beta}_w^{DD|X} = 80.01 \times 0.005 = 0.40$, or 73 percent of the difference across specifications.

Fig. 10 illustrates the controlled between term for the 1970 states compared to non-reform states, $\hat{\beta}^{2x2|d}_{1970,NRS}$. Panel A plots the treatment variable and the group-year means of predicted treatment status from the full-sample Frisch–Waugh–Lovell regression. $\tilde{p}_{kt}$ does not change much indicating that covariates do not predict treatment very well. In fact the R^2 from (22) is just 0.0067. Panel B plots differences in the group-level adjusted treatment variable $\tilde{d}_{1970,t} - \tilde{d}_{NRS,t}$ and differences in suicide rates. Because the controls do not absorb very much treatment variation, the controlled 2x2 term (−22.4) is almost the same as the uncontrolled one (−22.3). These control variables do not explain the adoption of no-fault divorce laws very well, but they are correlated with suicide rates across states that adopt these laws in the same year.[35]

Appendix B analyzes two common controls strategies: unit-specific linear time trends and region-by-year fixed effects. Column 6 of Table 2 shows that unit-specific trends change the unilateral divorce estimate to 0.59 (s.e. = 1.35), consistent with the observation that trends over-control for time-varying treatment effects (Lee and Solon, 2011; Neumark et al., 2014; Meer and West, 2016). I also propose a two-step strategy that fits linear trends by group in the pre-period only, subtracts them from the outcome in all periods, and then estimates an uncontrolled regression on the transformed outcome. This pre-trend-adjusted estimator is unaffected by effect dynamics and does not change the weights. Column 7 of Table 2 shows that adjusting for pre-trends only yields an estimate of −6.52 (s.e. = 2.98). The estimator with region-by-year fixed effects (column 8 of Table 2) preserves the form of Theorem 1, but essentially applies it within each region and then weights the 2x2s from different regions together by sample size. Note that 2x2s can drop out in this kind of specification if no region contains a given pair of timing groups.[36]

6. Conclusion

Difference-in-differences is perhaps the most widely applicable quasi-experimental research design, but it has primarily been understood in the context of the simplest two-group/two-period estimator. I show that when treatment timing varies across units, the TWFEDD estimator equals a weighted average of all possible simple 2x2 DDs that compare one group that changes treatment status to another group that does not. Many ways in which the theoretical interpretation of regression DD differs from the canonical model stem from the fact that these simple components are weighted together based both on sample sizes *and* the variance of their treatment dummy. The causal estimand that TWFEDD can identify is a variance-weighted average treatment effect on the treated (*VWATT*). It does so under two assumptions: that a variance-weighted average of untreated potential outcome changes equals zero ($VWCT = 0$) and that average treatment effects for each timing group do not change over time ($\Delta ATT = 0$). The assumption of constant treatment effects is necessary because already-treated units act as the control group in some 2x2 DD terms.

Researchers seeking to exploit variation in treatment timing to estimate causal effects should use TWFEDD with caution. The TWFEDD estimator only has a meaningful causal interpretation under strong assumptions on treatment effects and even then it yields a parameter that may differ from what researchers have in mind. If treatment effects are likely to vary over time one should not use TWFEDD to summarize the estimated effects. If the variance-weighted average of treatment effects is not of interest one should not use TWFEDD either. My results provide tools to help applied researchers evaluate both of these issues and judge whether TWFEDD can provide meaningful causal estimates. Fortunately, alternative estimators have recently been developed that can deliver causal estimates when TWFEDD cannot. They all carefully construct control groups to address the bias from time-varying treatment effects (Ben-Michael et al., 2019; Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020) and some allow control over the target parameter (Callaway and Sant'Anna, 2020) which improves interpretation of the results.

35 Appendix A analyzes the theoretical properties of a single controlled 2x2 DD ($\hat{\beta}_{kU}^{2x2|X}$) abstracting from the within-group term and differences in predicted treatment in the subsample versus full sample and additionally assuming that covariates are not affected by the treatment. When treatment effects are correlated with post-period changes in the covariates, controls absorb part of the treatment effect. This generalizes an existing point about unit-specific linear time trends (Lee and Solon, 2011). Any control variable could inappropriately absorb treatment effects. Moreover, when correctly and completely specified, controls do successfully partial out differential trends, but since $\boldsymbol{X_{it}}$ varies period-by-period even within the *PRE* (k) and *POST* (k), they also partial out period-by-period covariances between Y^0 and predicted treatment status that do not in themselves bias $\hat{\beta}_{kU}^{2x2}$. In sum, I find four main ways in which controlling for $\boldsymbol{X_{it}}$ in a regression does not address the bias in DD models. First, it introduces within-group comparisons that could not have biased $\hat{\beta}^{DD}$. Second, it extrapolates the full-sample predicted treatment variable onto the pairwise components. Third, it partials out period-by-period covariance between controls and untreated potential outcomes within the pre/post periods that could not have biased $\hat{\beta}^{DD}$. Lastly it nets out any part of the treatment effect that is correlated with differential covariate paths in the post period.

36 Appendix B also analyzes triple-difference models and shows that they also have a weighted average form. Appendix C briefly discusses treatment variables that turn off.

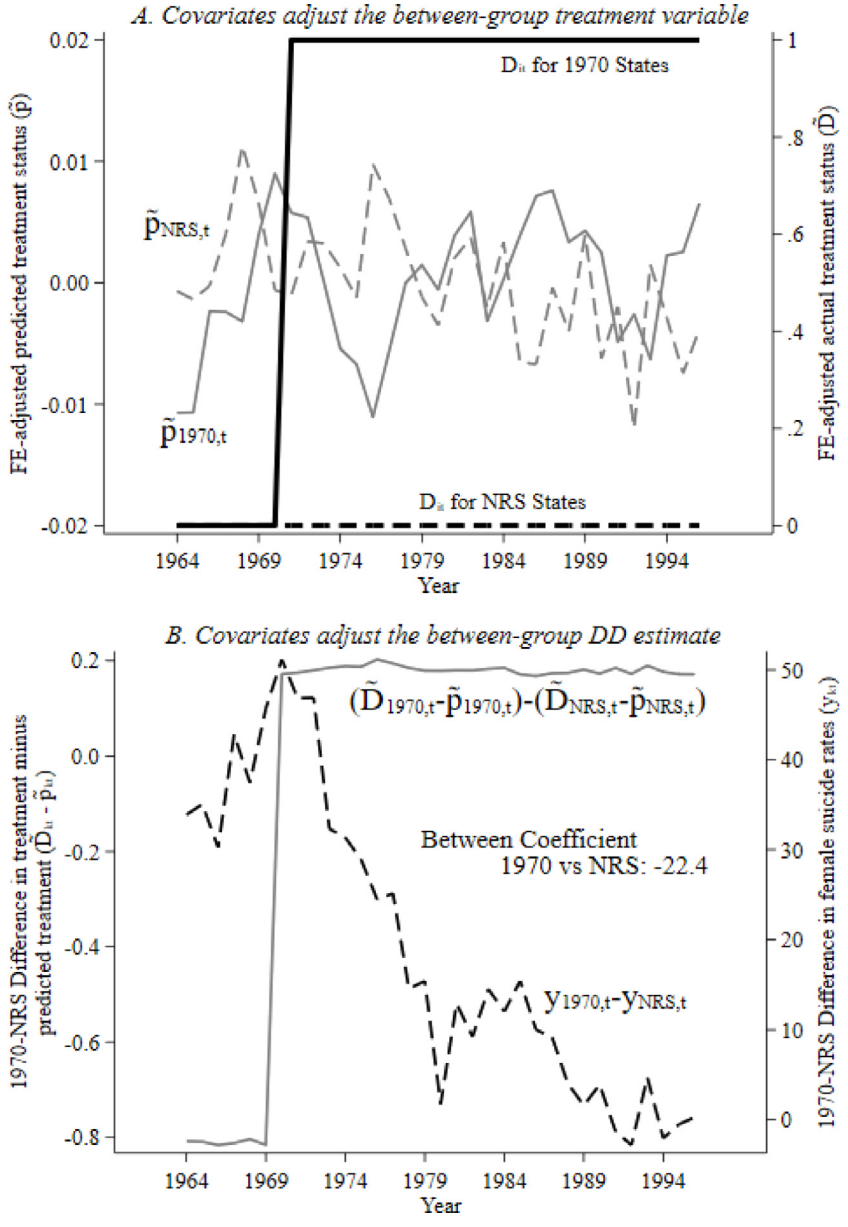

Fig. 10. Adding controls adjusts between-timing-group comparisons: An example with the 1970 no-fault divorce states compared to non-reform states. Notes: Panel A plots the treatment dummy and group-year averages of fitted values from the full-sample Frisch–Waugh–Lovell regression ($\tilde{p}_{kt}$) for the 1970 states and non-reform states. Panel B plots the difference in adjusted treatment variable, $\left(\tilde{D}_{1970,t} - \tilde{p}_{1970,t}\right) - \left(\tilde{D}_{NRS,t} - \tilde{p}_{NRS,t}\right)$, between the two groups and the same difference in female suicide rates. The covariates do not adjust the treatment dummy very much, so the controlled coefficient (−22.4) is almost identical to the uncontrolled 2x2 DD (−22.3). This is the part of the between term in (27) that comes from the 1970 versus non-reform comparison.

Acknowledgments

I thank Michael Anderson, Andrew Baker, Martha Bailey, Marianne Bitler, Brantly Callaway, Kitt Carpenter, Eric Chyn, Bill Collins, Scott Cunningham, John DiNardo, Andrew Dustan, Federico Gutierrez, Brian Kovak, Emily Lawler, Doug Miller, Austin Nichols, Sayeh Nikpay, Edward Norton, Jesse Rothstein, Pedro Sant'Anna, Jesse Shapiro, Gary Solon, Isaac Sorkin, Sarah West, and seminar participants at the Southern Economics Association, ASHEcon 2018, the University of California, Davis, University of Kentucky, University of Memphis, University of North Carolina Charlotte, the University of Pennsylvania, and Vanderbilt University. All errors are my own.

Appendix A. Proof of the DD decomposition theorem

The proofs involve sample covariances of demeaned variables, and rely on this Lemma:

Lemma 1. *The covariance between a variable z_{gt} and two-way-fixed-effects-adjusted variable $\tilde{x}_{kt} = (x_{kt} - \bar{x}_k) - \left(\bar{x}_t - \bar{\bar{x}}\right)$ equals a sum over every pair of observations ("dyads") of the period-by-period products of differences between groups in z_{kt} and $\tilde{x}_{kt}$.*

$$\sum_k n_k \frac{1}{T} \sum_t z_{kt} \left[(x_{kt} - \bar{x}_k) - \left(\bar{x}_t - \bar{\bar{x}}\right) \right] = \sum_k \sum_{\ell > k} n_\ell n_k \frac{1}{T} \sum_t (z_{kt} - z_{\ell t}) \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \tag{A.1}$$

Proof. Assume z_{kt} and x_{kt} are observed in cross-sectional units over time periods, t. The time means $\left(\bar{x}_t - \bar{\bar{x}}\right)$ are weighted averages across units, so:

$$(x_{kt} - \bar{x}_k) - \left(\bar{x}_t - \bar{\bar{x}}\right) = (x_{kt} - \bar{x}_k) - \sum_\ell n_\ell \left(x_{\ell t} - \bar{x}_\ell\right) \tag{A.2}$$

$$= \overbrace{(1 - n_k)}^{\sum_{\ell \neq k} n_\ell} (x_{kt} - \bar{x}_k) - \sum_{\ell \neq k} n_\ell \left(x_{\ell t} - \bar{x}_\ell\right) \tag{A.3}$$

$$= \sum_{\ell \neq k} n_\ell \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \tag{A.4}$$

Substituing into (A.1):

$$\frac{1}{T} \sum_t \sum_k \sum_{\ell \neq k} n_\ell n_k z_{kt} \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \tag{A.5}$$

$$= \sum_k \sum_{\ell > k} n_\ell n_k \frac{1}{T} \sum_t (z_{kt} - z_{kt}) \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \tag{A.6}$$

where (A.6) follows because every dyad (a, b) appears twice, once with $z_{at} \left[(x_{at} - \bar{x}_a) - (\bar{x}_{bt} - \bar{x}_b) \right]$ and once with $z_{bt} \left[(x_{bt} - \bar{x}_b) - (x_{at} - \bar{x}_a) \right]$. ■

Proof of Theorem 1. From Eq. (6) and the definition of $\tilde{D}_{it}$:

$$\frac{\hat{C}(y_{it}, \tilde{D}_{it})}{\hat{V}(\tilde{D}_{it})} = \frac{\frac{1}{NT} \sum_i \sum_t y_{it} \left[\left(D_{it} - \bar{D}_i\right) - \left(\bar{D}_t - \bar{\bar{D}}\right) \right]}{\hat{V}^D} \tag{A.7}$$

Add and subtract deviations of timing-group-by-time means from timing group-means $(\bar{D}_{k(i)t} - \bar{D}_{k(i)})$ in $\tilde{D}_{it}$. I use $k(i)$ to denote the group to which unit i belongs:

$$= \frac{\frac{1}{NT} \sum_i \sum_t y_{it} \left[\overbrace{\left(D_{it} - \bar{D}_i\right) - \left(\bar{D}_{k(i)t} - \bar{D}_{k(i)}\right)}^{=0} + \left(\bar{D}_{k(i)t} - \bar{D}_{k(i)}\right) - \left(\bar{D}_t - \bar{\bar{D}}\right) \right]}{\hat{V}^D} \tag{A.8}$$

The first terms in brackets equal zero because $\tilde{D}_{it}$ varies only at the timing-group-by-year level so $D_{it} = \bar{D}_{k(i)t}$. The covariance between y_{it} and $\left(\bar{D}_{k(i)t} - \bar{D}_{k(i)}\right) - \left(\bar{D}_t - \bar{\bar{D}}\right)$ then collapses to timing-group-by-year averages. I use k instead of $k(i)$ hereafter. Apply Lemma 1 to (A.8):

$$\frac{\sum_k \sum_{\ell > k} n_\ell n_k \frac{1}{T} \sum_t \left(\bar{y}_{kt} - \bar{y}_{\ell t}\right) \left[\left(\bar{D}_{kt} - \bar{D}_k\right) - \left(\bar{D}_{\ell t} - \bar{D}_\ell\right) \right]}{\hat{V}^D} \tag{A.9}$$

Now consider the possible values of $\frac{1}{T} \sum_t \left(\bar{y}_{kt} - \bar{y}_{\ell t}\right) \left[\left(\bar{D}_{kt} - D_k\right) - \left(\bar{D}_{\ell t} - \bar{D}_\ell\right) \right]$. When the control group is either never treated or always treated ($\ell = \infty$ or $\ell < 1$) we have $\left(\bar{D}_{Ut} - \bar{D}_U\right) = 0$ and:

$$-\frac{1}{T} \sum_{t<k} \left(\bar{y}_{kt} - \bar{y}_{Ut}\right) \bar{D}_k + \frac{1}{T} \sum_{t \geq k} \left(\bar{y}_{kt} - \bar{y}_{Ut}\right) \left(1 - \bar{D}_k\right) \tag{A.10}$$

$$= \left[\left(\bar{y}_{kt}^{POST(k)} - \bar{y}_{kt}^{PRE(k)}\right) - \left(\bar{y}_{Ut}^{POST(k)} - \bar{y}_{Ut}^{PRE(k)}\right)\right]\bar{D}_k\left(1-\bar{D}_k\right) \tag{A.11}$$

$$= \hat{\beta}_{kU}^{2x2}\bar{D}_k\left(1-\bar{D}_k\right) \tag{A.12}$$

When $k < \ell < T$, we have:

$$-\frac{1}{T}\sum_{t<k}\left(\bar{y}_{kt}-\bar{y}_{\ell t}\right)\left(\bar{D}_k-\bar{D}_\ell\right)+\frac{1}{T}\sum_{t\in[k,\ell)}\left(\bar{y}_{kt}-\bar{y}_{\ell t}\right)\left(1-\bar{D}_k+\bar{D}_\ell\right)$$
$$-\frac{1}{T}\sum_{t\geq\ell}\left(\bar{y}_{kt}-\bar{y}_{\ell t}\right)\left(\bar{D}_k-\bar{D}_\ell\right) \tag{A.13}$$

$$= -\left(\bar{y}_{kt}^{PRE(k)} - \bar{y}_{\ell t}^{PRE(k)}\right)\left(\bar{D}_k-\bar{D}_\ell\right)\left(1-\bar{D}_k\right)+\left(\bar{y}_{kt}^{MID(k,\ell)} - \bar{y}_{\ell t}^{MID(k,\ell)}\right)\left(\bar{D}_k-\bar{D}_\ell\right)\left(1-\bar{D}_k+\bar{D}_\ell\right)$$
$$-\left(\bar{y}_{kt}^{POST(\ell)} - \bar{y}_{\ell t}^{POST(\ell)}\right)\bar{D}_\ell\left(\bar{D}_k-\bar{D}_\ell\right) \tag{A.14}$$

$$= \left[\left(\bar{y}_{kt}^{MID(k,\ell)} - \bar{y}_{kt}^{PRE(k)}\right) - \left(\bar{y}_{\ell t}^{MID(k,\ell)} - \bar{y}_{\ell t}^{PRE(k)}\right)\right]\left(\bar{D}_k-\bar{D}_\ell\right)\left(1-\bar{D}_k\right)$$
$$+\left[\left(\bar{y}_{\ell t}^{POST(\ell)} - \bar{y}_{\ell t}^{MID(k,\ell)}\right) - \left(\bar{y}_{kt}^{POST(\ell)} - \bar{y}_{kt}^{MID(k,\ell)}\right)\right]\bar{D}_\ell\left(\bar{D}_k-\bar{D}_\ell\right) \tag{A.15}$$

$$= \left(1-\bar{D}_\ell\right)^2\hat{\beta}_{k\ell}^{2x2,k}\left(\frac{\bar{D}_k-\bar{D}_\ell}{1-\bar{D}_\ell}\right)\left(\frac{1-\bar{D}_k}{1-\bar{D}_\ell}\right)+\bar{D}_k^2\hat{\beta}_{k\ell}^{2x2,\ell}\left(\frac{\bar{D}_\ell}{\bar{D}_k}\right)\left(\frac{\bar{D}_k-\bar{D}_\ell}{\bar{D}_k}\right) \tag{A.16}$$

Substituting (A.12) and (A.16) into (A.9) and denoting untreated (or always treated) groups by U, earlier treated groups in a dyad by k, and later treated groups by ℓ, establishes Eq. (10a):

$$\frac{1}{\hat{V}^D}\left\{\sum_{k\neq U}\left(n_k+n_U\right)^2 n_{kU}\left(1-n_{kU}\right)\bar{D}_k\left(1-\bar{D}_k\right)\hat{\beta}_{kU}^{2x2}\right.$$
$$+\sum_{k\neq U}\sum_{\ell>k}\left[\left(\left(n_k+n_\ell\right)\left(1-\bar{D}_\ell\right)\right)^2 n_{k\ell}\left(1-n_{k\ell}\right)\left(\frac{\bar{D}_k-\bar{D}_\ell}{1-\bar{D}_\ell}\right)\left(\frac{1-\bar{D}_k}{1-\bar{D}_\ell}\right)\hat{\beta}_{k\ell}^{2x2,k}\right.$$
$$\left.\left.+\left(\left(n_k+n_\ell\right)\bar{D}_k\right)^2 n_{k\ell}\left(1-n_{k\ell}\right)\left(\frac{\bar{D}_\ell}{\bar{D}_k}\right)\left(\frac{\bar{D}_k-\bar{D}_k}{\bar{D}_k}\right)\hat{\beta}_{k\ell}^{2x2,\ell}\right]\right\} \tag{A.17}$$

$$\frac{\sum_{k\neq U}\left(n_k+n_U\right)^2\hat{V}_{kU}^D\hat{\beta}_{kU}^{2x2}+\sum_{k\neq U}\sum_{\ell>k}\left[\left(\left(n_k+n_\ell\right)\left(1-\bar{D}_\ell\right)\right)^2\hat{V}_{k\ell}^{D,k}\hat{\beta}_{k\ell}^{2x2,k}+\left(\left(n_k+n_\ell\right)\bar{D}_k\right)^2\hat{V}_{k\ell}^{D,\ell}\hat{\beta}_{k\ell}^{2x2,\ell}\right]}{\hat{V}^D} \tag{A.18}$$

The denominator, $\hat{V}^D$ the variance of $\tilde{D}_{it}$, equals the sum of the terms multiplying the $\hat{\beta}^{2x2}$'s in (A.18). This follows by substituting D_{it} for y_{it} in each $\hat{\beta}^{2x2}$ and noting that every term equals 1. Therefore, the weights sum to one: $\sum_{k\neq U}s_{kU}+\sum_{k\neq U}\sum_{\ell>k}\left[s_{k\ell}^k+s_{k\ell}^\ell\right]=1$. ■

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.03.014.

References

Abadie, Alberto, 2005. Semiparametric difference-in-differences estimators. Rev. Econom. Stud. 72 (1), 1–19.
Abadie, Alberto, Diamond, Alexis, Hainmueller, Jens, 2010. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. J. Amer. Statist. Assoc. 105 (490), 493–505. http://dx.doi.org/10.1198/jasa.2009.ap08746.
Allcott, Hunt, 2015. Site selection bias in program evaluation. Q. J. Econ. 130 (3), 1117–1165. http://dx.doi.org/10.1093/qje/qjv015.
Almond, Douglas, Hoynes, Hilary W., Schanzenbach, Diane, 2011. Inside the war on poverty: The impact of food stamps on birth outcomes. Rev. Econ. Stat. 93 (2), 387–403. http://dx.doi.org/10.2307/23015943.
Angrist, Joshua D., 1991. Grouped-data estimation and testing in simple labor-supply models. J. Econometrics 47 (2), 243–266. http://dx.doi.org/10.1016/0304-4076(91)90101-I.
Angrist, Joshua D., Krueger, Alan B., 1999. Chapter 23 - empirical strategies in labor economics. In: Ashenfelter, Orley C., Card, David (Eds.), HandBook of Labor Economics. Elsevier, pp. 1277–1366.
Angrist, Joshua D., Pischke, Jörn-Steffen, 2009. Mostly Harmless Econometrics : An Empiricist's Companion. Princeton University Press, Princeton.
Angrist, Joshua D., Pischke, Jörn-Steffen, 2015. Mastering 'Metrics : The Path from Cause to Effect. Oxford: Princeton University Press, Princeton.
Athey, Susan, Imbens, Guido W., 2006. Identification and inference in nonlinear difference-in-differences models. Econometrica 74 (2), 431–497.
Athey, Susan, Imbens, Guido W., 2018. Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption. Working Paper.
Ben-Michael, Eli, Feller, Avi, Rothstein, Jesse, 2019. Synthetic Controls and Weighted Event Studies with Staggered Adoption. Working Paper.
Bertrand, Marianne, Duflo, Esther, Mullainathan, Sendhil, 2004. How much should we trust differences-in-differences estimates?. Q. J. Econ. 119 (1), 249–275.

Bilinski, Alyssa, Hatfield, Laura, 2019. Nothing to See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions. Working Paper.

Bitler, Marianne P., Gelbach, Jonah B., Hoynes, Hilary W., 2003. Some evidence on race, welfare reform, and household income. Amer. Econ. Rev. 93 (2), 293–298. http://dx.doi.org/10.2307/3132242.

Blinder, Alan S., 1973. Wage discrimination: Reduced form and structural estimates. J. Hum. Resour. 8 (4), 436–455. http://dx.doi.org/10.2307/144855.

Borusyak, Kirill, Jaravel, Xavier, 2017. Revisiting Event Study Designs. Harvard University Working Paper.

Callaway, Brantly, Li, Tong, Oka, Tatsushi, 2018. Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. J. Econometrics 206 (2), 395–413. http://dx.doi.org/10.1016/j.jeconom.2018.06.008, https://www.sciencedirect.com/science/article/pii/S0304407618301027.

Callaway, Brantly, Sant'Anna, Pedro H.C., 2020. Difference-in-differences with multiple time periods. J. Econometrics http://dx.doi.org/10.1016/j.jeconom.2020.12.001.

Cameron, Colin, Trivedi, Pravin, 2005. Microeconometrics : Methods and Applications. Cambridge University Press, Cambridge ; New York.

Cengiz, Doruk, Dube, Arindrajit, Lindner, Attila, Zipperer, Ben, 2019. The effect of minimum wages on low-wage jobs*. Q. J. Econ. 134 (3), 1405–1454. http://dx.doi.org/10.1093/qje/qjz014.

de Chaisemartin, Clément, D'Haultfœuille, Xavier, 2018. Fuzzy differences-in-differences. Rev. Econom. Stud. 85 (2), 999–1028. http://dx.doi.org/10.1093/restud/rdx049.

de Chaisemartin, Clément, D'Haultfœuille, Xavier, 2020. Two-way fixed effects estimators with heterogeneous treatment effects. Amer. Econ. Rev. 110 (9), 2964–2996. http://dx.doi.org/10.1257/aer.20181169.

Chernozhukov, Victor, Fernández-Val, Iván, Hahn, Jinyong, Newey, Whitney, 2013. Average and quantile effects in nonseparable panel models. Econometrica 81 (2), 535–580. http://dx.doi.org/10.3982/ECTA8405.

Chyn, Eric, 2018. Moved to opportunity: The long-run effect of public housing demolition on labor market outcomes of children. Amer. Econ. Rev. 108 (10), 3028–3056. http://dx.doi.org/10.1257/aer.20161352.

Cunningham, Scott, 2021. Causal Inference: The Mixtape. Yale University Press.

Deaton, Angus, 1997. The Analysis of Household Surveys : A Microeconometric Approach to Development Policy. Johns Hopkins University Press, Baltimore, MD.

Deshpande, Manasi, Li, Yue, 2019. Who is screened out? Application costs and the targeting of disability programs. Amer. Econ. J.: Econ. Policy 11 (4), 213–248. http://dx.doi.org/10.3386/w23472.

Fadlon, Itzik, Nielsen, torben Heien, 2015. Family Labor Supply Responses to Severe Health Shocks. National Bureau of Economic Research Working Paper Series (21352), http://dx.doi.org/10.3386/w21352.

Frisch, Ragnar, Waugh, Frederick V., 1933. Partial time regressions as compared with individual trends. Econometrica 1 (4), 387–401. http://dx.doi.org/10.2307/1907330.

Gibbons, Charles, E., Serrato, Juan Carlos Suárez, Michael Urbancic, B., 2018. Broken or fixed effects?. J. Econometr. Methods.

Goodman, Joshua, 2017. The Labor of Division: Returns to Compulsory High School Math Coursework. National Bureau of Economic Research Working Paper Series (23063), http://dx.doi.org/10.3386/w23063.

Goodman-Bacon, Andrew, Goldring, Thomas, Nichols, Austin, 2019. Bacondecomp: Stata module for decomposing difference-in-differences estimation with variation in treatment timing. Stata Command.

Grosz, Michael, Miller, Douglas L., Shenhav, Na'ama, 2018. All in the Family: Assessing the External Validity of Family Fixed Effects Estimates and the Long Term Impact of Head Start. Working Paper.

Haines, Michael R., ICPSR, 2010. Historical, Demographic, Economic, and Social Data: The United States, 1790-2002. ICPSR [distributor].

Heckman, James J., Lalonde, Robert J., Smith, Jeffrey A., 1999. Chapter 31 - the economics and econometrics of active labor market programs. In: Ashenfelter, Orley C., Card, David (Eds.), HandBook of Labor Economics. Elsevier, pp. 1865–2097.

Holland, Paul W., 1986. Statistics and causal inference. J. Amer. Statist. Assoc. 81 (396), 945–960. http://dx.doi.org/10.2307/2289064.

Imai, K., Kim, I., 2021. On the use of two-way fixed effects regression models for causal inference with panel data. Polit. Anal. 29 (3), 405–415. http://dx.doi.org/10.1017/pan.2020.33.

Imbens, Guido W., Angrist, Joshua D., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475. http://dx.doi.org/10.2307/2951620.

Jacobson, Louis S., LaLonde, Robert J., Sullivan, Daniel G., 1993. Earnings losses of displaced workers. Amer. Econ. Rev. 83 (4), 685–709. http://dx.doi.org/10.2307/2117574.

Joseph Hotz, V., Imbens, Guido W., Mortimer, Julie H., 2005. Predicting the efficacy of future training programs using past experiences at other locations. J. Econometrics 125 (1), 241–270. http://dx.doi.org/10.1016/j.jeconom.2004.04.009.

Kitagawa, Evelyn M., 1955. Components of a difference between two rates. J. Amer. Statist. Assoc. 50 (272), 1168–1194. http://dx.doi.org/10.2307/2281213.

Krolikowski, Pawel, 2017. Choosing a control group for displaced workers. http://dx.doi.org/10.1177/0019793917743707, ILR Review:0019793917743707.

Lee, Jin Young, Solon, Gary, 2011. The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates. National Bureau of Economic Research Working Paper Series No. 16773.

Lovell, Michael C., 1963. Seasonal adjustment of economic time series and multiple regression analysis. J. Amer. Statist. Assoc. 58 (304), 993–1010. http://dx.doi.org/10.2307/2283327.

Malkova, Olga, 2017. Can maternity benefits have long-term effects on childbearing? Evidence from soviet Russia. Rev. Econ. Stat. http://dx.doi.org/10.1162/REST_a_00713.

Meer, Jonathan, West, Jeremy, 2016. Effects of the minimum wage on employment dynamics. J. Hum. Resour. 51 (2), 500–522. http://dx.doi.org/10.3386/w19262.

Meyer, Bruce D., 1995. Natural and quasi-experiments in economics. J. Bus. Econom. Statist. 13 (2), 151–161. http://dx.doi.org/10.2307/1392369.

Neumark, David, Ian Salas, J.M., Wascher, William, 2014. Revisiting the minimum wage—Employment debate: Throwing out the baby with the bathwater?. ILR Rev. 67 (3_suppl), 608–648. http://dx.doi.org/10.1177/00197939140670S307.

Oaxaca, Ronald, 1973. Male-female wage differentials in urban labor markets. Internat. Econom. Rev. 14 (3), 693–709. http://dx.doi.org/10.2307/2525981.

Oster, Emily, 2016. Unobservable selection and coefficient stability: Theory and evidence. J. Bus. Econom. Statist. 1–18. http://dx.doi.org/10.1080/07350015.2016.1227711.

Perron, Pierre, 2006. Dealing with Structural Breaks. Working Paper.

Rambachan, Ashesh, Roth, Jonathan, 2019. An Honest Approach to Parallel Trends. Working Paper.

Rosenbaum, Paul R., 1984. The consquences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Statist. Soc. Ser. A (Gen.) 147 (5), 656–666. http://dx.doi.org/10.2307/2981697.

Rubin, Donald B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66 (5), 688–701. http://dx.doi.org/10.1037/h0037350.

Sant'Anna, Pedro, Zhao, Jun, 2018. Doubly Robust Difference-in-Differences Estimators. Working Paper.
Sant'Anna, H.C., Zhao, Jun, 2020. Doubly robust difference-in-differences estimators. J. Econometrics 219 (1), 101–122. http://dx.doi.org/10.1016/j.jeconom.2020.06.003.
Shore-Sheppard, Lara D., 2009. Stemming the tide? The effect of expanding medicaid eligibility on health insurance coverage. B.E. J. Econ. Anal. Policy 8 (2).
Sloczynski, Tymon, 2017. A General Weighted Average Represetnation of the Ordinary and Two-Stage Least Squares EstimandS. Working Paper.
Snow, John, 1855. On the Mode of Communication of Cholera. Edited by John Churchill. Second ed. London.
Solon, Gary, Haider, Steven J., Wooldridge, Jeffrey M., 2015. What are we weighting for?. J. Hum. Resour. 50 (2), 301–316.
Stevenson, Betsey, Wolfers, Justin, 2006. Bargaining in the shadow of the law: Divorce laws and family distress. Q. J. Econ. 121 (1), 267–288.
Strezhnev, Anton, 2018. Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs. Working Paper.
Sun, Liyang, Abraham, Sarah, 2020. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J. Econometrics http://dx.doi.org/10.1016/j.jeconom.2020.09.006.
Surveillance, Epidemiology, Results, End, 2013. Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969-2011). Surveillance Research Program, Surveillance Systems Branch,
Walters, Christopher R., 2018. The demand for effective charter schools. J. Polit. Econ. 126 (6), 2179–2223.
Wolfers, Justin, 2006. Did unilateral divorce laws raise divorce rates? A reconciliation and new results. Amer. Econ. Rev. 96 (5), 1802–1820.
Wooldridge, Jeffrey M., 2001. Asymptotic properties of weighted M-estimators for standard stratified samples. Econometric Theory 17 (2), 451–470.
Wooldridge, Jeffrey M., 2005. Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. Rev. Econ. Stat. 87 (2), 385–390.
Wooldridge, Jeffrey M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed. MIT Press, Cambridge, Mass.