DIFFERENCE-IN-DIFFERENCES WITH VARIATION IN TREATMENT TIMING[*]

Andrew Goodman-Bacon

August 2020

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 general estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. This defines the DD estimand and identifying assumption, a generalization of common trends. I discuss how to interpret DD estimates and propose a new balance test. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls.

[*] Department of Economics, Vanderbilt University (email: andrew.j.goodman-bacon@vanderbilt.edu) and NBER. I thank Michael Anderson, 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.

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): $\left(\overline{y}_{TREAT}^{POST} - \overline{y}_{TREAT}^{PRE}\right) - \left(\overline{y}_{CONTROL}^{POST} - \overline{y}_{CONTROL}^{PRE}\right)$. 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:

$$y_{it} = \gamma + \gamma_i TREAT_i + \gamma_t POST_t + \beta^{2x2} TREAT_i \times POST_t + u_{it} \, . \qquad (1)$$

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. All econometrics textbooks and survey articles describe this structure,[2] and recent methodological extensions build on it.[3]

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 (α_i) and time periods (α_t), and a treatment dummy (D_{it}):

$$y_{it} = \alpha_{i\cdot} + \alpha_{\cdot t} + \beta^{DD} D_{it} + e_{it} \; . \qquad (2)$$

[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, but is not limited to, Angrist and Krueger (1999), Angrist and Pischke (2009), Heckman, Lalonde, and Smith (1999), Meyer (1995), Cameron and Trivedi (2005), Wooldridge (2010).
[3] Inverse propensity score reweighting: Abadie (2005), synthetic control: Abadie, Diamond, and Hainmueller (2010), changes-in-changes: Athey and Imbens (2006), quantile treatment effects: Callaway, Li, and Oka (forthcoming).
[4] Half of the 93 DD papers published in 2014/2015 in 5 general interest or field journals had variation in timing.

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, Duflo, and Mullainathan 2004, p. 250), and consequently lack well-defined strategies to test the validity of the DD design with timing. We have limited understanding of the treatment effect parameter that regression DD identifies. Finally, we often cannot evaluate when alternative specifications will work or why they change estimates.[6]

This paper shows that the two-way fixed effects DD estimator in (2) 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 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 DD estimates a variance-weighted average of treatment effect parameters sometimes with "negative weights" (Abraham and Sun 2018, Borusyak and Jaravel 2017, de Chaisemartin and D'HaultfŒuille forthcoming).[7] When treatment

[5] Imai, Kim, and Wang (2018) 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, Salas, and Wascher (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, Gelbach, and Hoynes 2003), or it provided simulation evidence (Meer and West 2013). Recent research on the weighting of heterogeneous treatment effects does not provide this intuition. de Chaisemartin and D'HaultfŒuille (forthcoming, p 7) and Borusyak and Jaravel (2017, p 10-11) describe these same weights as coming from an auxiliary

effects do not change over time, DD yields a variance-weighted average of cross-group treatment effects and all weights are positive. Negative weights only arise when effects vary over time. The DD decomposition shows why: when already-treated units act as controls, *changes* in their treatment effects over time get subtracted from the DD estimate. This does not imply a failure of the *design*, but it does caution against summarizing time-varying effects with a single-coefficient.

Next I use the DD decomposition to define "common trends" with timing variation. Each 2x2 DD relies on pairwise common trends in untreated potential outcomes, and the overall identifying assumption is an average of these terms using the variance-based decomposition weights. The extent to which a given 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. The earliest and/or latest treated units have low treatment variance, and can get *more* weight as controls than treatments. In designs without untreated units they always do. I construct a balance test derived from the estimator itself that improves on existing strategies that test between treated/untreated or earlier/later treated units.

Finally, I develop simple tools to describe the general DD design and evaluate why estimates change across specifications.[8] Plotting the 2x2 DDs against their weight displays heterogeneity in the estimated components and shows which terms or groups matter most. Summing the weights on the timing comparisons versus treated/untreated comparisons quantifies "how much" of the variation comes from timing (a common question in practice), and provides practical guidance on how well the two-way fixed effects estimator works compared to alternative

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, Goldring, and Nichols 2019).

estimators (Abraham and Sun 2018, Borusyak and Jaravel 2017, Callaway and Sant'Anna 2018, Imai, Kim, and Wang 2018, Strezhnev 2018, Ben-Michael, Feller, and Rothstein 2019). Comparing DD 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 implicitly introduce new unintended sources of variation such as comparisons between units with the same treatment but different covariates.

To demonstrate these methods I replicate Stevenson and Wolfers (2006) study of the effect of unilateral divorce laws on female suicide rates. The two-way fixed effects estimator suggest that unilateral divorce leads to 3 fewer suicides per million women. More than a third of the identifying variation comes from treatment timing and the rest comes from comparisons to states with no reforms during the sample period. Event-study estimates show that the treatment effects vary strongly over time, however, which biases many of the timing comparisons. The DD estimate (-3.08) is therefore a misleading summary of the average post-treatment effect (about -5). My proposed balance test detects higher per-capita income and male/female sex ratios in reform states, in contrast to joint tests across timing groups, which cannot reject the null of balance. Much of the sensitivity across specifications comes from changes in weights, or a small number of 2x2 DD's, and need not indicate bias.

I. 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 uses the two-way fixed effects regression in equation (2) (Cameron and Trivedi 2005 pg. 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 two-way fixed effects DD estimator into a weighted average of simple 2x2 DD estimators.

Figure 1 plots a simple data structure that includes treatment timing. Assume a a balanced panel dataset with T periods (t) and N cross-sectional units (i) that belong to either an untreated group, U; an early treatment group, k, which receives a binary treatment at t_k^*; and a late treatment group, ℓ, which receives the binary treatment at $t_\ell^* > t_k^*$.

Throughout the paper I use "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 group's sample share is n_k and the share of time it spends treated is $\overline{D}_k$. I use $\overline{y}_b^{POST(a)}$ to denote the sample mean of y_{it} for units in group b during group a's post period, $[t_a^*, T]$. ($\overline{y}_b^{PRE(a)}$ is defined similarly.)

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

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

[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.

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

The challenge in this setting has been to articulate how estimates of equation (2) compare the groups and times depicted in figure 1. We do, however, have clear intuition, for 2x2 designs in which one group's treatment status changes and another's does not. In the three-group case we could form four such designs estimable by equation (1) on subsamples of groups and time periods. Figure 2 plots them.

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

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

Note that I use 2x2 to refer to two *groups* of periods (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 group's treatment status changes. Before t_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, $MID(k, \ell)$, and group k's pre-period, $PRE(k)$:

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

The opposite situation, shown in panel D, arises after t_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 $POST(\ell)$ and $MID(k, \ell)$:

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

The already-treated units in 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 and only some time peroids. $\widehat{\beta}_{k\ell}^{2x2,k}$ uses group ℓ's pre-period so its share is $(n_k + n_\ell)(1 - \overline{D}_\ell)$, while $\widehat{\beta}_{k\ell}^{2x2,\ell}$ only uses group k's post-period so its share 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:

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

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

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

where $n_{ab} \equiv \frac{n_a}{n_a + n_b}$ is the relative size of groups in each pair. The first part of each pairwise variance measures how concentrated the groups are. 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, rescaled by the size of the relevant time window 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 any two-way fixed effects DD estimator is an average of well-understood 2x2 DD estimators, like those plotted in figure 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 $k = 1, \ldots, K$ groups of units ordered by the time when they receive a binary treatment, $t_k^* \in (1, T]$. There may be one group, U, that includes unites that never receives treatment or are always treated. The OLS estimate, $\widehat{\beta}^{DD}$, in a two-way fixed-effects regression (2) *is a weighted average of all possible two-by-two DD estimators.*

$$\widehat{\beta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\beta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{\ell > k} \left[s_{k\ell}^{k} \widehat{\beta}_{k\ell}^{2x2,k} + s_{k\ell}^{\ell} \widehat{\beta}_{k\ell}^{2x2,\ell} \right]. \qquad (10a)$$

Where the 2x2 DD estimators are:

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

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

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

The weights are:

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

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

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

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

Proof: See appendix A.

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

The weights on each 2x2 DD combine the absolute size of the subsample, the relative size of the treatment and control groups in the subsample, and the timing of treatment in the subsample.[10] The first part is the size of the subsample squared. The second part of each weight is the subsample variance from equations (7)-(9). The variance is larger when the two groups are closer in size ($n_{kU} \approx 0.5$) and when treatment occurs closer to the middle of the relevant 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 figure 2, the 2x2 DDs with 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 t_k^* and t_ℓ^* so that $\overline{D}_k = 0.66$ and $\overline{D}_\ell = 0.16$. For treated/untreated DDs, $s_{kU} > s_{\ell U}$ because group k is treated closer to the middle of the panel than group ℓ and therefore has a higher treatment variance: $\overline{D}_k(1-\overline{D}_k) = 0.22 > 0.13 = \overline{D}_\ell(1-\overline{D}_\ell)$. This is also true for the timing-only 2x2 DDs. Group k's treatment share within group ℓ's pre-period is $\frac{\overline{D}_k - \overline{D}_\ell}{1-\overline{D}_\ell} = \frac{0.66-0.16}{0.84} = 0.59$, but group ℓ's pre-period accounts for $1-\overline{D}_\ell = 0.84$ share of the observations. Group ℓ's treatment share within

[10] 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:

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

Similarly, two-stage least squares uses samples sizes and variances to "efficiently combine alternative Wald estimates" (Angrist 1991). Gibbons, Serrato, and Urbancic (2018) show a nearly identical 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.

group k's post-period, on the other hand, is $\frac{\overline{D}_\ell}{\overline{D}_k} = \frac{0.16}{0.66} = 0.24$, and group k's post-period accounts for $\overline{D}_k = 0.66$ share of the observations. Therefore, $s_{k\ell}^k > s_{k\ell}^\ell$ because $\hat{\beta}_{k\ell}^k$ has a higher variance from treatment timing alone and it uses more data: $(1-\overline{D}_\ell)^2 \frac{\overline{D}_k - \overline{D}_\ell}{1-\overline{D}_\ell}\frac{1-\overline{D}_k}{1-\overline{D}_\ell} = 0.17 > 0.08 = \overline{D}_k{}^2 \frac{\overline{D}_\ell}{\overline{D}_k}\frac{\overline{D}_k - \overline{D}_\ell}{\overline{D}_k}$. Scaling by the overall variance of $\widetilde{D}_{it}$ shows that the weights are $\{s_{kU}, s_{\ell U}, s_{k\ell}^k, s_{k\ell}^\ell\} = \{0.37, 0.22, 0.28, 0.13\}$.

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 figure 2. In that case, $\overline{D}_k = 0.83$ and $\overline{D}_\ell = 0.58$ and group ℓ is treated closer to the middle of the panel than group k. The weights change to $\{s_{kU}, s_{\ell U}, s_{k\ell}^k, s_{k\ell}^\ell\} = \{0.25, 0.43, 0.07, 0.25\}$. 2x2 DDs in which 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. Therefore panel length alone could change DD 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 figure 2:

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

Both 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. In (11), $\mu_{k\ell}$ simplifies to $\frac{1-\overline{D}_k}{1-(\overline{D}_k - \overline{D}_\ell)}$, which falls as $\overline{D}_k$ gets

closer to one (t_k^* gets closer to the first time period). In other words, the group treated close to the middle of the panel gets more weight. In the three group example $\mu_{k\ell} = 0.34/0.5 = 0.68$.[11]

A. Alternative Decompositions

Alternative algebraic decompositions are possible, but neither match the structure of the two-way fixed effects specification in (2) nor yield intuitive weights. Strezhnev (2018, equation 15) decomposes $\widehat{\beta}^{DD}$ 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 $\widehat{\beta}^{DD}$ into terms representing causal effects over different time-horizons. My "group-level" decomposition on the other hand yields well-defined intuitive weights by grouping 2x2 terms according to the identifying variation (pre/post, treatment/control) that unites them. See Appendix F for more details on the relationship between decompositions.

II. 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). The outcome is $y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$, where Y_{it}^1 is unit i's treated outcome at time t, and Y_{it}^0 is the corresponding untreated outcome. Following Callaway and Sant'Anna (2018, p 7) define the ATT for timing group k at time τ (the "group-time average treatment effect"): $ATT_k(\tau) \equiv E[Y_{i\tau}^1 - Y_{i\tau}^0 | k]$. Because regression DD averages outcomes in pre- and post-periods, I define the average $ATT_k(\tau)$ in a date range W (with T_W periods):

[11] 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, $POST(\ell)$.

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

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

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

Applying this notation to the 2x2 DDs in equations (4)-(6), adding and subtracting post-period counterfactual 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:

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

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

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

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

Substituting equations (14a)-(14c) into the DD decomposition theorem expresses the probability limit of the two-way fixed effects DD estimator (assuming that T is fixed and N grows) in terms of potential outcomes and separates the estimand from the identifying assumptions:

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

The first term in (15) is the two-way fixed effects DD estimand, which I call the "variance-weighted average treatment effect on the treated" (VWATT):

$$VWATT \equiv \sum_{k \neq U} \sigma_{kU} \, ATT_k\big(POST(k)\big)$$

$$+ \sum_{k \neq U} \sum_{\ell > k} \left[\sigma^k_{k\ell} ATT_k\big(MID(k,\ell)\big) + \sigma^\ell_{k\ell} ATT_\ell\big(POST(\ell)\big) \right] . \qquad (15a)$$

The σ terms are probability limits of the weights in (10a).[12] 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:

$$VWCT \equiv \sum_{k \neq U} \sigma_{kU} \left[\Delta Y^0_k\big(POST(k), PRE(k)\big) - \Delta Y^0_U\big(POST(k), PRE(k)\big) \right]$$

$$+ \sum_{k \neq U} \sum_{\ell > k} \left[\sigma^k_{k\ell} \left\{ \Delta Y^0_k\big(MID(k,\ell), PRE(k)\big) - \Delta Y^0_\ell\big(MID(k,\ell), PRE(k)\big) \right\} \right.$$

$$\left. + \sigma^\ell_{k\ell} \left\{ \Delta Y^0_\ell\big(POST(\ell), MID(k,\ell)\big) - \Delta Y^0_k\big(POST(\ell), MID(k,\ell)\big) \right\} \right] . \quad (15b)$$

Like VWATT, VWCT is also an average of the difference in counterfactual trends between pairs of 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 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 unit's post-period with respect to another unit's treatment timing:

[12] 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).

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

Because already-treated groups sometimes act as controls, the 2x2 estimators in equation (14c) subtract average changes in their untreated outcomes *and* their treatment effects. Of course $\Delta ATT \neq 0$ only if treatment effects vary over time, but when they do, equation (15c) defines the resulting bias in the DD. This does not mean that the research design is invalid. In this case specifications such as an event-study (Jacobson, LaLonde, and Sullivan 1993), "stacked DD" (Abraham and Sun 2018, Deshpande and Li 2017, Fadlon and Nielsen 2015), or reweighting estimators (Callaway and Sant'Anna 2018) may be more appropriate.[13]

A. Interpreting the DD Estimand

When the treatment effect is a constant, $ATT_k(W) = ATT$, $\Delta ATT = 0$, 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.

i. Effects that vary across units but not over time

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

[13] Recent DD research comes to related conclusions about DD with timing, but does not describe the full estimator as in equation (15). Abraham and Sun (2018), Borusyak and Jaravel (2017), and de Chaisemartin and D'HaultfŒuille (forthcoming) 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 equation (15), ensures an interpretable estimand ($VWATT$) and 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.

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

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) equal the sum of the decomposition weights for all the terms in which group k acts as the treatment group, defined as w_k^T.

In general, $w_k^T \neq n_k^*$, so the parameter does not equal the sample ATT.[14] 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. VWATT lies along the bias/variance tradeoff: the variance weights come from the fact that OLS combines 2x2 DDs efficiently but potentially moves the point estimate away from, say, the sample ATT. This tradeoff may be worthwhile. If estimates determine the level of some policy that affects social welfare, then the optimal estimator minimizes mean squared error (see appendix B and Kasy 2018). If VWATT is close to the ATT (for example) and has lower variance, it may be preferable by this criterion.

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, Imbens, and Mortimer 2005) and energy conservation programs (Allcott 2015) matches 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

[14] Abraham and Sun (2018), Borusyak and Jaravel (2017), Chernozhukov et al. (2013), de Chaisemartin and D'HaultfŒuille (forthcoming), Gibbons, Serrato, and Urbancic (2018), Wooldridge (2005) all make a similar observation. The DD decomposition theorem, provides a new solution for the relevant weights.

follow from "reverse Roy" selection where units with the smallest effects select into treatment first, which describes the take up of housing vouchers (Chyn forthcoming) and charter school applications (Walters forthcoming). 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 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 group is very large. Conversely, weighting matters less if the ATT_k's are similar, which one can evaluate by aggregating each group's 2x2 DD estimates from the decomposition theorem. Finally, one could directly compare VWATT to point estimates of a particular parameter of interest. Several alternative estimators give differently weighted averages of ATT's (Abraham and Sun 2018, Callaway and Sant'Anna 2018, de Chaisemartin and D'HaultfŒuille forthcoming).

ii. Effects that vary over time but not across units

Time-varying treatment effects generate heterogeneity across the 2x2 DDs by averaging over different post-treatment windows, up-weight short-run effects most likely to appear in the small windows between timing groups, and bias estimates away from VWATT because $\Delta ATT \neq 0$. Equations (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, figure 3 plots a case where counterfactual outcomes are identical, but the treatment effect is a linear trend-break, $Y_{it}^1 = Y_{it}^0 + \phi \cdot (t - t_i^* + 1)$ (see Meer and West 2013). $\hat{\beta}_{k\ell}^{2x2,k}$ uses group ℓ as a control group during its pre-period and identifies the ATT during

the middle window in which treatment status varies: $\phi\frac{(t_\ell^*-t_k^*+1)}{2}$. $\hat{\beta}_{k\ell}^{2x2,\ell}$ however, is biased because the control group (k) experiences a trend in outcomes due to the treatment effect:[15]

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

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

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

The entire two-group timing estimate can be wrong signed if there is sufficiently more weight on $\hat{\beta}_{k\ell}^{2x2,\ell}$ than $\hat{\beta}_{k\ell}^{2x2,k}$ (ie. $\sigma_{k\ell}^{\ell} > \sigma_{k\ell}^{k}$). In figure 3, for example, both units are treated equally close to the ends of the panel, so $\sigma_{k\ell}^{k} = \sigma_{k\ell}^{\ell}$ and the estimated DD effect equals $\frac{\phi}{2}$, even though both units experience treatment effects as large as $\phi\cdot[T-(t_k^*-1)]$. Summarizing time-varying effects using equation (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.[16]

---

[15] The average of the effects for group k during any set of positive event-times is just ϕ times the average event-time. The $MID(k,\ell)$ period contains event-times 0 through $t_\ell^*-(t_k^*-1)$ and the $POST(\ell)$ period contains event-times $t_\ell^*-(t_k^*-1)$ through $T-(t_k^*-1)$, so we have:

$$ATT_k\big(MID(k,\ell)\big) = \phi\frac{(t_\ell^*-t_k^*)(t_\ell^*-t_k^*+1)}{2(t_\ell^*-t_k^*)} = \phi\frac{t_\ell^*-t_k^*+1}{2},$$

$$ATT_k\big(POST(\ell)\big) = \phi(t_\ell^*-t_k^*) + \phi\frac{T-t_\ell^*+2}{2},$$

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

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

Another way to see this, as noted in figure 3, is that average outcomes in the treatment group are always below average outcomes in the early group in the $POST(\ell)$ period and the difference equals the maximum size of the treatment effect in group k at the end of the $MID(k,\ell)$ period: $\phi\cdot(t_\ell^*-t_k^*+1)$. Average outcomes for the late group are also below average outcomes in the early group in the $MID(k,\ell)$ period, but by the *average* amount of the treatment effect in group k during the $MID(k,\ell)$ period: $\phi\frac{(t_\ell^*-t_k^*+1)}{2}$. 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.

[16] Borusyak and Jaravel (2017) show that 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 for treatment effect estimation, though, because "effects" must occur after treatment. Job displacement provides a clear example (Jacobson, LaLonde, and Sullivan 1993, Krolikowski 2017). Comparisons based on displacement timing cannot identify whether all displaced workers

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 Abraham and Sun 2018). Fadlon and Nielsen (2015) and Deshpande and Li (2017) match treated units with controls that receive treatment a given amount of time later and gives an average of $\widehat{\beta}_{k\ell}^{2x2,k}$ terms with a fixed post-period (see similar proposals in Abraham and Sun 2018, Borusyak and Jaravel 2017, de Chaisemartin and D'HaultfŒuille forthcoming). Callaway and Sant'Anna (2018) discuss how to aggregate heterogeneous treatment effects and develop a reweighting estimator to do so.

B. What is the identifying assumption and how should we test it?

The preceding analysis maintained the assumption of *equal* counterfactual trends across groups, but (15) shows that when $\Delta ATT = 0$ identification of $VWATT$ only requires $VWCT = 0$. Assuming linear counterfactual trends ($\Delta Y_k^0 \equiv Y_{k,t}^0 - Y_{k,t-1}^0$) leads to a convenient approximation to $VWCT$:[17]

$$VWCT \approx \sum_{k \neq U} \Delta Y_k^0 \left[\sigma_{kU} + \sum_{j=1}^{k-1} \left(\sigma_{jk}^k - \sigma_{jk}^j \right) + \sum_{j=k+1}^{K} \left(\sigma_{kj}^k - \sigma_{kj}^j \right) \right] - \Delta Y_U^0 \sum_{k \neq U} \sigma_{kU}$$

$$= \sum_k \Delta Y_k^0 \left[w_k^T - w_k^C \right]. \quad (19)$$

Equation (19) generalizes the definition of common trends to the timing case and shows how a given group's counterfactual trend biases the overall estimate. To illustrate, assume there is a positive differential trend in group k only: $\Delta Y_k^0 > 0$. This will bias $\widehat{\beta}_{kU}^{2x2}$ by ΔY_k^0 which gets a weight of σ_{kU} in the full estimate. When comparing group k to other timing groups, however, biases offset each other. For the comparisons to group 1, for example, units in group k act as

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).

[17] 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 equation (19) holds.

treatments in $\widehat{\beta}_{1k}^{2x2,k}$ and the bias equals ΔY_k^0 and is weighted by σ_{1k}^k. But in $\widehat{\beta}_{1k}^{2x2,1}$, the units in group k act as controls so the bias equals $-\Delta Y_k^0$ and is weighted by σ_{1k}^1. On net the bias in $\widehat{\beta}_{1k}^{2x2}$ is ambiguous: $\Delta Y_k^0\left(\sigma_{1k}^k - \sigma_{1k}^1\right)$.

Similar expressions hold for the comparison of group k to every other group, and the total weight on each group's counterfactual trend equals the difference between the total weight it gets when it acts as a treatment group—w_k^T from equation (16)—minus the total weight it gets when it acts as a control group—$w_k^C \equiv \sum_{j=1}^{k-1} \sigma_{jk}^j + \sum_{j=k+1}^{K} \sigma_{kj}^j$. This difference is a new result that maps (linear) differential trends to bias.[18] A positive trend in group k induces positive bias when $w_k^T - w_k^C > 0$, negative bias when $w_k^T - w_k^C < 0$, and no bias when $w_k^T - w_k^C = 0$.[19]

Figure 4 plots $w_k^T - w_k^C$ 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 t^* approaches 1 or T, $w_k^T - w_k^C$ becomes negative; some timing groups effectively act as controls. This defines "the" control group in timing-only designs: all groups are controls in *some* terms, but the earliest and/or latest units necessarily get more weight as controls than treatments.

[18] 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, Hoynes, and Schanzenbach (2011, p 389-190) 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."

[19] 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). Moreover, 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. Equation (19) shows how to do so. With only one pre-treatment time period this test does not rely on linearity.

Is a DD design internally valid if $VWCT = 0$ but there is evidence that one or more groups have a differential trend? Consider an analogy to IV. The identifying assumption for a single Wald estimate comparing two values of a variable z_i is $E[\epsilon_i|z_i = a] - E[\epsilon_i|z_i = b] = 0$. When z_i takes many values it is unlikely that *every* possible Wald estimate satisfies this condition. A two-stage least squares estimator that uses z_i as an instrument, however, weights together all such Wald estimates and requires $E[z_i\epsilon_i] = 0$. This is the commonly stated identifying assumption for IV. Similarly, while each 2x2 DD requires pairwise common trends, the full DD estimator only requires $VWCT = 0$. Dropping a timing group with $\Delta Y_k^0 > 0$ but $w_k^T - w_k^C \approx 0$, for example, will "fix" the violation of equal trends without changing the estimate at all.

How should researchers test internal validity in a design with treatment timing? One approach is to test "pairwise" balance across groups. Existing approaches include estimating a linear relationships between a confounder, x_{it}, and t_k^* or comparing averages of x_{it} in "early" and "late" treated units (Almond, Hoynes, and Schanzenbach 2011, Bailey and Goodman-Bacon 2015).[20] The intuition is that if identification comes from comparing earlier- and later-treated units then covariates should be balanced between earlier- and later-treated units. Equation (19) shows that the effective control group can include *both* the earliest and latest treated units, so these tests could miss the relevant imbalance between the "most" and "least" treated units. Regressing x_{it} on a constant and dummies for timing groups tests the null of joint balance across groups (H_0: $\overline{x}_k - \overline{x}_U = 0,\ \forall k$), and plotting these means clarifies where any imbalance comes from. With many timing groups, though, this *F*-test will have low power and does not reflect how imbalance matters for bias in $\hat{\beta}^{DD}$: $(w_k^T - w_k^C)$.

[20] One can use time-varying confounders as outcomes (Freyaldenhoven, Hansen, and Shapiro 2018, Pei, Pischke, and Schwandt 2017), but this does not test for balance in levels, nor can it be used for sparsely measured confounders.

Equation (19) also suggests a single t-test of reweighted balance in x_{it}, a proxy for $VWCT = 0$:

1. Generate a dummy for the effective treatment group, $B_k = w_k^T - w_k^C > 0$.
2. Regress timing-group means, $\overline{x}_k$, on B_k weighting by $\left|w_k^T - w_k^C\right|$
3. The coefficient on B_k equals covariate differences weighted by the actual identifying variation, and its *t*-statistic tests the null of reweighted balance in (19).

One can use this strategy to test for pre-treatment trends in confounders (or the outcome) by regressing $\bar{x}_{kt}$ on B_k, year dummies, and their interaction, or the interaction of B_k with a linear trend using dates before any treatment starts. This procedure involves a single null hypothesis that maps directly to bias in the estimator, rather than a joint null without a clear relationship to bias.

III. 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. Figure 5 replicates their event-study result for female suicide using an unweighted specification with no covariates.[21] Our results match closely: suicide rates display no clear trend before the implementation of unilateral divorce laws, but begin falling soon

[21] Data on suicides by age, sex, state, and year come from the National Center for Health Statistics' Multiple Cause of Death files from 1964-1996, 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.

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.[22]

A. Describing the design

Figure 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.[23] 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 ($s_{k\ell}^{k}$ and $s_{k\ell}^{\ell}$) shows how much of $\hat{\beta}^{DD}$ comes from timing variation (37 percent). The large untreated group puts a lot of weight on $\hat{\beta}_{kU}^{2x2}$ terms, but more on those involving pre-1964 reform states (38.4 percent) than non-reform states (24 percent). Figure 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 figure 6. The average of the post-treatment event-study estimates in figure 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).[24] 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 -

[22] 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.

[23] 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 $(12^2 - 12)/2 = 66$ comparisons between an earlier switcher and a later non-switcher, and 66 comparisons between a later switcher and an earlier non-switcher

[24] 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 $\bar{D}_k = 1$ they can only act as an already-treated control group. If the effects for pre-1964 reform states had stabilized by 1969 they would not cause bias.

5.44—close to the average of the event-study coefficients. The DD decomposition theorem shows that one way to summarize effects in the presence of time-varying heterogeneity is simply to subtract the components of the DD estimate that are biased using the weights in equation (10a).

B. Testing the design

Figures 7 and 8 test for covariate balance in the unilateral divorce analysis. Figure 7 plots the balance test weights, $w_k^T - w_k^C$, from equation (19), the corresponding weights from a timing-only design, and each group's sample share. Because they have relatively low treatment variance, the earliest timing groups receive less weight than their sample shares imply.[25] In fact, the 1969 states effectively act as controls because $w_k^T - w_k^C < 0$.

Figure 8 implements both a joint balance test and the reweighted test using two potential determinants of marriage market equilibria in 1960: per-capita income and the male/female sex ratio. Panel A shows that average per-capita income in untreated states ($13,431) is lower than the average in every timing group except for those that implemented unilateral divorce in 1969 (which actually get more weight as controls) or 1985. The joint F-test, however, fails to reject the null hypothesis of equal means. It is not surprising that a test of 12 restrictions on 48 states observations fails to generate strong evidence against the null. The reweighed test, on the other hand, does detect a difference in per-capita income of $2,285 between effective treatment states—those that implemented unilateral divorce in 1970 or later—and effective control states—pre-1964 reform states, non-reform states, and the 1969 states. Panel B shows that the 1960 sex ratio is higher in almost all treatment states than in the control states. The joint test cannot reject the null of equal means, but the reweighted test does ($p = 0.06$).[26]

[25] Adding $5 \times year$ to the suicide rate for the 1970 states ($w_k^T - w_k^C = 0.0039$) changes the DD estimate from -3.08 to -2.75, but adding it to the 1973 group ($w_k^T - w_k^C = 0.18$) yields a very biased DD estimate of 12.28.

[26] One can run a joint test of balance across covariates using seemingly unrelated regressions (SUR), as suggested by Lee and Lemieux (2010) in the regression discontinuity context. The results of these χ^2 tests are displayed at the top of figure 8. As with the separate balance tests, I fail to reject the null of equal means across groups and covariates.

IV. 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. Any alternative specification that equals a weighted average can be written as a product a vector of 2x2 DDs and a vector of weights—$\hat{\beta}^{DD} = \boldsymbol{s}'\widehat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}$—so that the difference between two specifications has the form of a Oaxaca-Blinder-Kitagawa decomposition (Blinder 1973, Oaxaca 1973, Kitagawa 1955):

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

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 $\widehat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}_{\boldsymbol{alt}}$ against $\widehat{\boldsymbol{\beta}}^{\boldsymbol{2x2}}$ and $\boldsymbol{s}$ against $\boldsymbol{s}_{\boldsymbol{alt}}$.

A. Weighting

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

The joint reweighted balance test, however, does reject the null of equal weighted means between effective treatment and control groups. With 48 states and 12 timing groups, there are not sufficient degrees of freedom to implement a full joint test across many covariates. This is an additional rationale for the reweighted test.

[27] Grosz, Miller, and Shenhav (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

-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]

Figure 9 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]

B. 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, pp 770) observe that "an obvious extension is to include regressors" and Angrist and Pischke (2009, pp 236) highlight "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 . Most applications, however, include time-varying controls $\boldsymbol{X_{it}}$:

divorce estimate positive (2.42, s.e. = 1.81), but figure 6 suggests that this occurs not necessarily because of a problem with the design, but because *half* of the timing terms are biased by time-varying treatment effects.

[29] Solon, Haider, and Wooldridge (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 $\hat{\beta}_{kU}^{2x2}$ 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 $\hat{\beta}_{kU}^{2x2}$ much less negative, changing the overall DD estimate to 1.04. Callaway and Sant'Anna (2018) propose a generalized propensity score reweighted estimator to exploit timing variation.

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

But we have no theoretical guidance on how these controls adjust the 2x2 DDs, change the identifying variation, or when they eliminate violations of the identifying assumption.[32] This subsection derives a decomposition result for a general controlled DD specifications like (21).

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

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

The index of covariates, $\tilde{p}_{it} \equiv \hat{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{X}}_{\boldsymbol{it}}$ is predicted treatment status for unit i in period t based on the sample-wide relationship between $\widetilde{\boldsymbol{X}}_{\boldsymbol{it}}$ and $\widetilde{D}_{it}$ (see Sloczynski 2017). The covariate-adjusted treatment variable subtracts predicted treatment status from true treatment status: $\tilde{d}_{it} \equiv \left[(D_{it} - \overline{D}_i) - \left(\hat{\boldsymbol{\Gamma}}\overline{\boldsymbol{X}}_{\boldsymbol{it}} - \hat{\boldsymbol{\Gamma}}\overline{\boldsymbol{X}}_{\boldsymbol{i}}\right)\right] - \left[\left(\overline{D}_t - \overline{\overline{D}}\right) - \left(\hat{\boldsymbol{\Gamma}}\overline{\boldsymbol{X}}_{\boldsymbol{t}} - \hat{\boldsymbol{\Gamma}}\overline{\overline{\boldsymbol{X}}}\right)\right]$. The controlled DD coefficient comes from a regression of y_{it} on $\tilde{d}_{it}$:

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

Equation (23) shows that identification of $\hat{\beta}^{DD|X}$ comes from variation in $\widetilde{D}_{it}$ *and* $\tilde{p}_{it}$. $\widetilde{D}_{it}$ varies by timing group and time, but $\tilde{p}_{it}$ (generally) varies across units, even those in the *same* treatment timing group.

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 $\bar{d}_{kt} - \bar{d}_k = \left(\overline{D}_{kt} - \overline{D}_k\right) - \left(\hat{\boldsymbol{\Gamma}}\overline{\boldsymbol{X}}_{\boldsymbol{kt}} - \hat{\boldsymbol{\Gamma}}\overline{\boldsymbol{X}}_{\boldsymbol{k}}\right)$:

[32] de Chaisemartin and D'Haultfoeuille (2018) analyze a DD specification for the modified outcome variable $y_{it} - \boldsymbol{\Phi X_{it}}$, which has the same weights as the uncontrolled specification by definition, but does not account for the way that control variables change the identifying variation in $\widetilde{D}_{it}$.

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

Substituting (24) into (23) shows that controls both adjust the DD coefficient at the group-time level ($\tilde{d}_{kt}$), and introduce within-group comparisons ($\tilde{d}_{i(k)t}$):

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

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

The within coefficient, $\hat{\beta}_w^p \equiv \frac{\widehat{C}(y_{it}, \tilde{d}_{i(k)t})}{\widehat{V}_w^d}$, measures the relationship between y_{it} and changes over time in $\tilde{d}_{i(k)t}$ across units in the same timing group.[33] There is no variation in $\widetilde{D}_{it}$ within timing groups, though, so $\tilde{d}_{i(k)t}$ only varies because of predicted treatment status. $\hat{\beta}_w^p$ 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, $\hat{\beta}_b^d \equiv \frac{\widehat{C}(y_{it}, \tilde{d}_{kt})}{\widehat{V}_b^d}$, comes from timing-group-by-time-period variation, just as in Theorem 1. It contains the unadjusted DD coefficient $\hat{\beta}^{DD}$ and subtracts $\hat{\beta}_b^p$, the two-way fixed effects coefficient from a regression of y_{it} on $\widetilde{p}_{kt}$ (averages of the covariate index by group and time). Appendix C decomposes $\hat{\beta}_b^d$ into adjusted 2x2 DDs as in Theorem 1:

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

$$\hat{\beta}_b^d = \sum_k \sum_{\ell>k} \overbrace{(n_k + n_\ell)^2 \frac{\hat{V}_{b,k\ell}^d}{\hat{V}_b^d}}^{s_{k\ell}^{b|X}} \overbrace{\left[\frac{\hat{V}_{k\ell}^D \hat{\beta}_{k\ell}^{2x2} - \hat{V}_{b,k\ell}^p \hat{\beta}_{b,k\ell}^p}{\hat{V}_{b,k\ell}^d}\right]}^{\hat{\beta}_{b,k\ell}^d} . \quad (26)$$

The variances and coefficients in (26) parallel those in (25) but as the subscripts indicate they come from each two-group subsample.[34] Controls change the estimate for the two reasons highlighted in the Oaxaca-Blinder-Kitagawa expression: they change the weights because $\hat{V}_{b,k\ell}^d \neq \hat{V}_{kl}^D$ and they adjust each 2x2 estimate by the subsample relationship between y_{it} and $\tilde{p}_{kt}$: $\hat{V}_{b,k\ell}^p \hat{\beta}_{b,k\ell}^p$.[35]

Combining equations (25) and (26) gives the full decomposition for a controlled specification:

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

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

---

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

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

$$\hat{\beta}_{b,k\ell}^d = \frac{\hat{C}(y_{jt}, \widetilde{D}_{jt} - \tilde{p}_{jt}^{k\ell}) + \hat{C}(y_{jt}, \tilde{p}_{jt}^{k\ell} - \tilde{p}_{jt})}{\hat{V}_{b,k\ell}^d} \quad , j \in k, \ell$$

$$= \frac{(1 - R_{k\ell}^2)\hat{V}_{k\ell}^D \hat{\beta}_{k\ell}^{2x2|X} + \hat{V}_{b,k\ell}^{dp} \hat{\beta}_{b,k\ell}^{dp}}{(1 - R_{k\ell}^2)\hat{V}_{k\ell}^D + \hat{V}_{b,k\ell}^{dp}}$$

The superscript dp refers to the difference between subsample and full sample predicted treatment, $\tilde{p}_{jt}^{k\ell} - \tilde{p}_{jt}$. $\hat{V}_{b,k\ell}^{dp}$ is its variance and $\hat{\beta}_{b,k\ell}^{dp}$ is the regression coefficient relating it to y_{jt}. $R_{k\ell}^2$ comes from the subsample Frisch-Waugh regression. The smaller is $R_{k\ell}^2$ the more weight is put on the subsample controlled term and the closer it is to the unadjusted 2x2 DD. When $R_{k\ell}^2 = 1$, then $\tilde{p}_{jt}^{k\ell} = \widetilde{D}_{jt}$ and the estimate collapses back to $\hat{\beta}_{b,k\ell}^d$ 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 $\hat{V}_{b,k\ell}^{dp} \approx 0$ and $\hat{\beta}_{b,k\ell}^{2x2|d} \approx \hat{\beta}_{k\ell}^{2x2|X}$. Adding covariates extrapolates the full-sample Frisch-Waugh relationship to the pairs and therefore depends strongly on correctly specifying model (21).

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. Figure 10 illustrates the within variation for the two 1970 no-fault divorce states, California and Iowa. The two states have the same values of $\widetilde{D}_{it}$ 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, $\tilde{d}_{CA,t} - \tilde{d}_{IA,t} = \left(\widetilde{D}_{CA,t} - \tilde{p}_{CA,t}\right) - \left(\widetilde{D}_{IA,t} - \tilde{p}_{IA,t}\right) = \tilde{p}_{IA,t} - \tilde{p}_{CA,t}$, 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.

Figure 11 illustrates the controlled between term for the 1970 states compared to non-reform states, $\hat{\beta}_{1970,NRS}^{2x2|d}$. Panel A plots the treatment variable and the group-year means of predicted treatment status from the full-sample Frisch-Waugh 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.[36]

[36] Appendix C 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. When treatment effects are

Appendix D 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, Meer and West 2013, Neumark, Salas, and Wascher 2014). 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.[37]

V. CONCLUSION

Difference-in-differences is perhaps the most widely applicable quasi-experimental research design. Its transparency makes it simple to describe, test, interpret, and teach. This paper extends all of these advantages from canonical 2x2 DD estimator to general and much more common DD estimators with variation in the timing of treatment.

---

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, which can suffer from misspecification. 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.

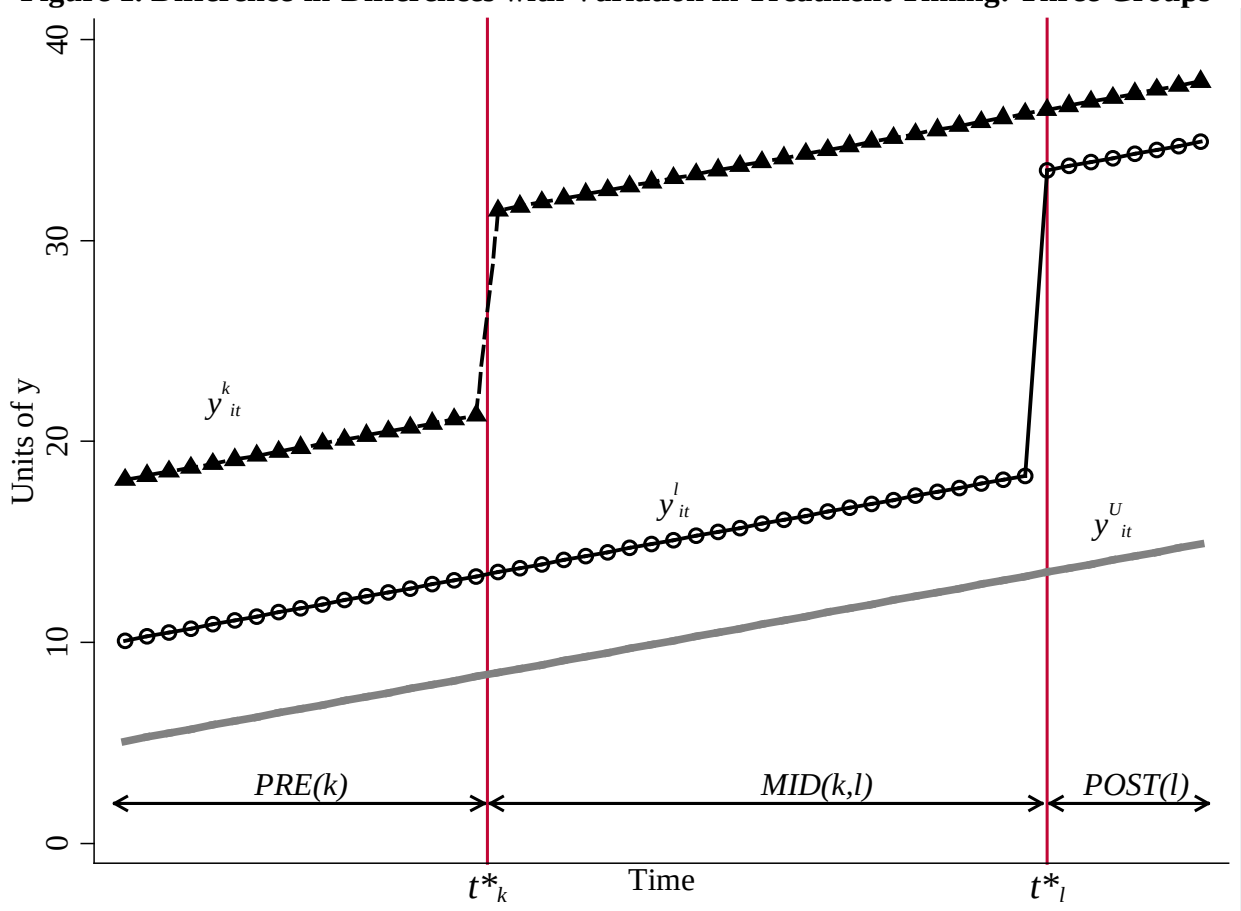
[37] Appendix D also analyzes triple-difference models and shows that they also have a weighted average form. Appendix E briefly discusses treatment variables that turn off.

The two-way fixed effects DD coefficient 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. This defines the DD estimand, the variance-weighted average treatment effect on the treated (VWATT), and generalizes the identifying assumption on counterfactual outcomes to variance-weighted common trends (VWCT). Moreover, because already-treated units act as controls in some 2x2 DD's, identification of VWATT requires an additional identifying assumption of time-invariant treatment effects.

The DD decomposition theorem also leads to several new tools for practitioners. Graphing the 2x2 DDs against their weight displays all the identifying variation in any DD application, and summing weights across types of comparisons quantifies "how much" of a given estimate comes from different sources of variation. I use the DD decomposition theorem to form a reweighted balance test that reflects this identifying variation, is easy to implement, tests fewer restrictions than joint balance tests, and shows how large and in what direction any imbalance occurs.

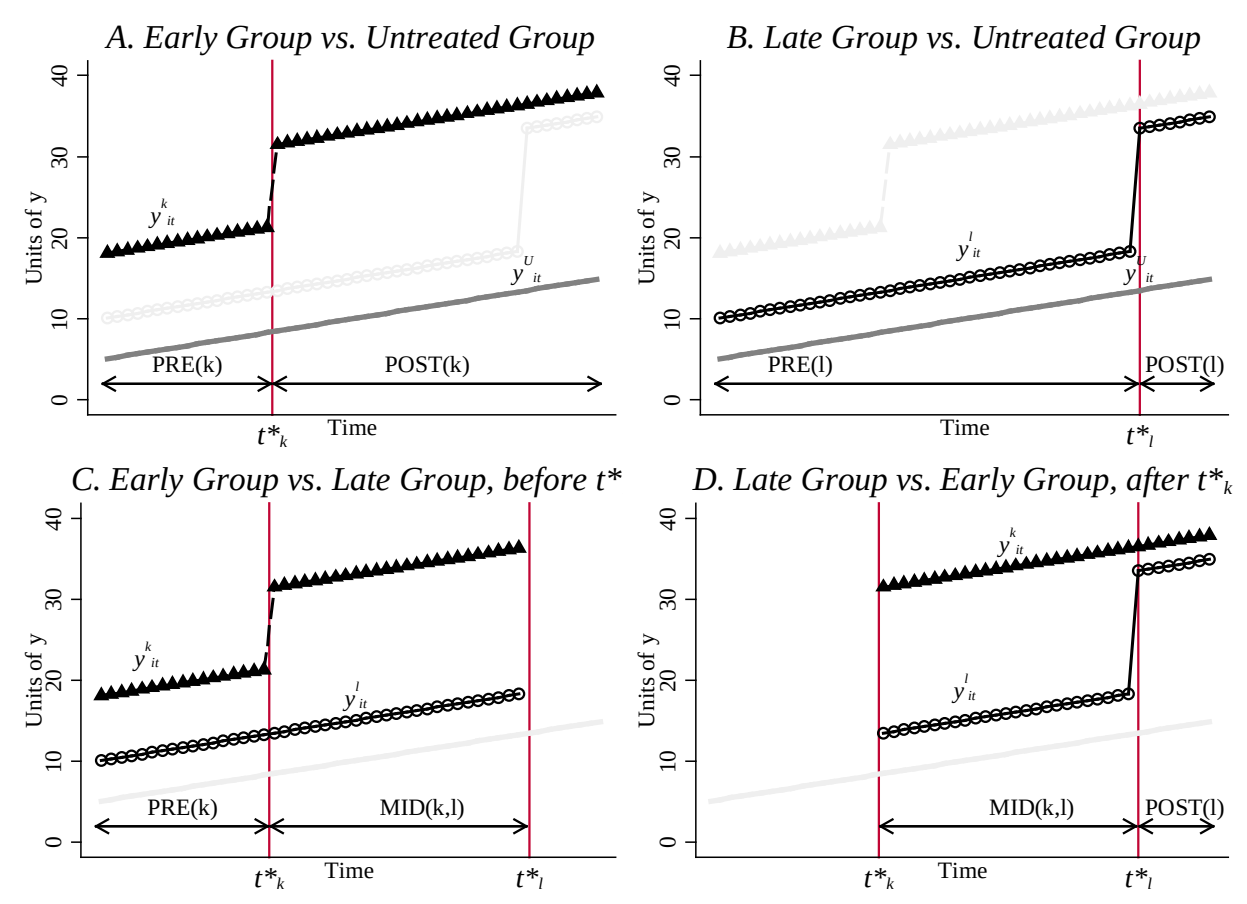
I suggest several simple methods to learn why estimates differ across alternative specifications. The weighted average representation leads to a Oaxaca-Blinder-Kitagawa-style decomposition that quantifies how much of the difference in estimates comes from changes in the 2x2 DD's, the weights, or both. Plots of the components or the weights across specifications show clearly where differences come from and can help researchers understand why their estimates changes and whether or not it is a problem. I also provide a new analysis of DD models that include covariates and quantify the extent to which identification is driven purely by variation in the controls, which is shown to matter in practice.

Figure 1. Difference-in-Differences with Variation in Treatment Timing: Three Groups



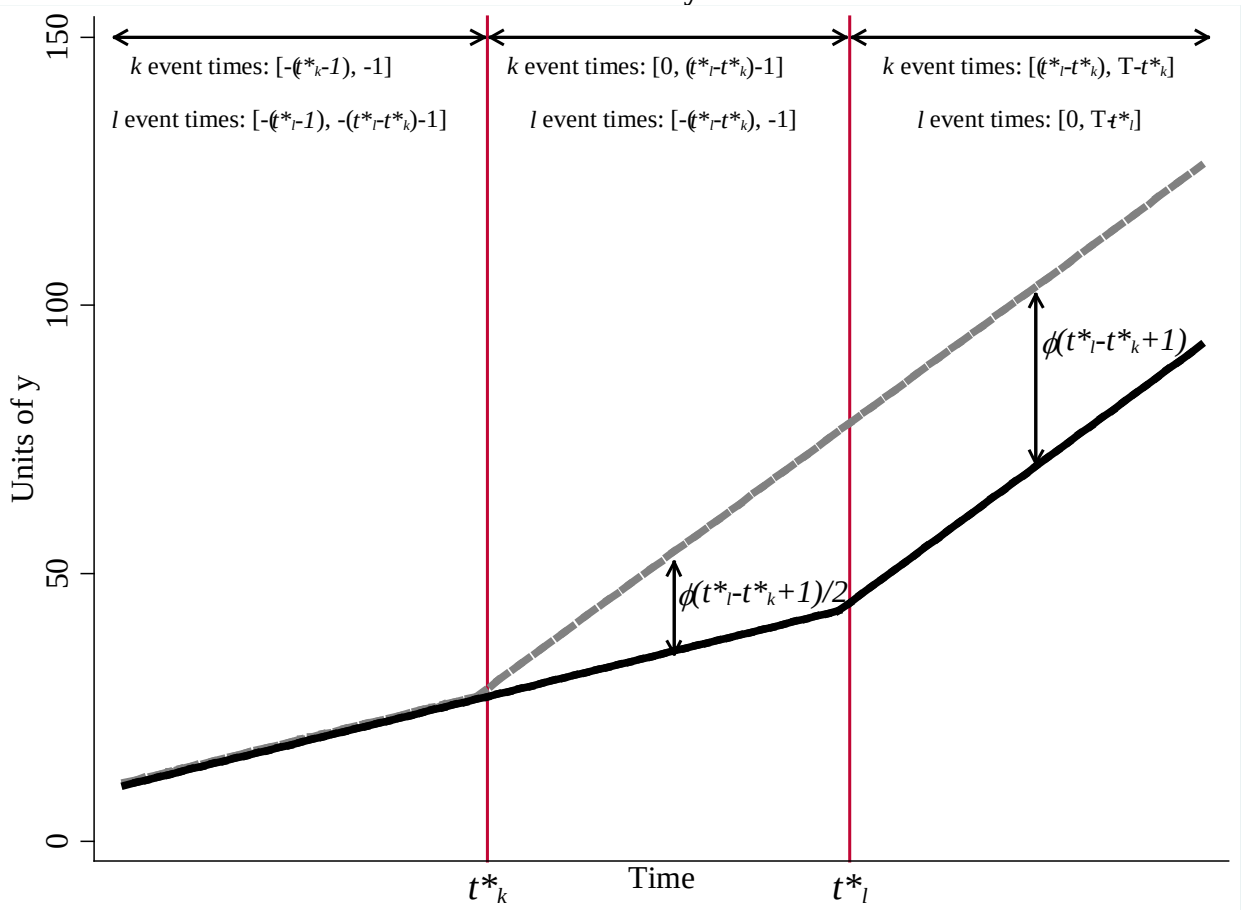
Notes: The figure plots outcomes in three groups: a control group, U, which is never treated; an early treatment group, E, which receives a binary treatment at $t_k^* = \frac{34}{100}T$; and a late treatment group, ℓ, which receives the binary treatment at $t_\ell^* = \frac{85}{100}T$. The x-axis notes the three sub-periods: the pre-period for group k, $[1, t_k^* - 1]$, denoted by $PRE(k)$; the middle period when group k is treated and group ℓ is not, $[t_k^*, t_\ell^* - 1]$, denoted by $MID(k, \ell)$; and the post-period for group ℓ, $[t_\ell^*, T]$, denoted by $POST(\ell)$. I set the treatment effect to 10 in group k and 15 in group ℓ.

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



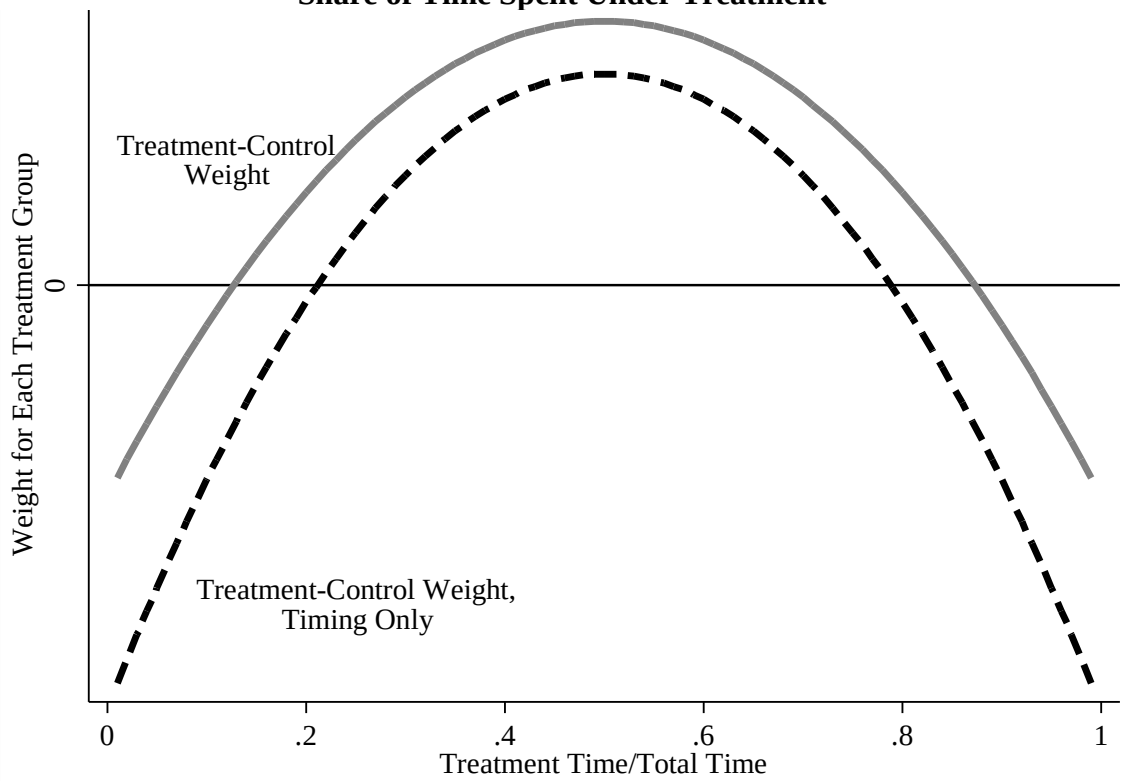
Notes: The figure plots the groups and time periods that generate the four simple 2x2 difference-in-difference estimates in the case with an early treatment group, a late treatment group, and an untreated group from Figure 1. Each panel plots the data structure for one 2x2 DD. Panel A compares early treated units to untreated units ($\widehat{\beta}_{kU}^{DD}$); panel B compares late treated units to untreated units ($\widehat{\beta}_{\ell U}^{DD}$); panel C compares early treated units to late treated units during the late group's pre-period ($\widehat{\beta}_{k\ell}^{DD,k}$); panel D compares late treated units to early treated units during the early group's post-period ($\widehat{\beta}_{k\ell}^{DD,\ell}$). 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.

Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



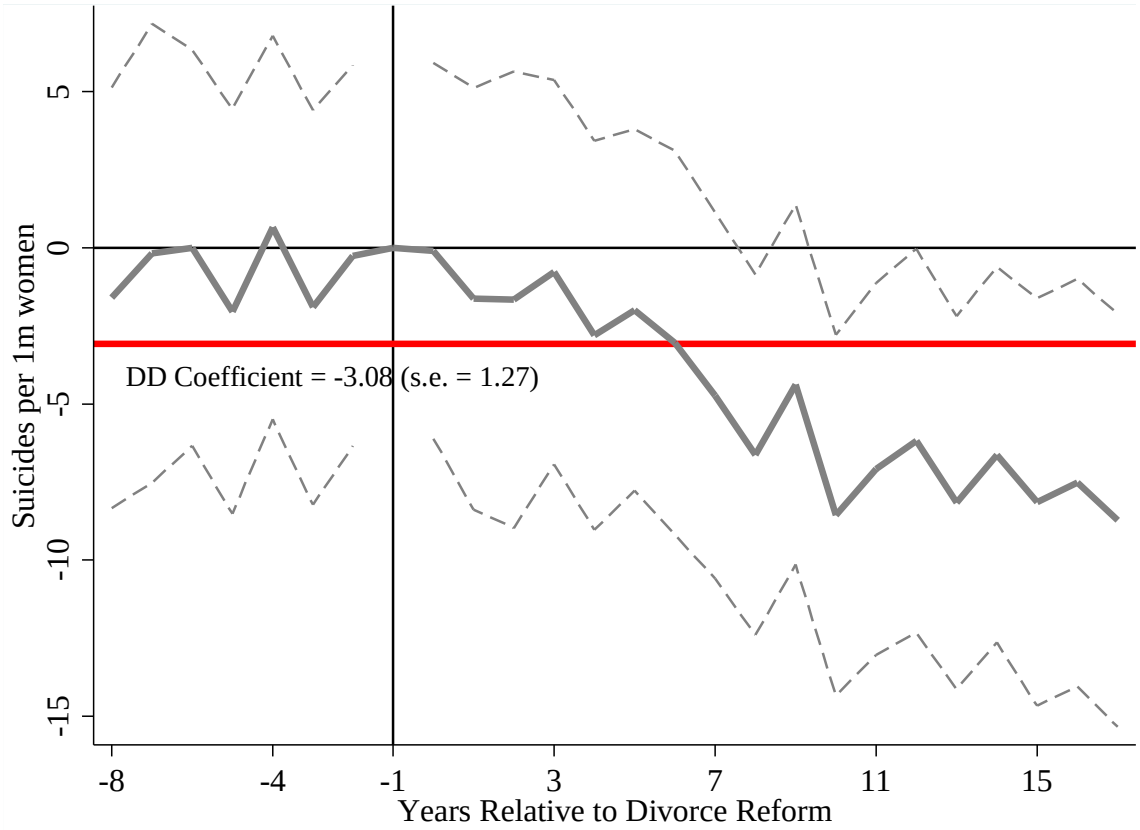
Notes: The figure plots a stylized example of a timing-only DD set up with a treatment effect that is a trend-break rather than a level shift (see Meer and West 2013). Following section II.A.ii, the trend-break effect equals $\phi \cdot (t - t^* + 1)$. The top of the figure notes which event-times lie in the $PRE(k)$, $MID(k, \ell)$, and $POST(\ell)$ periods for each unit. The figure also notes the average difference between groups in each of these periods. In the $MID(k, \ell)$ period, outcomes differ by $\frac{\phi}{2}(t_\ell^* - t_k^* + 1)$ on average. In the $POST(\ell)$ period, however, outcomes had already been growing in the early group for $t_\ell^* - t_k^*$ periods, and so they differ by $\phi(t_\ell^* - t_k^* + 1)$ on average. The 2x2 DD that compares the later group to the earlier group is biased and, in the linear trend-break case, weakly negative despite a positive and growing treatment effect.

Figure 4. Weighted Common Trends: The Treatment/Control Weights as a Function of the Share of Time Spent Under Treatment



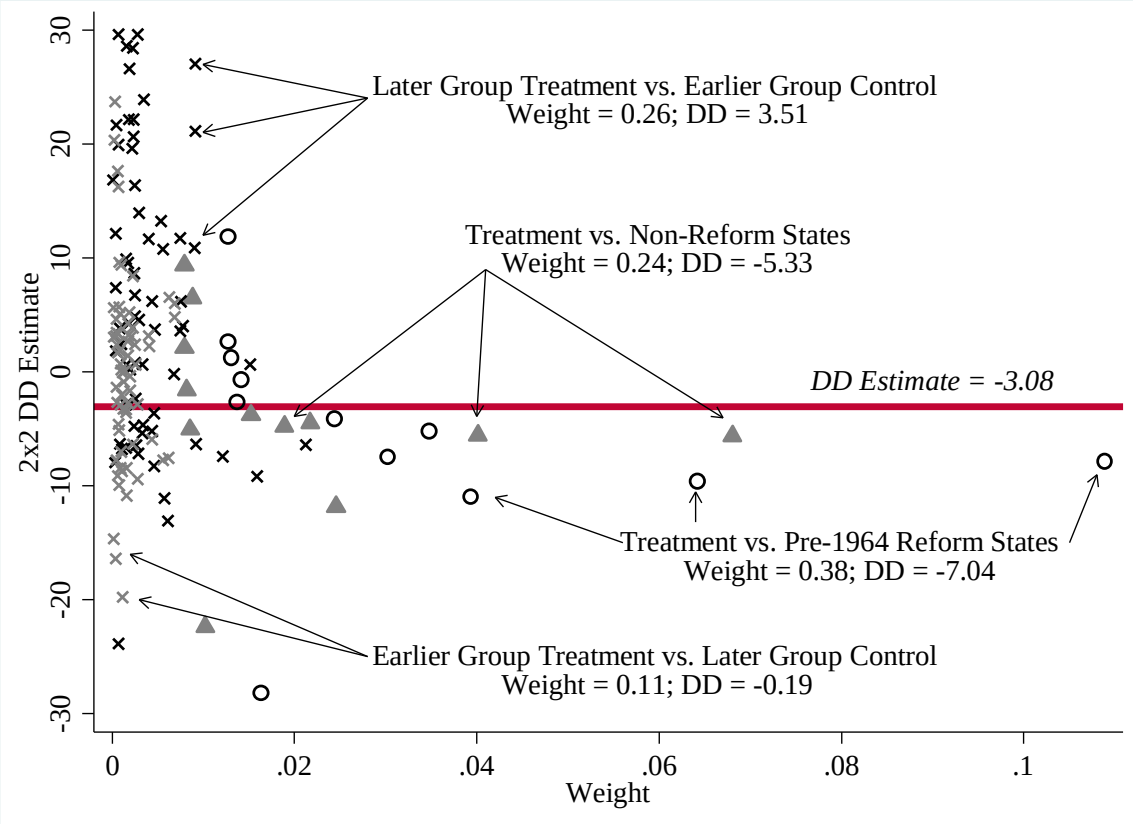
Notes: The figure plots the weights that determine each timing group's importance in the weighted common trends expression in equations (16) and (17).

Figure 5. Event-Study and Difference-in-Differences Estimates of the Effect of No-Fault Divorce on Female Suicide: Replication of Stevenson and Wolfers (2006)



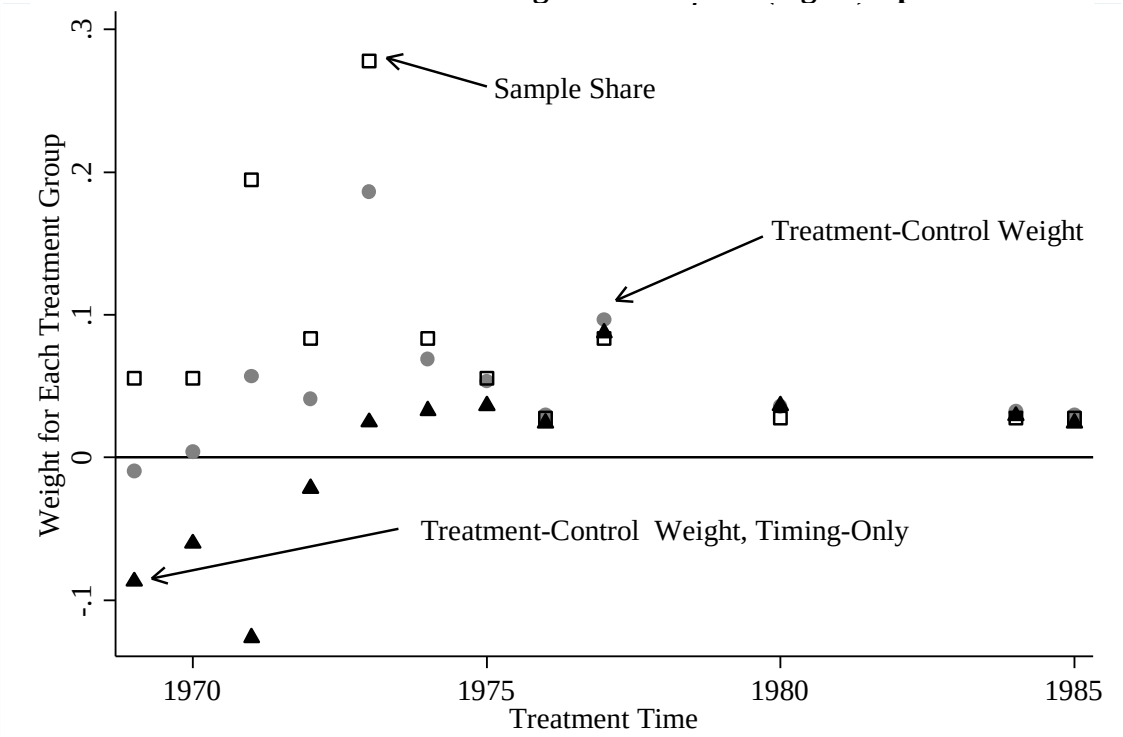
Notes: The figure plots event-study estimates from the two-way fixed effects regression equation on page 276 and plotted in figure 1 of Stevenson and Wolfers (2006), along with the DD coefficient. The specification does not include other controls and does not weight by population. Standard errors are robust to heteroskedasticity.

Figure 6. Difference-in-Differences Decomposition for Unilateral Divorce and Female Suicide



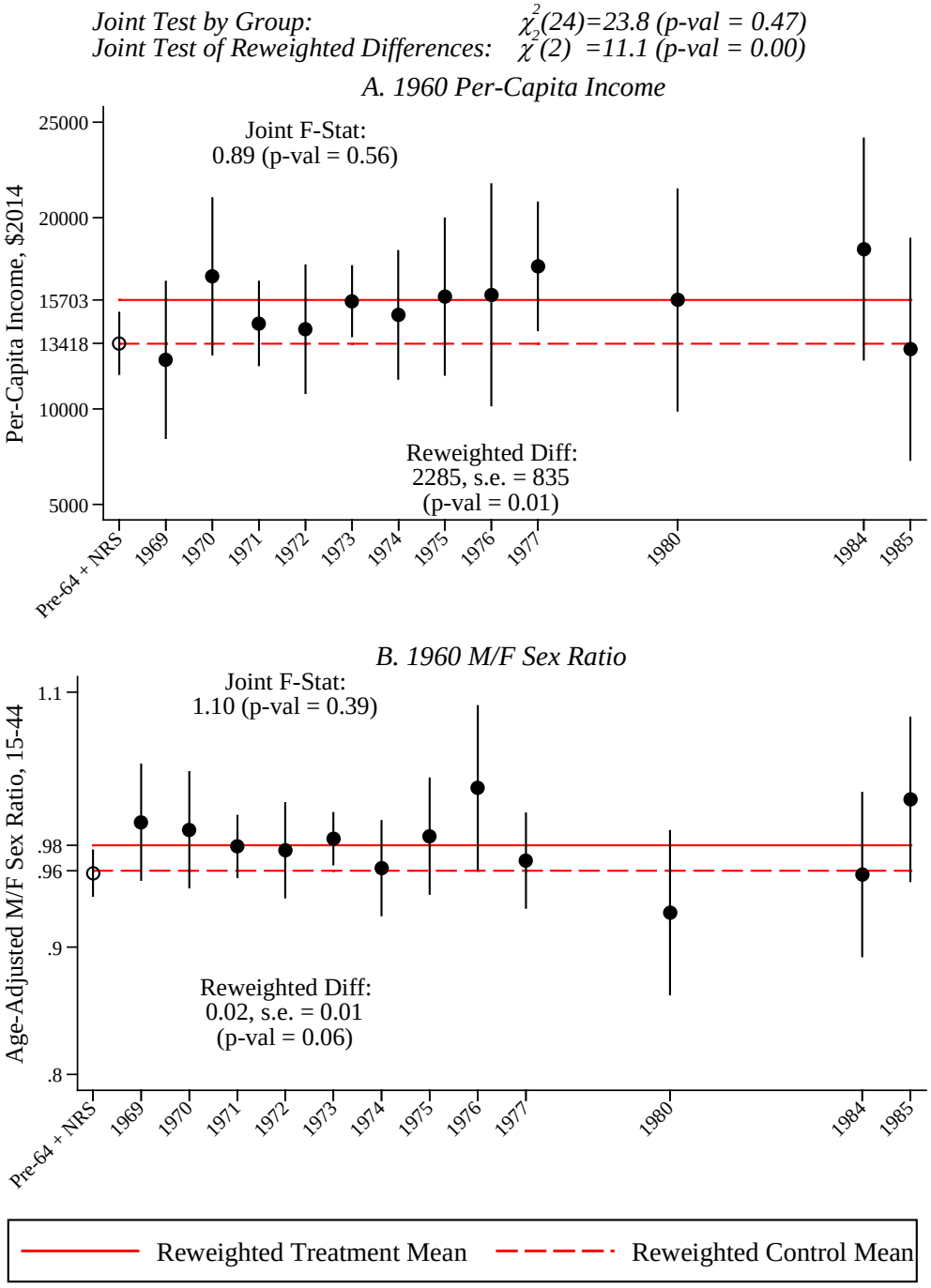
Notes: The figure plots each 2x2 DD components from the decomposition theorem against their weight for the unilateral divorce analysis. The open circles are terms in which one timing group acts as the treatment group and the pre-1964 reform states act as the control group. The closed triangles are terms 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.

Figure 7. Weighted Common Trends in the Unilateral Divorce Analysis: The Treatment/Control Weights on Each Timing Group



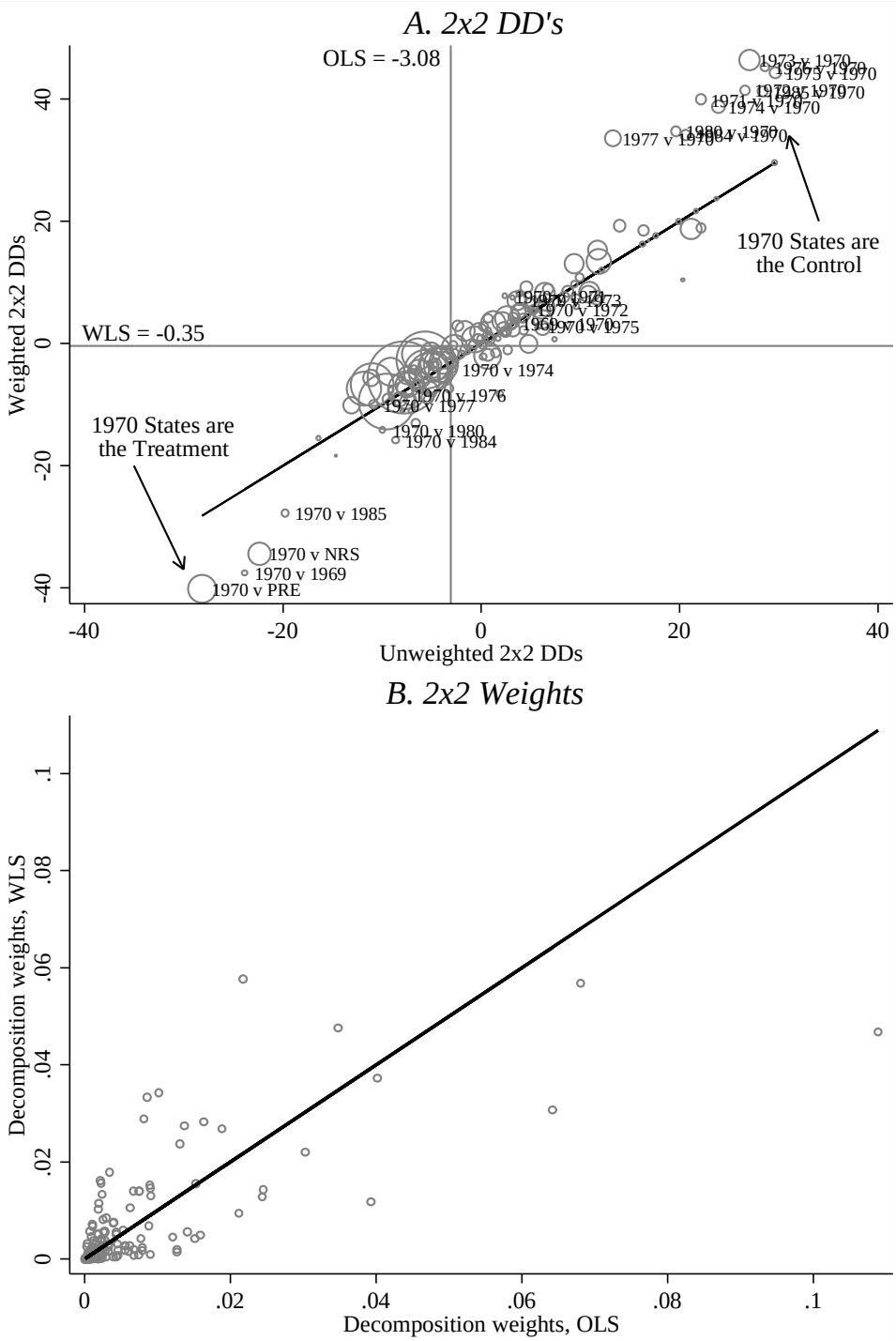
Notes: The figure plots the weights that determine each timing group's role in the weighted common trends expression. These are show in solid triangles and equal the difference between the total weight each treatment timing group receives in terms where it is the treatment group (w_k^T) and terms where 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 group's sample share.

Figure 8. Testing for Balance in a Difference-in-Differences Estimator with Timing: Reweighted Test versus Joint Test



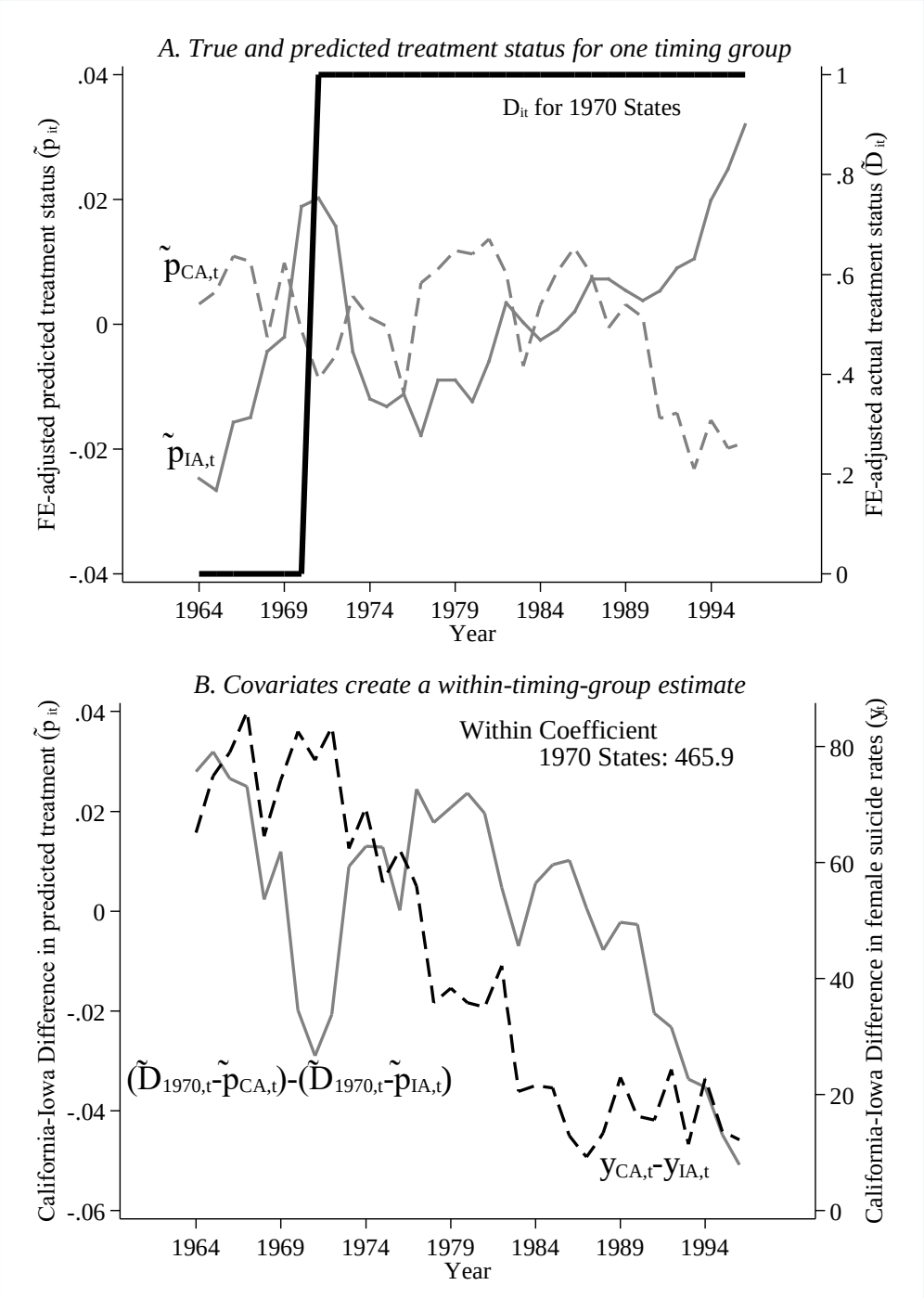
Notes: The figure plots average per-capita income and male/female sex ratio in 1960 for each timing group in the unilateral divorce analysis (combining the non-reform and pre-1964 reform states into the group labeled "Pre-64 + NRS"). The horizontal lines equal the average of these variables using the weights from figure 6 ($w_k^T - w_k^C$). Note that the 1969 states get more weight as a control group, so they are part of the reweighted control mean. Each panel reports the *F*-statistic and *p*-value from a joint test of equality across the means, and the reweighted difference, standard error, and *p*-value from the re-weighted balance test. The top of the figure reports $\chi^2(df)$ test-statistics for both covariates estimated using seemingly unrelated regressions.

Figure 9. Comparison of 2x2 DD Components and Decomposition Weights with and without Population Weights



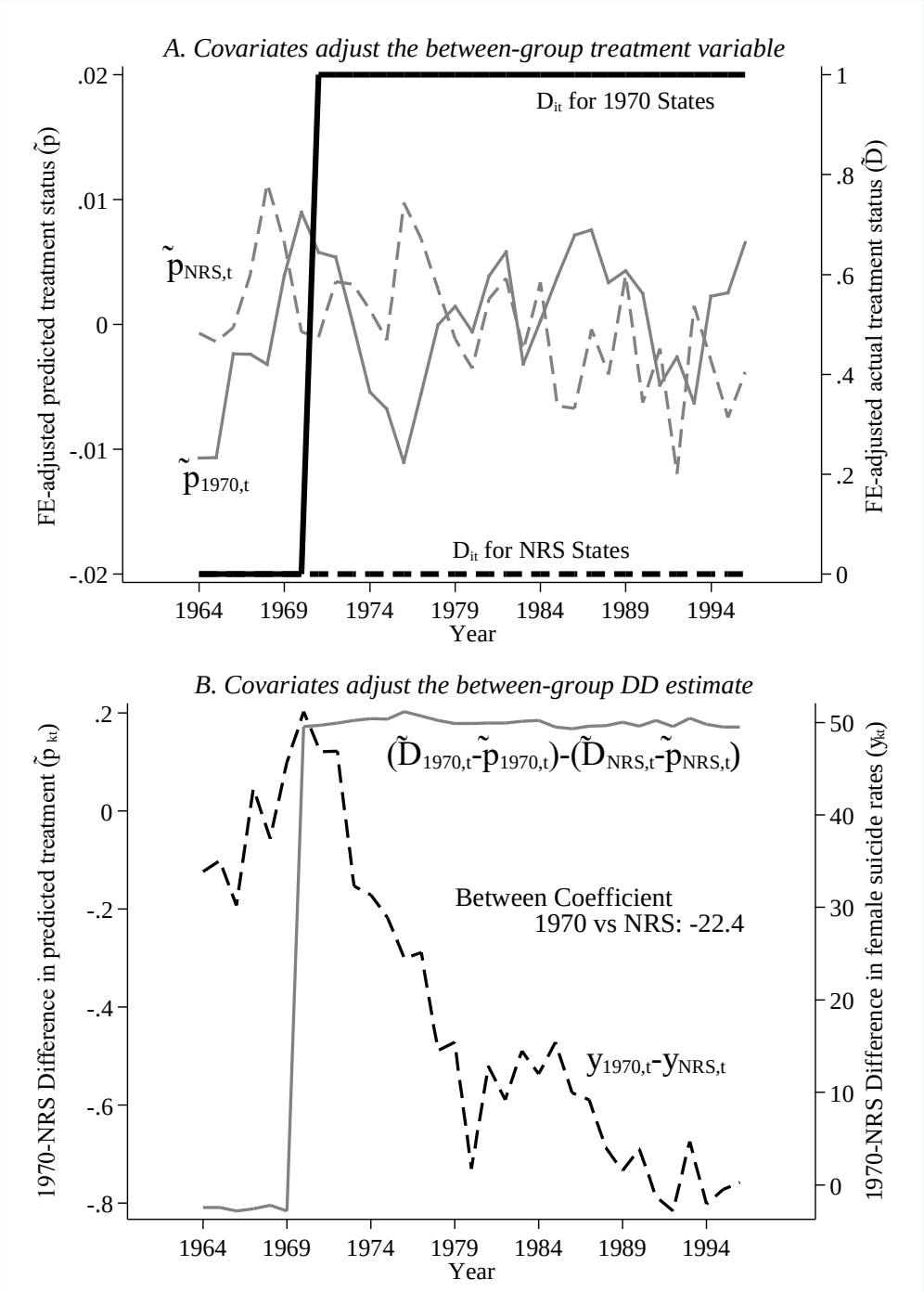
Notes: Panel A plots the 2x2 DD components from two-way fixed effects estimates that use population weights (*y*-axis) and do not (*x*-axis). The size of each point is proportional to its weight in an OLS version of equation (7). WLS estimates are much smaller than OLS estimates, and this figure shows 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.

Figure 10. Adding Controls Creates Within-Timing-Group Comparisons: An Example with the 1970 No-Fault Divorce States



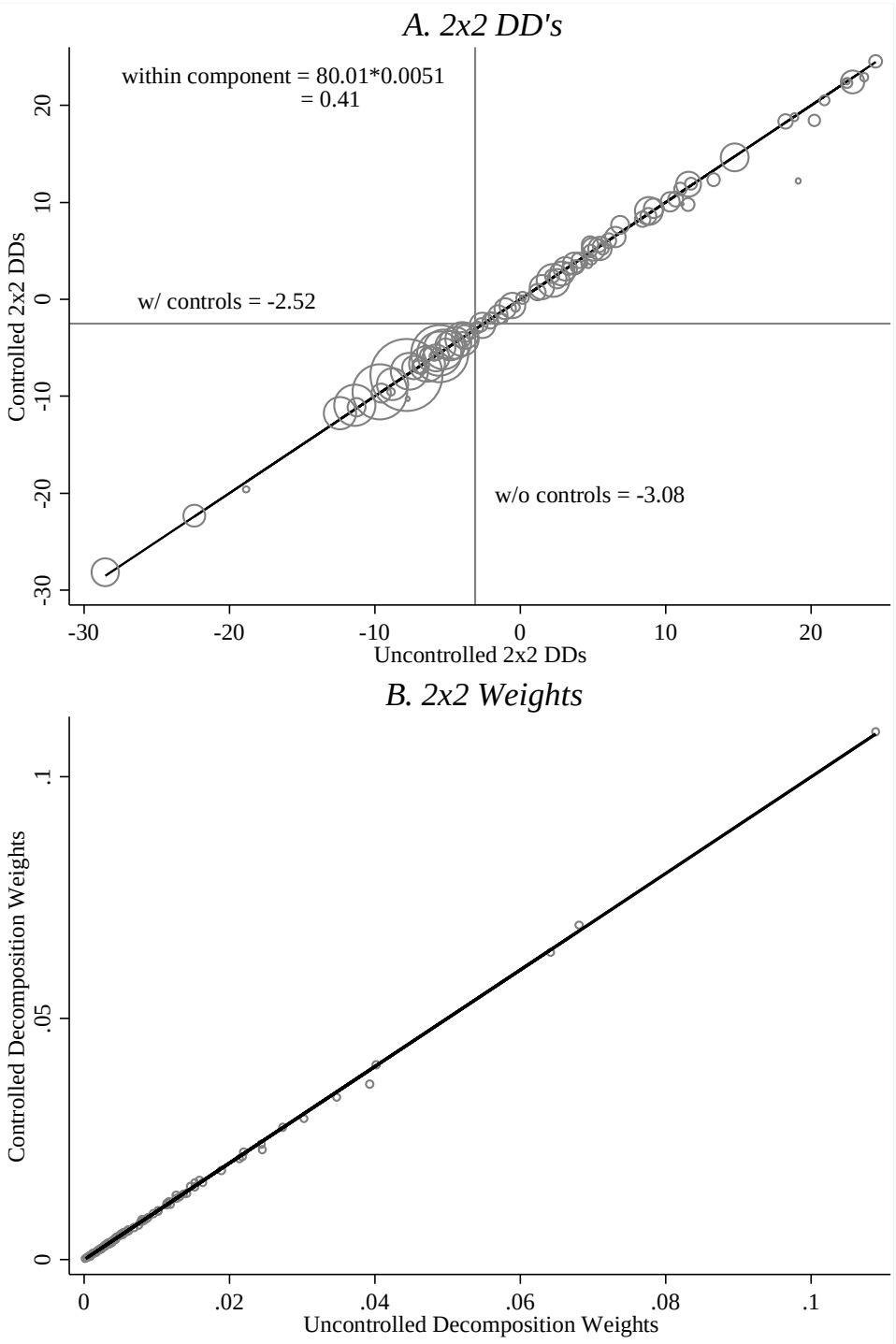
Notes: Panel A plots the treatment dummy and fitted values from the full-sample Frisch-Waugh regression (predicted treatment status, $\tilde{p}_{it}$) for the 1970 states, California and Iowa. Panel B plots the difference in adjusted treatment variable, $\tilde{D}_{it} - \tilde{p}_{it}$, between California 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.

Figure 11. 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 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 coefficient is -22.4. This is the part of the between term in (22) that comes from the 1970 versus non-reform comparison.

Figure 12. Comparison of 2x2 DD Components and Decomposition Weights with and without Controls



Notes: Panel A plots the two-group DDs from a regression that controls for per-capita income, female homicide rates, and welfare participation rates (*y*-axis) against those from the uncontrolled specification (*x*-axis). The size of each point is proportional to its weight in the controlled regression. Controls only change the estimate slightly, but almost all of this comes from the within-term: the comparisons between states in the same timing group (treatment status) but different predicted treatment status. Panel B is the same except that it plots the weights.

Table 1. The No-Fault Divorce Rollout: Treatment Times, Group Sizes, and Treatment Shares

No-Fault Divorce Year (t_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 1964-1996.

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 equation (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 1964-1968, subtracts them from the suicide rate, and estimates equation (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, an 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 equation (18).

VI. REFERENCES

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *The Review of Economic Studies* 72 (1):1-19.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490):493-505. doi: 10.1198/jasa.2009.ap08746.

Abraham, Sarah, and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Working Paper*.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3):1117-1165. doi: 10.1093/qje/qjv015.

Almond, Douglas, Hilary W. Hoynes, and Diane Whitmore Schanzenbach. 2011. "Inside the War On Poverty: The Impact of Food Stamps on Birth Outcomes." *The Review of Economics and Statistics* 93 (2):387-403. doi: 10.2307/23015943.

Angrist, Joshua D. 1991. "Grouped-data estimation and testing in simple labor-supply models." *Journal of Econometrics* 47 (2):243-266. doi: *https://doi.org/10.1016/0304-4076(91)90101-I*.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Chapter 23 - Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, 1277-1366. Elsevier.

Angrist, Joshua David, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics : an empiricist's companion*. Princeton: Princeton University Press.

Angrist, Joshua David, and Jörn-Steffen Pischke. 2015. *Mastering 'metrics : the path from cause to effect*. Princeton ; Oxford: Princeton University Press.

Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74 (2):431-497.

Athey, Susan, and Guido W. Imbens. 2018. "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." *Working Paper*.

Bailey, Martha J., and Andrew Goodman-Bacon. 2015. "The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans." *American Economic Review* 105 (3):1067-1104.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2019. "Synthetic Controls and Weighted Event Studies with Staggered Adoption." *Working Paper*.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1):249-275.

Bilinski, Alyssa, and Laura Hatfield. 2019. "Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions." *Working Paper*.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2003. "Some Evidence on Race, Welfare Reform, and Household Income." *The American Economic Review* 93 (2):293-298. doi: 10.2307/3132242.

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *The Journal of Human Resources* 8 (4):436-455. doi: 10.2307/144855.

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

Callaway, Brantly, Tong Li, and Tatsushi Oka. forthcoming. "Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with only Two Time Periods." *Journal of Econometrics*.

Callaway, Brantly, and Pedro Sant'Anna. 2018. "Difference-in-Differences With Multiple Time Periods and an Application on the Minimum Wage and Employment." *Working Paper*.

Cameron, Adrian Colin, and P. K. Trivedi. 2005. *Microeconometrics : methods and applications*. Cambridge ; New York: Cambridge University Press.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. 2013. "Average and Quantile Effects in Nonseparable Panel Models." *Econometrica* 81 (2):535-580. doi: 10.3982/ECTA8405.

Chyn, Eric. forthcoming. "Moved to Opportunity: The Long-Run Effect of Public Housing Demolition on Labor Market Outcomes of Children." *American Economic Review*.

de Chaisemartin, C., and X. D'HaultfŒuille. 2018. "Fuzzy Differences-in-Differences." *The Review of Economic Studies* 85 (2):999-1028. doi: 10.1093/restud/rdx049.

de Chaisemartin, C.;, and X. D'HaultfŒuille. forthcoming. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*.

de Chaisemartin, Clement, and Xavier D'Haultfoeuille. 2018. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *Working Paper*.
Deaton, Angus. 1997. *The Analysis of Household Surveys : a Microeconometric Approach to Development Policy*. Baltimore, MD: Johns Hopkins University Press.
Deshpande, Manasi, and Yue Li. 2017. "Who Is Screened Out? Application Costs and the Targeting of Disability Programs." *National Bureau of Economic Research Working Paper Series* No. 23472. doi: 10.3386/w23472.
Fadlon, Itzik, and Torben Heien Nielsen. 2015. "Family Labor Supply Responses to Severe Health Shocks." *National Bureau of Economic Research Working Paper Series* No. 21352. doi: 10.3386/w21352.
Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2018. "Pre-event Trends in the Panel Event-study Design." *National Bureau of Economic Research Working Paper Series* No. 24565. doi: 10.3386/w24565.
Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1 (4):387-401. doi: 10.2307/1907330.
Gibbons, Charles, E., Juan Carlos Suárez Serrato, and Michael Urbancic, B. 2018. Broken or Fixed Effects? In *Journal of Econometric Methods*.
Goodman-Bacon, Andrew, Thomas Goldring, and Austin Nichols. 2019. "bacondecomp: Stata module for Decomposing difference-in-differences estimation with variation in treatment timing." *Stata Command*.
Goodman, Joshua. 2017. "The Labor of Division: Returns to Compulsory High School Math Coursework." *National Bureau of Economic Research Working Paper Series* No. 23063. doi: 10.3386/w23063.
Grosz, Michael, Douglas L. Miller, and Na'ama Shenhav. 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., and ICPSR. 2010. Historical, Demographic, Economic, and Social Data: The United States, 1790-2002. ICPSR [distributor].
Heckman, James J., Robert J. Lalonde, and Jeffrey A. Smith. 1999. "Chapter 31 - The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, 1865-2097. Elsevier.
Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396):945-960. doi: 10.2307/2289064.
Imai, Kosuke, In Song Kim, and Erik Wang. 2018. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." *Working Paper*.
Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2):467-475. doi: 10.2307/2951620.
Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1993. "Earnings Losses of Displaced Workers." *The American Economic Review* 83 (4):685-709. doi: 10.2307/2117574.
Joseph Hotz, V., Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125 (1):241-270. doi: *https://doi.org/10.1016/j.jeconom.2004.04.009*.
Kasy, Maximilian. 2018. "Optimal taxation and insurance using machine learning — Sufficient statistics and beyond." *Journal of Public Economics* 167:205-219. doi: *https://doi.org/10.1016/j.jpubeco.2018.09.002*.
Kitagawa, Evelyn M. 1955. "Components of a Difference Between Two Rates." *Journal of the American Statistical Association* 50 (272):1168-1194. doi: 10.2307/2281213.
Krolikowski, Pawel. 2017. "Choosing a Control Group for Displaced Workers." *ILR Review*:0019793917743707. doi: 10.1177/0019793917743707.
Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2):281-355.
Lee, Jin Young, and Gary Solon. 2011. "The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates." *National Bureau of Economic Research Working Paper Series* No. 16773.
Malkova, Olga. 2017. "Can Maternity Benefits Have Long-Term Effects on Childbearing? Evidence From Soviet Russia." *The Review of Economics and Statistics*. doi: 10.1162/REST_a_00713.
Meer, Jonathan, and Jeremy West. 2013. "Effects of the Minimum Wage on Employment Dynamics." *National Bureau of Economic Research Working Paper Series* No. 19262. doi: 10.3386/w19262.
Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business & Economic Statistics* 13 (2):151-161. doi: 10.2307/1392369.
Neumark, David, J. M. Ian Salas, and William Wascher. 2014. "Revisiting the Minimum Wage—Employment Debate: Throwing Out the Baby with the Bathwater?" *ILR Review* 67 (3_suppl):608-648. doi: 10.1177/00197939140670S307.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14 (3):693-709. doi: 10.2307/2525981.
Oster, Emily. 2016. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*:1-18. doi: 10.1080/07350015.2016.1227711.
Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt. 2017. "Poorly Measured Confounders are More Useful on the Left Than on the Right." *National Bureau of Economic Research Working Paper Series* No. 23232. doi: 10.3386/w23232.
Perron, Pierre. 2006. "Dealing with Structural Breaks." *Working Paper*.
Rambachan, Ashesh, and Jonathan Roth. 2019. "An Honest Approach to Parallel Trends." *Working Paper*.
Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5):688-701. doi: 10.1037/h0037350.
Sant'Anna, Pedro, and Jun Zhao. 2018. "Doubly Robust Difference-in-Differences Estimators." *Working Paper*.
Shore-Sheppard, Lara D. . 2009. "Stemming the Tide? The Effect of Expanding Medicaid Eligibility On Health Insurance Coverage." *The B.E. Journal of Economic Analysis & 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, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2):301-316.
Stevenson, Betsey, and Justin Wolfers. 2006. "Bargaining in the Shadow of the Law: Divorce Laws and Family Distress." *The Quarterly Journal of Economics* 121 (1):267-288.
Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." *Working Paper*.
Surveillance, Epidemiology, and End Results (SEER). 2013. Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969-2011). edited by DCCPS National Cancer Institute, Surveillance Research Program, Surveillance Systems Branch.
Walters, Christopher R. forthcoming. "The Demand for Effective Charter Schools." *Journal of Political Economy*.
Wolfers, Justin. 2006. "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results." *American Economic Review* 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." *The Review of Economics and Statistics* 87 (2):385-390.
Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge, Mass.: MIT Press.