APPENDIX A: PROOF OF THE DD DECOMPOSITION THEOREM

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

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

$$\sum_k n_k \frac{1}{T} \sum_t z_{kt} \left[(x_{kt} - \bar{x}_k) - (\bar{x}_t - \bar{\bar{x}}) \right]$$

$$= \sum_k \sum_{\ell > k} n_\ell n_k \frac{1}{T} \sum_t (z_{kt} - z_{\ell t}) \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \qquad (A1)$$

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

$$(x_{kt} - \bar{x}_k) - (\bar{x}_t - \bar{\bar{x}}) = (x_{kt} - \bar{x}_k) - \sum_\ell n_\ell \, (x_{\ell t} - \bar{x}_\ell) \qquad (A2)$$

$$= \overbrace{(1 - n_k)}^{\sum_{\ell \neq k} n_\ell} (x_{kt} - \bar{x}_k) - \sum_{\ell \neq k} n_\ell \, (x_{\ell t} - \bar{x}_\ell) \qquad (A3)$$

$$= \sum_{\ell \neq k} n_\ell \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \qquad (A4)$$

Substitute in to (A1):

$$\frac{1}{T} \sum_t \sum_k \sum_{\ell \neq k} n_\ell \, n_k z_{kt} \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \qquad (A5)$$

$$= \sum_k \sum_{\ell > k} n_\ell n_k \frac{1}{T} \sum_t (z_{kt} - z_{kt}) \left[(x_{kt} - \bar{x}_k) - (x_{\ell t} - \bar{x}_\ell) \right] \qquad (A6)$$

Where (A6) follows because every dyad (a, b) appears twice, once with $z_{at}[(x_{at} - \bar{x}_a) - (\bar{x}_{bt} - \bar{x}_b)]$ and once with $z_{bt}[(x_{bt} - \bar{x}_b) - (x_{at} - \bar{x}_a)]$. ■

Proof of Theorem 1:

From equation (6) and the definition of $\widetilde{D}_{it}$:

$$\frac{\widehat{C}\,(y_{it},\widetilde{D}_{it})}{\widehat{V}\,(\widetilde{D}_{it})}=\frac{\frac{1}{NT}\sum_i\sum_t y_{it}\left[(D_{it}-\overline{D}_i)-\left(\overline{D}_t-\overline{\overline{D}}\right)\right]}{\widehat{V}^D} \tag{A7}$$

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

$$=\frac{\frac{1}{NT}\sum_i\sum_t y_{it}\left[\overbrace{(D_{it}-\overline{D}_i)-\left(\overline{D}_{k(i)t}-\overline{D}_{k(i)}\right)}^{=0}+\left(\overline{D}_{k(i)t}-\overline{D}_{k(i)}\right)-\left(\overline{D}_t-\overline{\overline{D}}\right)\right]}{\widehat{V}^D} \tag{A8}$$

The first terms in brackets equals zero because $\widetilde{D}_{it}$ varies only at the timing-group-by-year level so $D_{it}=\overline{D}_{k(i)t}$. The covariance between y_{it} and $\left(\overline{D}_{k(i)t}-\overline{D}_{k(i)}\right)-\left(\overline{D}_t-\overline{\overline{D}}\right)$ then collapses to group-by-year averages. I use k instead of $k(i)$ hereafter. Apply Lemma 1 to (A8):

$$\frac{\sum_k\sum_{\ell>k}n_\ell n_k\frac{1}{T}\sum_t(\bar{y}_{kt}-\bar{y}_{\ell t})[(\overline{D}_{kt}-\overline{D}_k)-(\overline{D}_{\ell t}-\overline{D}_\ell)]}{\widehat{V}^D} \tag{A9}$$

Now consider the possible values of $\frac{1}{T}\sum_t(\bar{y}_{kt}-\bar{y}_{\ell t})[(\overline{D}_{kt}-D_k)-(\overline{D}_{\ell t}-\overline{D}_\ell)]$. When k is treated at time t_k^* and $\ell=U$ is either never treated or always treated we have $(\overline{D}_{Ut}-\overline{D}_U)=0$ and:

$$-\frac{1}{T}\sum_{t<t_k^*}(\bar{y}_{kt}-\bar{y}_{Ut})\overline{D}_k+\frac{1}{T}\sum_{t\geq t_k^*}(\bar{y}_{kt}-\bar{y}_{Ut})(1-\overline{D}_k) \tag{A10}$$

$$=\left[\left(\bar{y}_{kt}^{POST(k)}-\bar{y}_{kt}^{PRE(k)}\right)-\left(\bar{y}_{Ut}^{POST(k)}-\bar{y}_{Ut}^{PRE(k)}\right)\right]\overline{D}_k(1-\overline{D}_k) \tag{A11}$$

$$=\hat{\beta}_{kU}^{2x2}\overline{D}_k(1-\overline{D}_k) \tag{A12}$$

When k is treated at time t_k^* and ℓ is treated at time $t_\ell^*>t_k^*$, we have:

$$-\frac{1}{T}\sum_{t<t_k^*}(\bar{y}_{kt}-\bar{y}_{\ell t})(\bar{D}_k-\bar{D}_\ell)+\frac{1}{T}\sum_{t\in[t_k^*,t_\ell^*)}(\bar{y}_{kt}-\bar{y}_{\ell t})(1-\bar{D}_k+\bar{D}_\ell)$$

$$-\frac{1}{T}\sum_{t\geq t_\ell^*}(\bar{y}_{kt}-\bar{y}_{\ell t})(\bar{D}_k-\bar{D}_\ell) \qquad (A13)$$

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

$$-\left(\bar{y}_{kt}^{POST(\ell)}-\bar{y}_{\ell t}^{POST(\ell)}\right)\bar{D}_\ell(\bar{D}_k-\bar{D}_\ell) \qquad (A14)$$

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

$$+\left[\left(\bar{y}_{\ell t}^{POST(\ell)}-\bar{y}_{\ell t}^{MID(k,\ell)}\right)-\left(\bar{y}_{kt}^{POST(\ell)}-\bar{y}_{kt}^{MID(k,\ell)}\right)\right]\bar{D}_\ell(\bar{D}_k-\bar{D}_\ell) \qquad (A15)$$

$$=(1-\bar{D}_\ell)^2\hat{\beta}_{k\ell}^{2x2,k}\left(\frac{\bar{D}_k-\bar{D}_\ell}{1-\bar{D}_\ell}\right)\left(\frac{1-\bar{D}_k}{1-\bar{D}_\ell}\right)+\bar{D}_k^2\hat{\beta}_{k\ell}^{2x2,\ell}\left(\frac{\bar{D}_\ell}{\bar{D}_k}\right)\left(\frac{\bar{D}_k-\bar{D}_\ell}{\bar{D}_k}\right) \qquad (A16)$$

Substituting (A12) and (A16) into (A9) and denoting untreated (or always treated) groups by U, earlier treated groups in a dyad by k, and later treated groups by ℓ, establishes equation (10a):

$$\frac{1}{\widehat{V}^D}\left\{\sum_{k\neq U}(n_k+n_U)^2n_{kU}(1-n_{kU})\bar{D}_k(1-\bar{D}_k)\,\widehat{\beta}_{kU}^{2x2}\right.$$

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

$$\left.\left.+\left((n_k+n_\ell)\bar{D}_k\right)^2n_{k\ell}(1-n_{k\ell})\left(\frac{\bar{D}_\ell}{\bar{D}_k}\right)\left(\frac{\bar{D}_k-\bar{D}_k}{\bar{D}_k}\right)\hat{\beta}_{k\ell}^{2x2,\ell}\right]\right\} \qquad (A17)$$

$$\frac{\sum_{k\neq U}(n_k+n_U)^2\hat{V}_{kU}^D\,\widehat{\beta}_{kU}^{2x2}+\sum_{k\neq U}\sum_{\ell>k}\left[\left((n_k+n_\ell)(1-\bar{D}_\ell)\right)^2\hat{V}_{k\ell}^{D,k}\hat{\beta}_{k\ell}^{2x2,k}+\left((n_k+n_\ell)\bar{D}_k\right)^2\hat{V}_{k\ell}^{D,\ell}\hat{\beta}_{k\ell}^{2x2,\ell}\right]}{\widehat{V}^D} \qquad (A18)$$

The denominator, $\widehat{V}^D$ the variance of $\widetilde{D}_{it}$, equals the sum of the terms multiplying the $\hat{\beta}^{2x2}$'s in (A18). This follows by substituting D_{it} for y_{it} in each $\hat{\beta}^{2x2}$ and noting that every term equals 1. Therefore, the weights sum to one: $\sum_{k\neq U}s_{kU}+\sum_{k\neq U}\sum_{\ell>k}\left[s_{k\ell}^k+s_{k\ell}^\ell\right]=1$. ■

APPENDIX B. WHEN IS VWATT OPTIMAL?

Imagine a social welfare function that is a weighted sum of individual value functions:

$$W(e) = \sum_i \eta_i V_i\big(Y_i, \boldsymbol{p}, \theta(e)\big)$$

η_i are social welfare weights (Hendren 2019), V_i is the value function, which takes as arguments income (Y_i), prices ($\boldsymbol{p}$), and a policy $\theta(e)$, which is set optimally according to *some* parameter, e.

Welfare analyses (ie. sufficient statistics papers) typically proceed by defining the parameter that one would need to set policy to maximize $W(e)$ and estimating it by some plausibly valid research design. This, for example, is what can sometimes give rise to the notion of a "parameter of interest". It is "of interest" because it comes out of an optimal policy problem.[1]

In estimating e we focus on internal validity of the point estimate. In other words, if e^* is the true value of the parameter, the role of research design in this literature is to provide consistent point estimates. Assume that optimal policy is set based on the ATT. Consider an estimator, e_0, perhaps a small RCT, whose probability limit equals $e^* = ATT$. Consistency and asymptotic normality implies:

$$e_0 \sim N(e^*, \sigma_0^2)$$

But often there are several estimators available. For example, consider a DD estimator with timing, e_1, whose probability limit does not equal the ATT; it equals VWATT:

$$e_1 \sim N(e^* + x, \sigma_1^2)$$

Should we ever use e_1 instead of e_0? Expected social welfare differs across estimators because of the point estimate and how far off a given estimate is from its probability limit (variance):

[1] Of course other definitions of "interest" are less theoretical. We want to know the average treatment effect on the treated not because of a microeconomic problem, but because it is understandable. Similarly, marginal treatment effects need not come from a social welfare function analysis, but just refer to the effect for those on the current margin of program expansion.

$$W(e_j) = \int \left[\sum_i \eta_i V_i\left(Y_i, \boldsymbol{p}, \theta(\epsilon)\right) \right] f_j(\epsilon) d\epsilon \qquad j = 0,1$$

If we take a second order approximation of V_i around $\epsilon = e^*$, we know that it must be falling in both directions (because it's at a maximum at e^*):

$$V_i\left(Y_i, \boldsymbol{p}, \theta(e^*)\right) + \overbrace{\frac{\partial V_i}{\partial \epsilon}}^{0 \text{ by envelope}} (\epsilon - e^*) + \overbrace{\left.\frac{\partial^2 V_i}{(\partial \epsilon)^2}\right|_{\epsilon = e^*}}^{<0 \text{ by SOC}} (\epsilon - e^*)^2$$

We can write welfare using estimator j as:

$$W(e_j) \approx W(e^*) + \int \left[\sum_i \eta_i \left.\frac{\partial^2 V_i}{(\partial \epsilon)^2}\right|_{\epsilon = e^*} (\epsilon - e^*)^2 \right] \phi_j(\epsilon) d\epsilon \qquad j = 0,1$$

Where ϕ_j is the asymptotic normal distribution of e_j (the analysis does not rely on asymptotic normality, though). Bring the integral inside:

$$W(e_j) \approx W(e^*) + \sum_i \eta_i \left.\frac{\partial^2 V_i}{(\partial \epsilon)^2}\right|_{\epsilon = e^*} \int (\epsilon - e^*)^2 \phi_j(\epsilon) d\epsilon \qquad j = 0,1$$

Note that when $plim\ e_j = e^*$, the integral is the asymptotic variance or e_j. So for e_0:

$$W(e_0) \approx W(e^*) + \sigma_0^2 \overbrace{\sum_i \eta_i \left.\frac{\partial^2 V_i}{(\partial \epsilon)^2}\right|_{\epsilon = e^*}}^{\equiv \nu < 0}$$

$$= W(e^*) + \sigma_0^2 \nu$$

The sum, defined as ν, weights together how responsive welfare is (in its second derivative) to the policy using the social welfare weights. The presence of σ_0^2 immediately clarifies the social benefits from more precise estimators: we are less likely to be wrong and set policies that are far from e^*.

But what about e_1, the estimator that was not consistent for e^*?

$$W(e_1) \approx W(e^*) + \sum_i \eta_i \frac{\partial^2 V_i}{(\partial \epsilon)^2}\bigg|_{\epsilon = e^*} \int (\epsilon - (e^* + x) + x)^2 f_1(\epsilon) d\epsilon$$

The integral simplifies easily:

$$\overbrace{\int \big(\epsilon - (e^* + x)\big)^2 f_1(\epsilon) d\epsilon}^{\sigma_1} + x^2 \overbrace{\int f_1(\epsilon) d\epsilon}^{1} + 2x \overbrace{\int \big(\epsilon - (e^* + x)\big) f_1(\epsilon) d\epsilon}^{0}$$

So we have:

$$W(e_1) \approx W(e^*) + \overbrace{(\sigma_1^2 + x^2)}^{MSE(e_1)} \nu$$

The bias is costly and so is the variance. But, crucially, now we can compare the two. Specifically, the biased estimator is better in an expected social welfare sense if $W(e_1) - W(e_0)$, or:

$$(\sigma_1^2 + x^2)\nu > \ \sigma_0^2 \nu$$

$$\overbrace{\sigma_1^2 + x^2}^{MSE_1} < \ \overbrace{\sigma_0^2}^{MSE_0}$$

If an estimator that is not consistent for "the parameter of interest" is sufficiently precise, it may be preferable in this sense. The social welfare context shows that the optimal estimator is the one that minimizes mean-squared error, not necessarily one that is consistent for a given parameter of interest.

APPENDIX C: PROOF OF THE CONTROLLED DD DECOMPOSITION

This section derives equation (26). The covariance between y_{it} and $\tilde{d}_{kt}$ is:

$$\hat{C}(y_{it}, \tilde{d}_{kt}) = \frac{1}{NT}\sum_k \left(\sum_{i \in k} \sum_t y_{it}\left[(\bar{d}_{kt} - \bar{d}_k) - (\bar{d}_t - \bar{\bar{d}}) \right] \right) \quad (C1)$$

$$= \frac{1}{T}\sum_k n_k \sum_t \bar{y}_{kt}\left[(\bar{d}_{kt} - \bar{d}_k) - (\bar{d}_t - \bar{\bar{d}}) \right] \quad (C2)$$

$$= \sum_k \sum_{\ell > k} n_k n_\ell \frac{1}{T} \sum_t (\bar{y}_{kt} - \bar{y}_{\ell t})\left[(\bar{d}_{kt} - \bar{d}_k) - (\bar{d}_{\ell t} - \bar{d}_\ell) \right] \quad (C3)$$

Where the last equality follows from Lemma 1. By the definition of $\tilde{d}_{it}$, though, this covariance can be written as a function of the pairwise two-way fixed effects coefficients by rescaling each term by the variance of either $\bar{D}_{kt}$ or $\bar{p}_{kt}$—averages of $\tilde{p}_{it}$ by group and time period—adjusted for fixed effects in the dyad:

$$\sum_k \sum_{\ell > k} (n_k + n_\ell)^2 \left[\hat{V}^D_{k\ell} \hat{\beta}^{2x2}_{k\ell} - \hat{V}^p_{b,k\ell} \hat{\beta}^p_{b,k\ell} \right] \quad (C4)$$

Where $\hat{\beta}^{2x2}_{k\ell}$ is defined in Theorem 1, and $\hat{\beta}^p_{b,k\ell} \equiv \frac{\sum_t (\bar{y}_{kt} - \bar{y}_{\ell t})[(\bar{p}_{kt} - \bar{p}_k) - (\bar{p}_{\ell t} - \bar{p}_\ell)]}{\sum_t [(\bar{p}_{kt} - \bar{p}_k) - (\bar{p}_{\ell t} - \bar{p}_\ell)]^2}$ is the coefficient from a subsample regression of $\bar{y}_{jt}$ on $\bar{p}_{jt}$ with unit and time fixed effects. The variances, $\hat{V}^D_{k\ell} \equiv \frac{n_k}{n_k + n_\ell} \frac{n_\ell}{n_k + n_\ell} \frac{1}{T} \sum_t [(\bar{D}_{kt} - \bar{D}_k) - (\bar{D}_{\ell t} - \bar{D}_\ell)]^2$ and $\hat{V}^p_{b,k\ell} \equiv \frac{n_k}{n_k + n_\ell} \frac{n_\ell}{n_k + n_\ell} \frac{1}{T} \sum_t [(\bar{p}_{kt} - \bar{p}_k) - (\bar{p}_{\ell t} - \bar{p}_\ell)]^2$ are the residual variances from regressions of $\widetilde{D}_{jt}$ of $\tilde{p}_{jt}$ on group and year fixed effects in the sample of groups k and ℓ. This establishes equation (26) in the text.

A. How do covariates affect the estimand and identifying assumption?

Here I focus on the theoretical implications for a 2x2 DD that compares a timing group to an untreated group. I make the following assumptions:

1. Covariates do not vary within groups. This eliminates the within term in (25) and simplifies the analysis.
2. No heterogeneity in the Frisch-Waugh coefficients. This eliminates $\hat{V}_{b,k\ell}^{dp}\hat{\beta}_{b,k\ell}^{dp}$ in footnote 36, the term that reflects misspecification in the covariates for subgroup Frisch-Waugh regressions and ensures the controlled decomposition weights together 2x2 coefficients that themselves come from controlled regressions.

Under these assumptions, a given 2x2 involving group U is:

$$\hat{\beta}_{kU}^{2x2|X} = \frac{\hat{C}\left(y_{it}, \widetilde{D}_{jt}\right) - \hat{C}\left(y_{it}, \tilde{p}_{jt}^{kU}\right)}{(1 - R_{kU}^2)\hat{V}_{kU}^{D}} \tag{C5}$$

Substituting in potential outcomes we have:

$$\hat{C}\left(y_{it}, \widetilde{D}_{jt}\right) = \hat{V}_{kU}^{D} \frac{\overbrace{\sum_{t \geq t_k^*}\left(\bar{Y}_{kt}(1) - \bar{Y}_{kt}(0)\right)}^{ATT_k(POST(k))}}{T - (t_k^* - 1)}$$

$$+ \hat{V}_{kU}^{D} \overbrace{\left[\Delta Y_k^0(POST(k), PRE(k)) - \Delta Y_U^0\left(POST(k), PRE(k)\right)\right]}^{Common\ Trends} \tag{C6}$$

And:

$$\hat{C}\left(y_{it}, \tilde{p}_{jt}^{kU}\right) = \frac{n_{kU}(1 - n_{kU})}{T} \sum_t \left(\bar{Y}_{kt}(0) + D_{kt}\left(\bar{Y}_{kt}(1) - \bar{Y}_{kt}(0)\right) - \bar{Y}_{Ut}(0)\right)\left(\left(\bar{p}_{kt}^{kU} - \bar{p}_k^{kU}\right)\right.$$

$$\left. - \left(\bar{p}_{Ut}^{kU} - \bar{p}_U^{kU}\right)\right) \tag{C7}$$

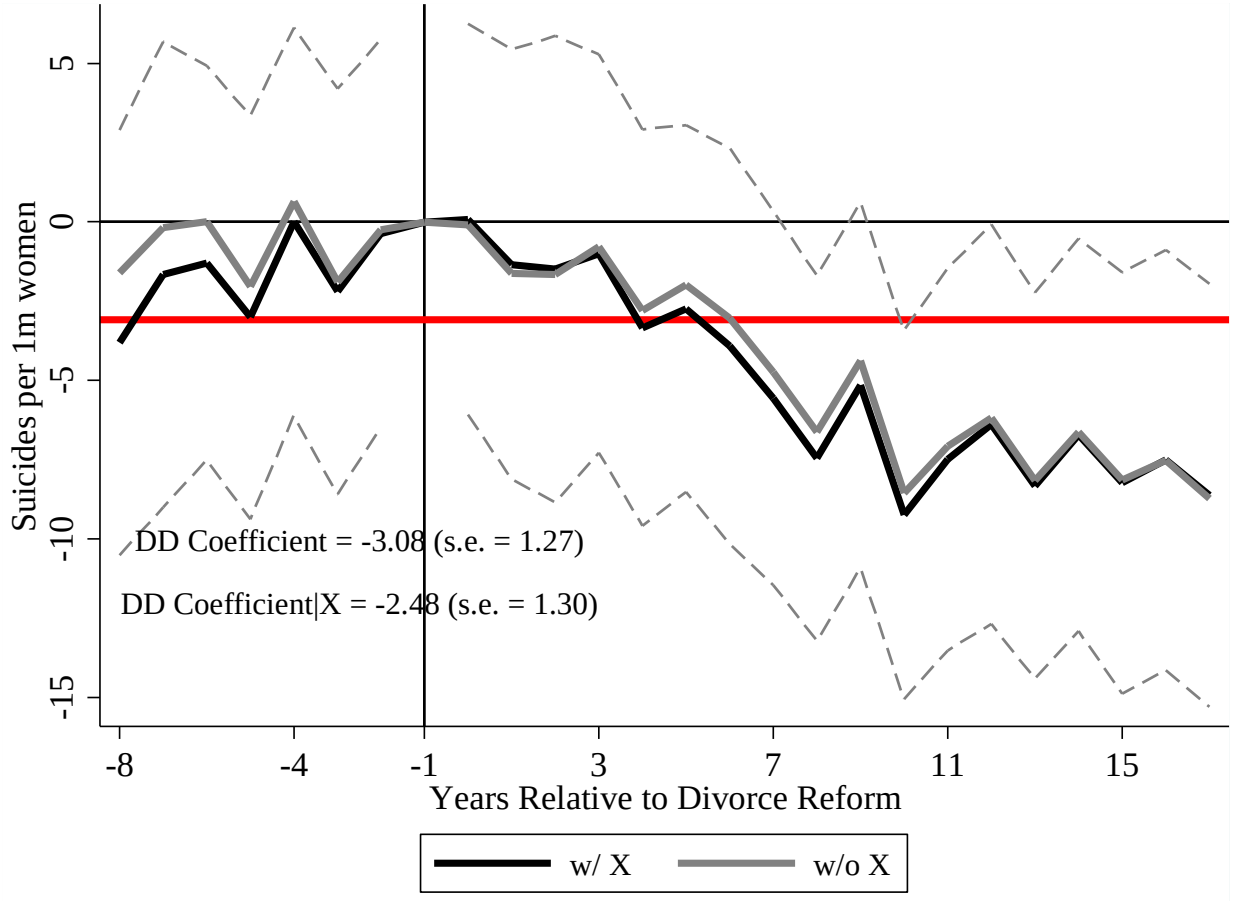
$$= \overbrace{\frac{n_{kU}(1-n_{kU})}{T}\sum_{t}\left(\bar{Y}_{kt}(0)-\bar{Y}_{Ut}(0)\right)\left(\left(\bar{p}_{kt}^{kU}-\bar{p}_{k}^{kU}\right)-\left(\bar{p}_{Ut}^{kU}-\bar{p}_{U}^{kU}\right)\right)}^{Covariance\ between\ controls\ and\ differential\ trends}$$

$$+ \overbrace{\frac{n_{kU}(1-n_{kU})}{T-\left(t_g^*-1\right)}\sum_{t\geq t_k^*}\left(\bar{Y}_{kt}(1)-\bar{Y}_{kt}(0)\right)\left(\left(\bar{p}_{kt}^{kU}-\bar{p}_{k}^{kU}\right)-\left(\bar{p}_{Ut}^{kU}-\bar{p}_{U}^{kU}\right)\right)}^{year-to-year\ covariance\ between\ controls\ and\ group-time\ treatment\ effects} \qquad (C8)$$

Even with no additional assumptions about counterfactual outcomes, we can already see that part of what covariates do is to absorb variation in treatment effects. The second term is the extent to which relative changes in the covariate index in the treatment and control groups in the post period are correlated with changing treatment effects. (Note that it is not a proper covariance demeaned on the post period.) Covariates can therefore change the 2x2 DDs even with no bias because they "control for" changing treatment effects.

This has been observed for specific controls, like unit-specific time trends (discussed below), but not for general covariates. Equation (C8) suggests that a potentially pervasive reason why controls change specifications is by subtracting treatment effect parameters rather than counterfactual trends. A simple way to analyze this is to compare changes in event-study coefficients and DD coefficients with and without controls. In the unilateral divorce analysis, I estimate both the event-study and DD coefficients with and without controls averaged at the group-by-year level (to eliminate the "within" comparisons that drive the estimate with state-by-year controls). The magnitude of the post-period event-study estimates with covariates are on average 98 of the uncontrolled coefficients, but the DD estimate with (averaged) controls is 80 percent as large, consistent with the second line of equation (C8).

Appendix Figure C.1. Event-Study and DD Estimates with and without Covariates



The main point of adding controls, though, is to address differential trends. Do the two common trends terms in (C6) and (C8) cancel? First, note that common trends in the unadjusted DD coefficient only involves pre/post means of $\bar{Y}_{kt}(0) - \bar{Y}_{Ut}(0)$, not period-to-period changes. The control term, however, is a covariance between $\bar{Y}_{kt}(0) - \bar{Y}_{Ut}(0)$ and differences in covariates across all periods because $\boldsymbol{X_{it}}$ itself must vary period-to-period in order for its coefficients to be identified in the main estimating equation. Breaking up the sum, adding and subtracting averages of $\bar{p}_{kt}^{kU}$ and $\bar{p}_{Ut}^{kU}$, and simplifying using the definition $\bar{p}_{k}^{kU} = (1 - \bar{D}_k)\bar{p}_{k,PRE(k)}^{kU} + \bar{D}_k\bar{p}_{k,POST(k)}^{kU}$ expresses the control term into a piece that varies just like the common trends term (pre/post t_k^*)

and another term measuring the period-to-period changes within the pre/post periods. The control term then nets out the following function of counterfactual trends:

$$\left[\frac{1}{T}\sum_{t<t_k^*}\left(\bar{Y}_{kt}(0)-\bar{Y}_{Ut}(0)\right)\overbrace{\left(\left(\bar{p}_{kt}^{kU}-\bar{p}_{k}^{kU}\right)-\left(\bar{p}_{Ut}^{kU}-\bar{p}_{U}^{kU}\right)\right)}^{+/-\,\bar{p}_{k,PRE(k)}^{kU}\,and\,+/-\,\bar{p}_{U,PRE(k)}^{kU}}\right.$$

$$\left.+\frac{1}{T}\sum_{t\geq t_k^*}\left(\bar{Y}_{kt}(0)-\bar{Y}_{Ut}(0)\right)\overbrace{\left(\left(\bar{p}_{kt}^{kU}-\bar{p}_{k}^{kU}\right)-\left(\bar{p}_{Ut}^{kU}-\bar{p}_{U}^{kU}\right)\right)}^{+/-\,\bar{p}_{k,POST(k)}^{kU}\,and\,+/-\,\bar{p}_{U,POST(k)}^{kU}}\right] \qquad (C9)$$

$$\bar{D}_k(1-\bar{D}_k)\left(\left(\bar{p}_{k,POST(k)}^{kU}-\bar{p}_{k,PRE(k)}^{kU}\right)-\left(\bar{p}_{U,POST(k)}^{kU}-\bar{p}_{U,PRE(k)}^{kU}\right)\right)\left[\Delta Y_k^0(POST(k),PRE(k))\right.$$

$$\left.-\Delta Y_U^0\left(POST(k),PRE(k)\right)\right]$$

$$+\left[\frac{1}{T}\sum_{t<t_k^*}\left(\bar{Y}_{kt}(0)-\bar{Y}_{Ut}(0)\right)\left(\left(\bar{p}_{kt}^{kU}-\bar{p}_{k,PRE(k)}^{kU}\right)-\left(\bar{p}_{Ut}^{kU}-\bar{p}_{U,PRE(k)}^{kU}\right)\right)\right.$$

$$\left.+\frac{1}{T}\sum_{t\geq t_k^*}\left(\bar{Y}_{kt}(0)-\bar{Y}_{Ut}(0)\right)\left(\left(\bar{p}_{kt}^{kU}-\bar{p}_{k,POST(k)}^{kU}\right)-\left(\bar{p}_{Ut}^{kU}-\bar{p}_{U,POST(k)}^{kU}\right)\right)\right] \qquad (C10)$$

The first part of (C10) reflects pre/post averages in counterfactual outcomes just like the potential bias in the unadjusted 2x2 DD. When the covariates exactly match the underlying model of D_{it} the predicted values $\left(\left(\bar{p}_{k,POST(k)}^{kU}-\bar{p}_{k,PRE(k)}^{kU}\right)-\left(\bar{p}_{U,POST(k)}^{kU}-\bar{p}_{U,PRE(k)}^{kU}\right)\right)\approx 1$ and this cancels out the differential trend reproduced in (C6). Equation (C10) makes it clear why measurement error or misspecification limit the value of controls. In these cases $\left(\left(\bar{p}_{k,POST(k)}^{kU}-\bar{p}_{k,PRE(k)}^{kU}\right)-\left(\bar{p}_{U,POST(k)}^{kU}-\bar{p}_{U,PRE(k)}^{kU}\right)\right)\neq 1$ and does not cancel out differential trends in Y^0.

Even with perfectly measured and specified covariates, the second part of (C10) which measures the covariance between period-to-period changes in $\bar{Y}_{kt}(0) - \bar{Y}_{Ut}(0)$ and the controls within the pre- and post-periods could be non-zero. The numerator of $\hat{\beta}_{kU}^{2x2|X}$ subtracts this term, yet it does not match the bias in $\hat{\beta}_{kU}^{2x2}$, which comes only from changes in average untreated outcomes in the pre/post periods. This is analogous to the within term in equation (25) because it comes from the fact that $\boldsymbol{X_{it}}$ varies differently (period-to-period) than D_{it} (pre/post). When we control for a variable like this in a regression framework we need this variation to identify the control coefficients but then necessarily subtract it from our 2x2 DD estimate.[2]

I have focused on one kind of 2x2 comparison, but the DD decomposition for controlled specifications shows how these kinds of conclusions—over-controlling for heterogeneous treatment effects, the influence of misspecification or measurement error, and partialling out period-by-period effects—would aggregate up to the full DD estimate. I have also made simplifying assumptions that eliminated the within-group comparisons across $\tilde{p}_{it}$ and the potential for subsample heterogeneity in $\tilde{p}_{it}^{kU}$, both of which would require their own potential outcomes analysis. Lastly, I have not discussed what is required for a given covariate to be an admissible control. The would come from an explicit statement about the potential outcomes for X_{it} itself and the assumption that X_{it} is not caused by D_{it}.

[2] In the simplest possible 2x2 DD—only 4 data points—any control variable is perfectly collinear with $\widetilde{D}_{it}$. Therefore the within-group variation and period-to-period variation are necessary to identify the control coefficients, but then this variation partly drives the estimated DD.

APPENDIX D: ALTERNATIVE SPECIFICATIONS

A. Triple-Difference Estimator

When some units should not be (as) affected by a given treatment, they can be used as a falsification test. Assume that units belong to either an affected group ($G_i = 1$) or an unaffected group ($G_i = 0$). The simplest way to incorporate the "third difference", G_i, would be to estimate separate DD coefficients in each sub-sample: $\widehat{\beta}_0^{DD}$ and $\widehat{\beta}_1^{DD}$. One could equivalently estimate the following triple-difference specification (DDD) on the pooled sample including interactions of G_i with all variables from equation (2):[3]

$$y_{it} = \alpha_i + \alpha_t + \beta_0^{DD} D_{it} + \alpha_t G_i + \beta^{DDD} D_{it} G_i + e_{it} \qquad (D1)$$

$\widehat{\beta}_0^{DD}$ is the two-way fixed effects DD estimate for the $G_i = 0$ sample and $\widehat{\beta}^{DDD}$ equals the difference between the sub-sample DD coefficients: $\widehat{\beta}_1^{DD} - \widehat{\beta}_0^{DD}$.

One problem with this estimator is that $\widehat{\beta}_1^{DD}$ equals an average weighted by the cross-sectional distribution in the in the $G_i = 1$ sample, but $\widehat{\beta}_0^{DD}$ uses the cross-sectional distribution of the $G_i = 0$ sample. If G_i were an indicator for black respondents, then 2x2 DD's that included Southern states would get more weight in the black than the white sample while the opposite would be true for Vermont and New Hampshire. Estimates of (D1) difference out cross-state/cross-year changes in white outcomes weighted by white populations, and so may not capture relative trends by race *within* states. A null result for $\widehat{\beta}_0^{DD}$, which is typically reassuring, could be driven by completely different 2x2 DD's than the ones that matter most for $\widehat{\beta}_1^{DD}$.

[3] In this set up the third difference partitions units, so $\alpha_i G_i$ is collinear with α_i.

DDD specifications that include a more saturated set of fixed effects overcome this problem. If treatment rolled out by state (s), for example, a DDD can include state-by-time fixed effects ($\alpha_s\alpha_t$):

$$y_{it} = \alpha_i + G_i\alpha_s + G_i\alpha_t + \alpha_s\alpha_t + \beta^{DDD}D_{it}G_i + e_{it} \qquad (D2)$$

The DDD estimate from (D2) does equal a weighted average of 2x2 DD's. $\widehat{\beta}^{DDD}$ is equivalent to first collapsing the data to mean differences between G-groups within (s,t) cells, $\bar{y}_{st}^{G=1} - \bar{y}_{st}^{G=0}$, then estimating a DD weighted by cell sizes times the $\widehat{var}(G_i|s,t) = g_{st}(1-g_{st})$, where g_{st} is the mean of G_i by s and t. Unlike (D1), estimates from (D2) *do* net out changes across G_i within state/year cells. This changes the weights, though, because the introduction of variation across the third difference within a cell leads to the typical OLS result that cells with more variation get more weight. In this case, "more variation" in sample membership within a cell means approximately equal numbers of units with $G_i = 1$ and $G_i = 0$.[4]

Recasting this version of DDD as a DD on differences by G_i implies that the decomposition theorem holds, albeit with a slight change to the calculation of the weights. All the results and diagnostic tools derived above apply to specifications like (20) by defining the outcome as $\bar{y}_{st}^{G=1} - \bar{y}_{st}^{G=0}$ and using the proper weights.

B. Unit-specific linear time trends

Researchers control for unit-specific linear time trends to allow "treatment and control states to follow different trends" (Angrist and Pischke 2009, p 238), and view it as "an important check on the causal interpretation of any set of regression DD estimates" (Angrist and Pischke 2015, p. 199). The specification is:

[4] Collapsing the data to the within-cell mean differences $\bar{y}_{st}^{G=1} - \bar{y}_{st}^{G=0}$ and running OLS on the aggregated data (or WLS using cell populations) would eliminate the $g_{st}(1-g_{st})$ from the decomposition weights.

$$y_{it} \quad = \quad \alpha_i + \alpha_t + A^{k(i)}\alpha_i\left(t - \overline{t}\right) + \beta_{trend}^{DD} D_{it} + e_{it} \qquad (D3)$$

Unit-specific trends are just a particular type of control variable that does not vary across units within a timing group—since the fitted trend through $\widetilde{D}_{it}$ is the same for all units in the same group—and for which the full-sample Frisch-Waugh relationship is the same as any subsample relationship—since the trends are specific to each unit. In the notation of section IV.B, $\Omega = 0$ and $\tilde{p}_{jt} = \tilde{p}_{jt}^{k\ell}$, so that every 2x2 DD also comes from a subsample detrended regression specification.

The specific way that unit-specific trends change each 2x2 component fit with previous intuition. They essentially subtract the cross-group difference in averages of y before and after the middle of the panel, $\bar{t}$, but these differences are weighted by absolute distance to $\bar{t}$: $|t - \bar{t}|$. This is, as Lee and Solon (2011) point out, akin to a regression discontinuity design in that the estimator relies less on variation at the beginning or end of the panel because this variation is absorbed by the trends. Unfortunately, trends tend to absorb time-varying treatment effects that are necessarily larger at the end of the panel, and in these cases they over control. Unit-specific trends also increase the weight on units treated at the extremes of the panel, changing estimates for this reason as well.

i. What is the linear trend through $\widetilde{D}_{it}$ for each 2x2 DD?

The Frisch-Waugh expression for the linear trend through $\widetilde{D}_{it}$ for each unit i in a two-group subsample is:

$$A^j = \frac{\frac{1}{T}\sum_t \widetilde{D}_{it}\left(t - \overline{t}\right)}{\frac{1}{T}\sum_t \left(t - \overline{t}\right)^2} \qquad \forall\ i \in k, \ell \qquad (D4)$$

The numerator of each of those regressions is:

$$\begin{aligned}
\frac{1}{T}\sum_t \widetilde{D}_{it}\,(t - \overline{t}) \quad &= \quad \frac{1}{T}\left[\sum_t \left(D_{it} - \overline{\overline{D}}\right)(t - \overline{t}) - \sum_t \left(\overline{D}_i - \overline{\overline{D}}\right)(t - \overline{t}) - \sum_t \left(\overline{D}_t - \overline{\overline{D}}\right)(t - \overline{t})\right] \\
&= \quad \frac{1}{T}\left[\sum_t D_{it}(t - \overline{t}) - \left(\overline{D}_i - \overline{\overline{D}}\right)\overbrace{\sum_t (t - \overline{t})}^{0} - \sum_t \overline{D}_t(t - \overline{t})\right] \\
&= \quad \frac{1}{T}\sum_t \left(D_{it} - \overline{D}_t\right)(t - \overline{t}) \qquad (D5)
\end{aligned}$$

The path of D_{it} is obvious for all groups and the path of $\overline{D}_t$ depends on the two groups in the subsample. For groups k and U we have:

$$A_{kU}^{k} = \frac{\frac{1}{T}\sum_{t=t_k^*}^{T}(1-n_{kU})\left(t-\overline{t}\right)}{\frac{1}{T}\sum_t\left(t-\overline{t}\right)^2} \qquad (D6)$$

The sum in the numerator can be written as:

$$\begin{aligned}
&\sum_{t=t_k^*}^{T}\left(t-\overline{t}\right) = \\
&\quad = \left(\sum_{t=t_k^*}^{T} t\right) - \overline{t}\left(T - \left(t_k^*-1\right)\right) \\
&\quad = \left(\sum_{j=1}^{T-(t_k^*-1)} j + (t_k^*-1)\right) - \overline{t}\left(T - \left(t_k^*-1\right)\right) \\
&\quad = \frac{\left(T-(t_k^*-1)\right)(T-(t_k^*-1)+1)}{2} + (t_k^*-1)(T-(t_k^*-1)) - \overline{t}\left(T - \left(t_k^*-1\right)\right) \\
&\quad = \left[\frac{(t_k^*-1)(T-(t_k^*-1))}{2}\right] \\
&= \frac{T^2}{2}\overline{D}_k(1-\overline{D}_k) \qquad (D7)
\end{aligned}$$

The denominator equals $\frac{(T+1)(T-1)}{12}$.[5] So he have:

$$A_{kU}^{k} = \frac{6T}{(T+1)(T-1)}(1-n_{kU})\overline{D}_k(1-\overline{D}_k) \qquad (D8)$$

The fitted trend through group U in this subsample is the same but of the opposite sign:

$$A_{kU}^{U} = -\frac{6T}{(T+1)(T-1)}(1-n_{kU})\overbrace{\overline{D}_k(1-\overline{D}_k)}^{V_k} \qquad (D9)$$

Where I use $V_k \equiv \overline{D}_k(1-\overline{D}_k)$ to save space. A similar derivation shows that for the comparison of two timing groups k and ℓ the fitted trends are:

---

[5] The variance of $t-\overline{t}$ can be rewritten as the sums over demeaned time not as sums from 1 to T, but from $1-\overline{t} = 1-\frac{T+1}{2} = -\frac{T-1}{2} = -(\overline{t}-1)$ to $T-\overline{t} = \frac{2T}{2} - \frac{T+1}{2} = \frac{T-1}{2} = (\overline{t}-1)$. The variance then is just the sum of squared integers, so the negative and the positive portions are equal:

$$\frac{1}{T}\sum_{j=-(\overline{t}-1)}^{(\overline{t}-1)} j^2 = \frac{2}{T}\sum_{j=1}^{(\overline{t}-1)} j^2 = 2\left[\frac{\overline{t}(\overline{t}-1)(2\overline{t}-1)}{6T}\right] = \left[\frac{\overline{t}(\overline{t}-1)}{3}\right] = \frac{(T+1)(T-1)}{12}$$

Where the last equality follows because our assumption that T is odd implies that $\left(2\overline{t}-1\right) = T$, and $\overline{t}\left(\overline{t}-1\right) = \left(\frac{T+1}{2}\right)\left(\frac{T-1}{2}\right) = \frac{(T+1)(T-1)}{4}$.

$$A^k_{k\ell} = \frac{6T}{(T+1)(T-1)}(1-n_{k\ell})\overbrace{\left(\overline{D}_k(1-\overline{D}_k) - \overline{D}_\ell(1-\overline{D}_\ell)\right)}^{V_k - V_\ell} \quad (D10)$$

$$A^\ell_{k\ell} = -\frac{6T}{(T+1)(T-1)} n_{k\ell}\left(\overline{D}_k(1-\overline{D}_k) - \overline{D}_\ell(1-\overline{D}_\ell)\right) \quad (D11)$$

ii. What do linear trends partial out of each two-group estimator?

With the definitions of these trend coefficients, it is straightforward to calculate the regression coefficient that relates the outcome to the fitted trends. For 2x2 estimators that compare a treated group to the untreated group, we have

$$\hat{\beta}^{trend}_{kU} = \frac{\frac{1}{NT}\sum_{i\in k,U}\sum_t A^i_{kU}(t-\bar{t})\, y_{it}}{\frac{1}{NT}\sum_i\sum_t\left[A^i_{kU}(t-\bar{t})\right]^2} \quad (D12)$$

$$= \frac{\frac{1}{T} n_k A^k_{kU}\sum_t (t-\bar{t})\,\bar{y}_{kt} + \frac{1}{T} n_U A^U_{kU}\sum_t (t-\bar{t})\,\bar{y}_{Ut}}{\left(n_{kU}{A^k_{kU}}^2 + (1-n_{kU}){A^U_{kU}}^2\right)\frac{1}{T}\sum_t[(t-\bar{t})]^2} \quad (B13)$$

Where the second line follows because the sums over units, i, collapse to weighted sums over groups, k.

The sum of each group's time means, $\bar{y}_{kt}$, times $(t-\bar{t})$ has a positive part when $t > \bar{t}$, and a negative part when $t < \bar{t}$. Each portion equals a mean of y before or after $\bar{t}$ that use $|t-\bar{t}|$ as weights. We have:

$$\sum_t (t-\bar{t})\,\bar{y}_{jt} = \sum_{t=\bar{t}}^{T}(t-\bar{t})\left[\frac{\overbrace{\sum_{t=\bar{t}}^{T}(t-\bar{t})\,\overline{y}_{jt}}^{post-\bar{t}}}{\sum_{t=\bar{t}}^{T}(t-\bar{t})} - \frac{\overbrace{\sum_{t=1}^{\bar{t}}(\bar{t}-t)\,\overline{y}_{kt}}^{pre-\bar{t}}}{\sum_{t=1}^{\bar{t}}(\bar{t}-t)}\right]$$

$$= \frac{(T+1)(T-1)}{8}\Delta^*\bar{y}_j \quad (D14)$$

Where the second line uses $\Delta^*\bar{y}_j$ to denote the weighted differences in y before and after $\bar{t}$, and the fact that the sum of integers $\sum_{j=1}^{\bar{t}-1} j = \frac{\bar{t}(\bar{t}-1)}{2} = \frac{(T+1)(T-1)}{8}$. The denominator also contains the

variance of t, which was defined above. Substituting into the expression for a regression of y_{it} on fitted trends in a pair yields:

$$\hat{\beta}_{kU}^{trend} = \frac{\frac{(T+1)(T-1)}{8T}\left[n_{kU}A_{kU}^{k}\Delta^{*}\bar{y}_{k} + (1-n_{kU})A_{kU}^{U}\Delta^{*}\bar{y}_{U}\right]}{\frac{(T+1)(T-1)}{12}\left(n_{kU}{A_{kU}^{k}}^{2} + (1-n_{kU}){A_{kU}^{U}}^{2}\right)} \quad (D15)$$

Substituting the definitions of fitted trends shows that $\hat{\beta}_{kU}^{trend}$ is:

$$\hat{\beta}_{kU}^{trend} = \frac{\frac{3}{4}n_{kU}(1-n_{kU})V_{k}[\Delta^{*}\bar{y}_{k} - \Delta^{*}\bar{y}_{U}]}{\frac{(T+1)(T-1)}{12}\frac{36T^{2}}{(T+1)^{2}(T-1)^{2}}(n_{kU}(1-n_{kU})V_{k}^{2})}$$

$$\frac{(T+1)}{T}\frac{(T-1)}{T}\frac{V^{*}}{V_{k}}(\Delta^{*}\bar{y}_{k} - \Delta^{*}\bar{y}_{U}) \quad (D16)$$

I replace the $\frac{1}{4}$ with V^{*}, which I define as the variance of a dummy that equals one after $\bar{t}$.

A similar derivation for a two-group timing-only estimator yields:

$$\hat{\beta}_{k\ell}^{trend} = \frac{(T+1)}{T}\frac{(T-1)}{T}\frac{V^{*}}{V_{k} - V_{\ell}}(\Delta^{*}\bar{y}_{k} - \Delta^{*}\bar{y}_{\ell}) \quad (D17)$$

Equations (B18) and (B19) show that the coefficient that relates outcomes and a trend line fitted through each group's treatment variable equals the difference in a time-weighted average before and after time $\bar{t}$, scaled by a term measuring how close to the middle of the panel they are treated.

iii. What are the detrended 2x2 estimators?

The definition of $\hat{\beta}_{b,k\ell}^{d}$ in equation (26) in the paper collapses to $\hat{\beta}_{k\ell}^{2x2,trend} = \frac{\hat{\beta}_{k\ell}^{2x2} - R_{k\ell}^{2}\hat{\beta}_{k\ell}^{trend}}{(1-R_{k\ell}^{2})}$. The only term left to solve for is the $R_{k\ell}^{2}$ from each two-group Frisch-Waugh regression. That is just the ratio of the variance of the fitted trend to the variance of the treatment dummy. The variance of the fitted trend is:

$$\widehat{var}\left(A_{ku}^{j}(t-\bar{t})\right) = \frac{(T+1)(T-1)}{12}\left(n_{kU}{A_{kU}^{k}}^{2} + (1-n_{kU}){A_{kU}^{U}}^{2}\right)$$

$$= \frac{(T+1)(T-1)}{12}\frac{36T^2}{(T+1)^2(T-1)^2}n_{kU}(1-n_{kU})[\bar{D}_k(1-\bar{D}_k)]^2$$

$$= 3\frac{T}{T+1}\frac{T}{T-1}n_{kU}(1-n_{kU})[\bar{D}_k(1-\bar{D}_k)]^2 \qquad (D18)$$

But we also know from (7) that $\hat{V}_{kU}^{D} = n_{kU}(1-n_{kU})\bar{D}_k(1-\bar{D}_k)$, so:

$$R_{kU}^2 = 3\frac{T}{T+1}\frac{T}{T-1}\bar{D}_k(1-\bar{D}_k) = 3\frac{T}{T+1}\frac{T}{T-1}V_k \qquad (D19)$$

Therefore,

$$\hat{\beta}_{kU}^{2x2,trend} = \left(\bar{y}_k^{POST(k)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_U^{POST(k)} - \bar{y}_U^{PRE(k)}\right)$$
$$+ \frac{3\frac{T}{T+1}\frac{T}{T-1}V_k}{1-3\frac{T}{T+1}\frac{T}{T-1}V_k}\left[\left(\left(\bar{y}_k^{POST(k)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_U^{POST(k)} - \bar{y}_U^{PRE(k)}\right)\right)\right.$$
$$\left. - \frac{(T+1)}{T}\frac{(T-1)}{T}\frac{V^*}{V_k}(\Delta^*\bar{y}_k - \Delta^*\bar{y}_U)\right] \qquad (D20)$$

$$\hat{\beta}_{k\ell}^{2x2,trend} = \left[\frac{1-\bar{D}_k}{1-(\bar{D}_k-\bar{D}_\ell)}\right]\left[\left(\bar{y}_k^{MID(k,\ell)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_\ell^{MID(k,\ell)} - \bar{y}_\ell^{PRE(k)}\right)\right]$$
$$+ \left[\frac{\bar{D}_\ell}{1-(\bar{D}_k-\bar{D}_\ell)}\right]\left[\left(\bar{y}_\ell^{POST(\ell)} - \bar{y}_\ell^{MID(k,\ell)}\right) - \left(\bar{y}_k^{POST(\ell)} - \bar{y}_k^{MID(k,\ell)}\right)\right]$$
$$+ \frac{3\frac{T}{T+1}\frac{T}{T-1}\frac{(V_k-V_\ell)^2}{(\bar{D}_k-\bar{D}_\ell)(1-(\bar{D}_k-\bar{D}_\ell))}}{1-3\frac{T}{T+1}\frac{T}{T-1}\frac{(V_k-V_\ell)^2}{(\bar{D}_k-\bar{D}_\ell)(1-(\bar{D}_k-\bar{D}_\ell))}}\left[\left[\frac{1-\bar{D}_k}{1-(\bar{D}_k-\bar{D}_\ell)}\right]\left[\left(\bar{y}_k^{MID(k,\ell)}\right.\right.\right.$$
$$\left. - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_\ell^{MID(k,\ell)} - \bar{y}_\ell^{PRE(k)}\right)\Big]$$
$$+ \left[\frac{\bar{D}_\ell}{1-(\bar{D}_k-\bar{D}_\ell)}\right]\left[\left(\bar{y}_\ell^{POST(\ell)} - \bar{y}_\ell^{MID(k,\ell)}\right) - \left(\bar{y}_k^{POST(\ell)} - \bar{y}_k^{MID(k,\ell)}\right)\right]$$
$$\left. - \frac{(T+1)}{T}\frac{(T-1)}{T}\frac{V^*}{V_k-V_\ell}(\Delta^*\bar{y}_k - \Delta^*\bar{y}_\ell)\right] \qquad (D21)$$

Note that because $R^2_{k\ell}$ has $V_k - V_\ell$ in the numerator, it equals zero when the variance of treatment is the same for two timing groups. This happens when they are treated equally close to the ends the panel, in which case unit-specific linear trends have no effect on the 2x2 point estimate.

The weights closely resemble the weights in the main DD decomposition theorem, but also incorporate how well linear trends fit the treatment variable in each pair via the $\left(1 - R^2_{jU}\right)$. Trends thus change a DD estimate in two ways. First, they change the value of each 2x2 component by netting out a trend. Second, they alter the weights placed on each two-group comparison. Because it is a function of the variance of the treatment dummy, $R^2_{jU} = 3\frac{T}{T+1}\frac{T}{T-1}V_j$, will be highest for groups treated toward the middle of the panel. Trends thus downweight these terms relative to an unadjusted estimator.

C. Group-Specific Linear Pre-Trends

A simple strategy to address counterfactual trends is to estimate *pre*-treatment trends in Y^0_k directly and partial them out of the full panel (cf. Bhuller et al. 2013, Goodman-Bacon 2018). This is what we hope that unit-specific trends do, it does not depend on the treatment effect pattern, and it does not change the weighting of the 2x2 DD's. Specifically, using data from before t^*_1, one can estimate a pre-trend in y_{it} for each timing group.[6] The slope will equal the linear component of unobservables plus a linear approximation to trend deviations before before t^*_1. Removing this trend from the full panel yields an outcome variable that is robust to linear trends, unaffected by time-varying treatment effects, and weights each 2x2 DD component in the same way as the

[6] One could estimate pre-trends for each *unit*, but this yields identical point estimates to the group-specific pre-trends. Moreover, group-specific trends reduce the variability of the estimator because trend deviations that would bias *unit* specific pre-trends cancel out to some extent when averaged by timing group. This matters because this strategy extrapolates potentially many periods into the future, magnifiying specification error in the pre-trends. This point applies to unit-specific linear trend specifications as well. Point estimates are identical when this specification includes linear trends for each group rather than each unit.

unadjusted estimator. Like the unit-specific time trend control strategy, group-specific pre-trends are sensitive to non-linear unobservables in the pre-period. While inference is outside the scope of this paper (see Athey and Imbens 2018), note that this two-step strategy necessarily involves a partly estimated outcome variable so second-stage standard errors are incorrect.

Using data from before t_1^*, estimate a pre-trend in y_{it} for each timing group.[7] The potential outcome model in (B37) shows that this slope will equal the linear component of unobservables plus a linear approximation to trend deviations before before t_1^*:

$$\frac{\widehat{cov}(\bar{y}_{kt}, t-\bar{t}|t<t_1^*)}{\widehat{var}(t-\bar{t}|t<t_1^*)} = a^k + \frac{\overbrace{\widehat{cov}\left(\overline{dY}_{kt}^0, t-\bar{t}|t<t_1^*,\right)}^{da^k}}{\widehat{var}(t-\bar{t}|t<t_1^*)} \qquad (D22)$$

The fitted pre-trends equal the linear component of unobservables, a^k, plus any non-linear deviations that vary systematically with time in the pre-period, da^k. Removing this trend from the full panel means that the outcome variable will be:

$$\bar{y}_{kt}^* = -da^k(t-\bar{t}) + \overline{dY}_{kt}^0 + D_{kt}ATT_k(t) \qquad (D23)$$

Where $\overline{da}^k$ is the amount by which group k's pre-trend estimate deviates from the true linear component, a^k. The two-group estimate then equals:

$$\hat{\beta}_{kU}^{2x2,pretrend} = ATT_k(POST(k)) - \frac{\left(\overline{da}^k - \overline{da}^U\right)T}{2} + \Delta_{kU}dY^0\big(POST(k), PRE(k)\big) \quad (D24)$$

Bias from non-linear unobservables remains because a linear detrending strategy does not attempt to control for it. (Controlling for unit-specific linear trends suffers from this source of bias as well, so partialling out pre-trends is not obviously any worse in this regard.) Bias from linear pre-trends, however, is addressed. Unless pre-treatment nonlinearities lead to the wrong slope estimates

[7] This detrending, and actually the strategy of adding linear trends as a control, only needs to happen at the group level, not the individual level. Point estimates are the same while standard errors are generally smaller, especially errors from misspecification of the trend during the extrapolated period which well tend to be smaller for groups of units compared to individual units.

(which becomes less likely with more pre-tratement periods), $\frac{(\overline{da}^k - \overline{da}^U)\,T}{2}$ should be smaller than $\frac{(a^k - a^U)\,T}{2}$.

D. Disaggregated Time Fixed Effects

If counterfactual outcomes evolve differently by a category, R, to which units belong, one can model these changes flexibly by including separate time fixed effects for each category:

$$y_{it} = \alpha_i + \alpha_t^{R(i)} + \beta_{R\times t}^{DD} D_{it} + e_{it} \qquad (D25)$$

The coefficient $\widehat{\beta}_{R\times t}^{DD}$ equals an average of two-way fixed effects estimates by values of R weighted by the share of units in each R (n^R) and the within-R variance of $\widetilde{D}_{it}$. For simplicity, I refer to R as "region", but the analysis is not specific to region-by-time fixed effects. When treatments vary by county or city, for example, studies include state-by-time fixed effects (e.g. Almond, Hoynes, and Schanzenbach 2011, Bailey and Goodman-Bacon 2015); when treatments vary by firm studies include industry-by-year fixed effects (e.g. Kovak, Oldenski, and Sly 2018).

Since the decomposition theorem holds for each within-region DD estimate, it also holds for $\widehat{\beta}_{R\times t}^{DD}$. Each 2x2 DD averages the region-specific 2x2 DD's ($\widehat{\beta}_{kU,R}^{2x2}$ and $\widehat{\beta}_{k\ell,R}^{2x2}$) but the weights reflect region size and the within-region distribution of timing groups:

$$\widehat{\beta}_{kU,R\times t}^{2x2} = \sum_R \frac{n^R n_k^R n_U^R}{\sum_R n^R n_k^R n_U^R} \widehat{\beta}_{kU,R}^{2x2} \qquad (D26)$$

$$\widehat{\beta}_{k\ell,R\times t}^{2x2} = \sum_R \frac{n^R n_k^R n_\ell^R}{\sum_R n^R n_k^R n_\ell^R} \left[\mu_{k\ell}\, \widehat{\beta}_{k\ell,R\times t}^{2x2,k} + (1-\mu_{k\ell})\, \widehat{\beta}_{k\ell,R\times t}^{2x2,\ell}\right] \qquad (D27)$$

2x2 DD's from large regions get more weight (via the n^R), but the distribution of timing groups within region (the $n_a^R n_b^R$ terms) also determine the importance of each region. Regions with no units in group k contribute nothing to 2x2 DD's involving that group, in contrast to the simpler

estimator where that region would contribute controls for group k. If no region contains units from a given pair of groups that pair drops out of the disaggregated time effects specification. These within-region 2x2 DD's in (D26) and (D27) are weighted together using the cross-region average of the sample share products, $\sum_R n^R n_k^R n_\ell^R$ as well as treatment variances.

Adding region-by-year effects to the unilateral divorce analysis cuts the estimated effect by a factor of three (-1.16). Figure 11 plots the 2x2 DD's and the weights from this specification against those from the baseline results. 43 of the 156 2x2 terms in the baseline regression drop out of the within-region specification, and table 2 shows that about three quarters of the difference in the estimates comes from the way these fixed effects change the weights.

With disaggregated time fixed effects, the only difference is that there are now R-specific year effects taken out, so the covariance equals:

$$\widehat{cov}\left(\widetilde{y}_{it}, \widetilde{D}_{it}\right) = \sum_R \frac{N_R}{N} [\frac{1}{N_R T} \sum_{i \in R} \sum_t \left(y_{it} - \overline{\overline{y}}_R\right)\left(D_{it} - \overline{\overline{D}}_R\right) - \frac{1}{N_R} \sum_{i \in R} \left(\overline{y}_i - \overline{\overline{y}}_R\right)\left(\overline{D}_i - \overline{\overline{D}}_R\right)$$

$$- \frac{1}{T} \sum_t \left(\overline{y}_t^R - \overline{\overline{y}}_R\right)\left(\overline{D}_t^R - \overline{\overline{D}}_R\right)] \qquad (D28)$$

The simplification does not change, but now it happens within each R. The first two terms are:

$$\sum_R n^R \left[\sum_k n_k^R \overline{D}_k \left(1 - \overline{D}_k\right)\left(\overline{y}_{k,R}^{POST(k)} - \overline{y}_{k,R}^{PRE(k)}\right)\right] \qquad (D29)$$

And the third term is:

$$\sum_R n^R \left[\sum_k \sum_{\ell \neq U} n_\ell^R n_k^R \overline{D}_\ell \left(1 - \overline{D}_\ell\right) \Delta(y_{\ell,R}, t_k^*)\right] \qquad (D30)$$

Combining equations (D2) and (D3) shows that the covariance equals:

$$\sum_R n^R \left[\sum_{k \neq U} n_k^R \overline{D}_k (1 - \overline{D}_k) \Delta(y_{k,R}, t_k^*) - \sum_\ell \sum_{k \neq C} n_\ell^R n_k^R \overline{D}_k (1 - \overline{D}_k) \Delta(y_{\ell,R}, t_k^*)\right] \qquad (D31)$$

Following the derivation in appendix A, we know the covariance within each region essentially follows the decomposition theorem and that the whole estimate is divided by $\hat{V}(\widetilde{D}_{it}^{/R})$, the variance of D conditional on unit and region-by-year fixed effects:

$$\sum_R \frac{n^R}{\hat{V}(\widetilde{D}_{it}^{/R})}\Big[\sum_{k\neq U} \overline{D}_k(1-\overline{D}_k)n_k^R n_U^R\, \widehat{\beta}^{2x2}_{kU,R} + \sum_{k\neq U}\sum_{\ell>k}(\overline{D}_k-\overline{D}_\ell)\big(1-(\overline{D}_k-\overline{D}_\ell)\big)n_k^R n_\ell^R\left[\mu_{k\ell}\,\widehat{\beta}^{2x2,k}_{k\ell,R} + (1-\mu_{k\ell})\,\widehat{\beta}^{2x2,\ell}_{k\ell,R}\right] \tag{D32}$$

Consider the weight on the terms that compare group k to group U within each region:

$$\frac{\overline{D}_k(1-\overline{D}_k)}{\hat{V}(\widetilde{D}_{it}^{/R})}\sum_R n^R n_k^R n_U^R\,\widehat{\beta}^{2x2}_{kU,R} = \frac{\overline{D}_k(1-\overline{D}_k)\sum_R n^R n_k^R n_U^R}{\hat{V}(\widetilde{D}_{it}^{/R})}\sum_R \frac{n^R n_k^R n_U^R}{\sum_R n^R n_k^R n_U^R}\widehat{\beta}^{2x2}_{kU,R}$$

$$= \overbrace{\frac{\overline{D}_k(1-\overline{D}_k)\sum_R n^R n_k^R n_U^R}{\hat{V}(\widetilde{D}_{it}^{/R})}}^{s_{kU}^{R\times t}} \overbrace{\sum_R \rho_{kU}^R\,\widehat{\beta}^{2x2}_{kU,R}}^{\widehat{\beta}^{2x2}_{kU,R\times t}} \tag{D33}$$

The same type of expression holds for the other two-group terms. The analogy of equation (7) for the disaggregated fixed effects specification is:

$$\widehat{\beta}^{DD}_{R\times t} = \sum_{k\neq U} s_{kU}^{R\times t}\,\widehat{\beta}^{2x2,R\times t}_{kU} + \sum_{k\neq U}\sum_{\ell>k} s_{k\ell}^{R\times t}\left[\mu_{k\ell}\,\widehat{\beta}^{2x2,k}_{k\ell,R\times t} + (1-\mu_{k\ell})\,\widehat{\beta}^{2x2,\ell}_{k\ell,R\times t}\right] \tag{D34}$$

APPENDIX E: TREATMENTS THAT TURN OFF

Many treatments turn on *and* off. To characterize how these models work, write the treatment dummy as the difference between a dummy for whether the treatment has ever turned on and a dummy for whether the treatment has ever turned off: $W_{it} = D_{it}^{ON} - D_{it}^{OFF}$. The fixed-effects adjusted version of W_{it} is just $\widetilde{D}_{it}^{ON} - \widetilde{D}_{it}^{OFF}$, so the DD estimate equals:

$$\hat{\beta}_{ON,OFF}^{DD} = \frac{\widehat{cov}\left(\widetilde{y}_{it}, \widetilde{D}_{it}^{ON}\right) - \widehat{cov}\left(\widetilde{y}_{it}, \widetilde{D}_{it}^{OFF}\right)}{\widehat{var}\left(\widetilde{W}_{it}\right)} \qquad (E1)$$

The fact that the decomposition theorem applies to both terms in the numerator shows that DD estimates combine estimated effects when treatments turn on and when they turn off. To show how these comparisons are made, I consider a 2x2 comparison between a group, j, whose treatment turns on and off and another, U, that never receives treatment. Both covariances have the DD form, and are multiplied by the product of the sample shares and the treatment variance:

$$n_j n_U \bar{D}_j^{ON}\left(1 - \bar{D}_j^{ON}\right)\left[\left(\overline{y}_j^{POST-ON(j)} - \overline{y}_j^{PRE(j)}\right) - \left(\overline{y}_U^{POST-ON(j)} - \overline{y}_U^{PRE(j)}\right)\right] +$$

$$n_j n_U \bar{D}_j^{OFF}\left(1 - \bar{D}_j^{OFF}\right)\left[\left(\overline{y}_j^{OFF(j)} - \overline{y}_j^{PRE-OFF(j)}\right) - \left(\overline{y}_U^{OFF(j)} - \overline{y}_U^{PRE-OFF(j)}\right)\right] \qquad (E2)$$

The time periods used in these comparisons overlap, though. A portion $\frac{\bar{D}_j^{ON} - \bar{D}_j^{OFF}}{\bar{D}_j^{ON}}$ the period when $D_{it}^{ON} = 1$ contains the "middle window" when $W_{it} = D_{it}^{ON} - D_{it}^{OFF} = 1$ and a portion $\frac{\bar{D}_j^{OFF}}{\bar{D}_j^{ON}}$ contains the period after treatment turns off, $W_{it} = 1 - 1 = 0$. Similarly, the period when $D_{it}^{OFF} = 0$ contains the same "middle window" and the period before treatment starts and $W_{it} = 0 - 0 = 0$. This matches the two-group timing-only estimator, and equation (E2) can be written in the same way:

$$\hat{\beta}^{2x2,kU}_{ON,OFF} = \mu^{*}_{jU} \overbrace{\left[\left(\overline{y}^{ON(j)}_{j} - \overline{y}^{PRE(j)}_{j}\right) - \left(\overline{y}^{ON(j)}_{U} - \overline{y}^{PRE(j)}_{U}\right)\right]}^{\hat{\beta}^{2x2,ON}_{jU}} +$$

$$\left(1 - \mu^{*}_{jU}\right) \overbrace{\left[\left(\overline{y}^{ON(j)}_{j} - \overline{y}^{OFF(j)}_{j}\right) - \left(\overline{y}^{ON(j)}_{U} - \overline{y}^{OFF(j)}_{U}\right)\right]}^{\hat{\beta}^{2x2,OFF}_{jU}} \quad (E3)$$

$$\mu^{*}_{jU} = \frac{1 - \overline{D}^{ON}_{j}}{1 - \left(\overline{D}^{ON}_{j} - \overline{D}^{OFF}_{j}\right)} \quad (E4)$$

Equations (E3) and (E4) show that a DD model with a treatment that turns off averages the 2x2 DD that compares the middle window to the pre-treatment period and the 2x2 DD that compares the middle window to the period when treatment has turned off. The weights come from how close to the middle of the panel is the treatment's start date versus its end date. The analogous version with two treated groups would be more complicated and depend on the overlap between the periods when each unit's treatment has not yet turned on, is on, or has already turned off.

Note that while $\hat{\beta}^{2x2,ON}_{jU}$ is a typical 2x2 estimator, $\hat{\beta}^{2x2,OFF}_{jU}$ compares outcomes in the period when treatment is turned on to a period when treatment is turned off but, by definition, has been on in the past. If treatment effects persist after treatment stops, we do not observe the counterfactual outcome in the $OFF(j)$ period. Outcomes may not actually change when treatment turns off, making $\hat{\beta}^{2x2,OFF}_{jU}$ and therefore the overall DD estimate too small. A version of the DD decomposition theorem and plots like figure 4 could be used to analyze the extent to which the estimates based on treatments turning off differ systematically from estimates based on treatments turning on.

APPENDIX F. ALTERNATIVE DECOMPOSITIONS

This section discusses the connection between Theorem 1 and other decompositions of either the estimator or estimand that have been proposed in the literature.

A. A very special case

In one knife-edge case the TWFE estimator equals an average of 2x2 DD's that only use untreated units and periods as control groups and that only estimate the treatment effect in the first post-treatment period. Consider the case with three periods (1, 2, and 3) and three groups: E (treated in time 2), L (treated in time 3), and U (never treated). Assume equal group sizes so that all the cross-sectional shares, $n_E = n_L = n_U$, drop out. Theorem 1 shows that:

$$\frac{1}{\widehat{V}^D}\left[\overline{D}_E(1-\overline{D}_E)\,\widehat{\beta}_{EU}^{2x2} + \overline{D}_L(1-\overline{D}_L)\,\widehat{\beta}_{EU}^{2x2} + (\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)\,\widehat{\beta}_{EL}^{2x2,E} + \overline{D}_L(\overline{D}_E-\overline{D}_L)\,\widehat{\beta}_{EL}^{2x2,L}\right]$$

Rewrite the full DD decomposition in terms of mean outcomes in every period using $\bar{y}_j^{POST(E)} = \frac{\overline{D}_L}{\overline{D}_E}\bar{y}_j^{POST(L)} + \frac{(\overline{D}_E-\overline{D}_L)}{\overline{D}_E}\bar{y}_j^{MID(E,L)}$ and $\bar{y}_j^{PRE(L)} = \frac{1-\overline{D}_E}{1-\overline{D}_L}\bar{y}_j^{PRE(E)} + \frac{(\overline{D}_E-\overline{D}_L)}{1-\overline{D}_L}\bar{y}_j^{MID(E,L)}$:

$$\begin{aligned}
\frac{1}{\widehat{V}^D}\Bigg[&\overline{D}_E(1-\overline{D}_E)\Bigg[\left(\frac{\overline{D}_L}{\overline{D}_E}\bar{y}_E^{POST(L)} + \frac{(\overline{D}_E-\overline{D}_L)}{\overline{D}_E}\bar{y}_E^{MID(E,L)} - \bar{y}_E^{PRE(E)}\right)\\
&\quad - \left(\frac{\overline{D}_L}{\overline{D}_E}\bar{y}_U^{POST(L)} + \frac{(\overline{D}_E-\overline{D}_L)}{\overline{D}_E}\bar{y}_U^{MID(E,L)} - \bar{y}_U^{PRE(E)}\right)\Bigg]\\
&+ \overline{D}_L(1-\overline{D}_L)\Bigg[\left(\bar{y}_L^{POST(L)} - \frac{1-\overline{D}_E}{1-\overline{D}_L}\bar{y}_L^{PRE(E)} - \frac{(\overline{D}_E-\overline{D}_L)}{1-\overline{D}_L}\bar{y}_L^{MID(E,L)}\right)\\
&\quad - \left(\bar{y}_U^{POST(L)} - \frac{1-\overline{D}_E}{1-\overline{D}_L}\bar{y}_U^{PRE(E)} - \frac{(\overline{D}_E-\overline{D}_L)}{1-\overline{D}_L}\bar{y}_U^{MID(E,L)}\right)\Bigg]\\
&+ (\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)\left[\left(\overline{y}_E^{MID(E,L)} - \overline{y}_E^{PRE(E)}\right) - \left(\overline{y}_L^{MID(E,L)} - \overline{y}_L^{PRE(E)}\right)\right]\\
&+ \overline{D}_L(\overline{D}_E-\overline{D}_L)\left[\left(\overline{y}_L^{POST(L)} - \overline{y}_L^{MID(E,L)}\right) - \left(\overline{y}_E^{POST(L)} - \overline{y}_E^{MID(E,L)}\right)\right]\Bigg]
\end{aligned}$$

The coefficient on each $\overline{y}$ (inside the square brackets) after grouping terms is shown below as well as its value after substituting in the fact that $1-\overline{D}_E=\overline{D}_L=\frac{1}{3}$:

$$\left(\overline{D}_L(1-\overline{D}_E)-(\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)\right)\bar{y}_E^{POST(L)}=0$$

$$\left((\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)+\ (\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)+\overline{D}_L(\overline{D}_E-\overline{D}_L)\right)\bar{y}_E^{MID(E,L)}=\frac{1}{3}\bar{y}_E^{MID(E,L)}$$

$$-\left(\overline{D}_E(1-\overline{D}_E)+(\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)\right)\bar{y}_E^{PRE(E)}=\ -\frac{1}{3}\bar{y}_E^{PRE(E)}$$

$$-\left((1-\overline{D}_E)\overline{D}_L+\ \overline{D}_L(1-\overline{D}_L)\right)\bar{y}_U^{POST(L)}=-\frac{1}{3}\bar{y}_U^{POST(L)}$$

$$-\left((1-\overline{D}_E)(\overline{D}_E-\overline{D}_L)-\overline{D}_L(\overline{D}_E-\overline{D}_L)\right)\bar{y}_U^{MID(E,L)}=0$$

$$\left(\overline{D}_E(1-\overline{D}_E)+\overline{D}_L(1-\overline{D}_E)\right)\bar{y}_U^{PRE(E)}=\frac{1}{3}\bar{y}_U^{PRE(E)}$$

$$\left(\overline{D}_L(1-\overline{D}_L)+\ \overline{D}_L(\overline{D}_E-\overline{D}_L)\right)\bar{y}_L^{POST(L)}=\frac{1}{3}\ \bar{y}_L^{POST(L)}$$

$$-\left(\overline{D}_L(\overline{D}_E-\overline{D}_L)+\overline{D}_L(\overline{D}_E-\overline{D}_L)+(\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)\right)\bar{y}_L^{MID(E,L)}=-\frac{1}{3}\ \bar{y}_L^{MID(E,L)}$$

$$\left((\overline{D}_E-\overline{D}_L)(1-\overline{D}_E)-\overline{D}_L(1-\overline{D}_E)\right)\overline{y}_L^{PRE(E)}=\underbrace{\frac{1}{3}\overline{y}_L^{PRE(E)}-\frac{1}{3}\ \overline{y}_L^{PRE(E)}}_{0}$$

The DD estimator equals:

$$\frac{1}{\hat{V}^D}\frac{1}{3}\left\{\left[\left(\bar{y}_E^{MID(E,L)}-\bar{y}_E^{PRE(E)}\right)-\left(\bar{y}_L^{MID(E,L)}-\overline{y}_L^{PRE(E)}\right)\right]\right.$$
$$\left.+\left[\left(\bar{y}_L^{POST(L)}-\overline{y}_L^{PRE(E)}\right)-\left(\bar{y}_U^{POST(L)}-\bar{y}_U^{PRE(E)}\right)\right]\right\}$$

This is an equally weighted average of one 2x2 comparing group E to group L in periods 1 and 3 and another comparing group L to group U in periods 3 and 1. Neither DD uses an already-treated control group. Under a common trends assumption (but no assumption on heterogeneous effects) they both capture the treatment effect in the first period after treatment. But this is not a general

feature of the DD decomposition, it is a special case that *only* arises when group shares are equal and treatment shares balance in this exact way. There is no alternative way to decompose $\hat{\beta}^{DD}$ so that it converges to a positively weighted average of treatment effects under a common trends assumption only. (The fact that $\hat{\beta}^{DD}$ can have the opposite sign of all the $ATT(g,t)$ also shows that there is no general decomposition that yields a positive weighted average of treatment effects.)

B. Average of 2x2s by individual unit and time period

Proposition 1 in Strezhnev (2018) decomposes $\hat{\beta}^{DD}$ into an unweighted average of DD comparisons between two individual units (rather than timing groups) in two individual time periods (instead of pre- and post- windows defined by the two units' treatment timing). Using A_{it} to denote the treatment indicator, his result is:

$$\widehat{\beta}^{DD} = \frac{\sum_t \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{[Y_{it} - Y_{it'}] - [Y_{jt} - Y_{jt'}]\}}{\sum_t \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{1 - (A_{it'} - A_{jt'})\}}$$

This is a disaggregated version of the DD Decomposition Theorem.

Consider a timing-only case as in equation (11) with three periods and two equally sized groups treated at time 2 (group E) and time 3 (group L). The DD Decomposition theorem shows that $\widehat{\beta}^{DD}$ is an equally weighted average of $\widehat{\beta}^{2x2,E}_{EL}$ and $\widehat{\beta}^{2x2,L}_{EL}$:

$$\widehat{\beta}^{DD} = \frac{\overbrace{(\overline{D}_E - \overline{D}_L)(1 - \overline{D}_E)}^{\frac{1}{2}}}{\widehat{V}^D}[(\overline{y}^2_E - \overline{y}^1_E) - (\overline{y}^2_L - \overline{y}^1_L)] + \frac{\overbrace{(\overline{D}_E - \overline{D}_L)\overline{D}_L}^{\frac{1}{2}}}{\widehat{V}^D}[(\overline{y}^3_L - \overline{y}^2_L) - (\overline{y}^3_E - \overline{y}^2_E)]$$

Equal weighting follows from the fact that $(1 - \overline{D}_E) = \overline{D}_L$.

The numerator in Proposition 1 in Strezhnev (2018) is:

$$\widehat{\beta}^{DD} = \sum_{i:A_{i2}=1} \sum_{j:A_{j2}=0} \sum_{t' \neq 2} \{[Y_{E2} - Y_{Et'}] - [Y_{L2} - Y_{Lt'}]\}$$

The only observations for which $A_{i2} = 1$ are those in group E and for which $A_{j2} = 0$ are those in group L. There are no units with $A_{i3} = 0$.

$$= \sum_{i \in E} \sum_{j \in L} \left(\{ [Y_{i2} - Y_{i1}] - [Y_{j2} - Y_{j1}] \} - \{ [Y_{j3} - Y_{j2}] - [Y_{i3} - Y_{i2}] \} \right)$$

$$= N_E N_L (\{ [\bar{Y}_{E2} - \bar{Y}_{E1}] - [\bar{Y}_{L2} - Y_{L1}] \} + \{ [\bar{Y}_{E2} - \bar{Y}_{E3}] - [\bar{Y}_{L2} - \bar{Y}_{L3}] \})$$

$$= N_E N_L \left(\overbrace{\{ [\bar{Y}_{E2} - \bar{Y}_{E1}] - [\bar{Y}_{L2} - Y_{L1}] \}}^{\hat{\beta}_{EL}^{2x2,E}} + \overbrace{\{ [\bar{Y}_{L3} - \bar{Y}_{L2}] - [\bar{Y}_{E3} - \bar{Y}_{E2}] \}}^{\hat{\beta}_{EL}^{2x2,L}} \right)$$

The denominator is:

$$\sum_{i \in E} \sum_{j \in L} \left(1 - \overbrace{(A_{i1} - A_{j1})}^{0} + 1 - \overbrace{(A_{i3} - A_{j3})}^{0} \right) = 2 N_E N_L$$

So Strezhnev (2018) also implies that $\hat{\beta}^{DD} = \frac{1}{2} \hat{\beta}_{EL}^{2x2,E} + \frac{1}{2} \hat{\beta}_{EL}^{2x2,L}$. Which shows that the two decompositions are equivalent.

I argue that there are several advantages of a decomposition in terms of timing-group means and groups of periods. It makes the weighting explicit and intuitive, as I describe. In the more disaggregated decomposition the weights on the group-level 2x2 DDs are implicit. Groups are the more natural level at which to explain $\hat{\beta}^{DD}$ because they represent the identifying variation: whether and when collections of units are treated. Finally, Theorem 1 uses 2x2 DDs that match the canonical definition: the treatment group moves from untreated to treated and the control group's treatment status remains unchanged. This sometimes means that already-treated units act as the control group. Strezhnev's decomposition describes this using the idea of "future matches" in which two units whose treatment status differ at one date are "compared to" each other at a future date when both treated. This switches both the temporal order of the comparison (earlier

periods are compared to later periods) and the definition of the treatment group (the unit treated earlier is said to be the treatment and the unit untreated earlier is said to be the control).

C. Weights on individual treatment effect parameters in the estimand

de Chaisemartin and D'HaultfŒuille (forthcoming) present a decomposition of the TWFE estimand in terms of each $ATT(g,t)$ and conclude that the estimator "is more likely to assign a negative weight to periods where a large fraction of groups are treated, and to groups treated for many periods."[8] They describe the weights in terms of the residual from a regression that partials fixed effects out of the treatment dummy, not in terms of sample composition and treatment variance as in Theorem 1. On the other hand, equations (15a) and (15c) characterize the TWFE estimand in terms of averages of $ATT(g,t)$ over different post-treatment windows but with easily interpretable weights. This section shows reconciles the two approaches.

To see the connection between these results consider the three places that a given group-time effect, $ATT(j,\tau)$ (where $\tau > j$) appears in equations (15a) and (15c) and what weight it receives:

1. It is part of $ATT(j, POST(j))$ identified by the 2x2 comparison between group j and group U and in the 2x2 comparisons between group j and each already-treated control group k. Following equation (12), $ATT(j,\tau)$ accounts for a share $\frac{1}{T-(j-1)}$ of $ATT(j, POST(j))$ and each term that identifies gets a weight defined in Theorem 1. Therefore, the total weight from these terms is $\frac{1}{T-(j-1)}\left[\sigma_{jU} + \sum_{k<j} \sigma_{kj}^{j}\right]$.
2. It is also part of $ATT(j, MID(j,\ell))$ in comparisons between group j and each future-treated control group ℓ as long as ℓ is far enough away from j to include effects in period τ (that

[8] Borusyak and Jaravel (2017) conclude the same thing.

is, $\ell > \tau$). In these cases $ATT(j,\tau)$ accounts for a share $\frac{1}{\ell-j}$ of $ATT(j, MID(j,\ell))$, and the total weight on these terms is: $\sum_{\ell>\tau} \frac{1}{\ell-j}\sigma_{j\ell}^{j}$.

3. Most importantly, $ATT(j,\tau)$ appears as heterogeneity bias in ΔATT. For $\hat{\beta}_{j\ell}^{2\times 2,\ell}$, the heterogeneity bias is $ATT\big(j, POST(\ell)\big) - ATT\big(j, MID(j,\ell)\big) = \frac{1}{T-(j-1)}\sum_{t\geq\ell} ATT(j,t) - \frac{1}{\ell-j}\sum_{t\geq j}^{\ell-1} ATT(j,t)$. The group-time effect at time τ appears in the first sum if $\tau \geq \ell$ and the second if $\tau < \ell$. Therefore the weight on all $ATT(j,\tau)$ across all heterogeneity bias terms is: $-\left[\sum_{\substack{j<\ell \\ \ell\leq\tau}} \frac{\sigma_{j\ell}^{\ell}}{T-(\ell-1)} - \sum_{\ell>\tau}\frac{\sigma_{j\ell}^{\ell}}{\ell-j}\right]$.

Collecting all the terms that multiply $ATT(j,t)$ maps the definition of the estimand from equation (15) where the weights are easily interpretable functions of sample sizes and treatment variances, to the weighted average in terms of easily interpretable group-time average effects. Specifically, in the notation of de Chaisemartin and D'HaultfŒuille (forthcoming) the weight on $ATT(j,t)$, denoted $\frac{N_{j,\tau}}{N_1} w_{j,\tau}$, is:

$$\frac{N_{j,\tau}}{N_1} w_{j,\tau} = \frac{\sigma_{jU} + \sum_{k<j}\sigma_{kj}^{j}}{T-(j-1)} + \sum_{\ell>\tau}\frac{\sigma_{j\ell}^{j} + \sigma_{j\ell}^{\ell}}{\ell - j} - \overbrace{\sum_{\substack{\ell>j \\ \ell\leq\tau}} \frac{\sigma_{j\ell}^{\ell}}{T-(\ell-1)}}^{source\ of\ negative\ weight}$$

For a given treatment group j, weights are smaller and possibly negative as τ gets bigger. A larger τ, meaning a longer-run treatment effect parameter, means fewer terms added in the second sum and more terms subtracted in the third sum. Later treated groups, denoted by a larger j, are less likely to have negative weight because there are more terms added in the first sum and fewer terms are subtracted in the third term.

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 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. 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.

Bhuller, Manudeep, Tarjei Havnes, Edwin Leuven, and Magne Mogstad. 2013. "Broadband Internet: An Information Superhighway to Sex Crime?" *The Review of Economic Studies* 80 (4 (285)):1237-1266.

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

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

Goodman-Bacon, Andrew. 2018. "The Long-Run Effects of Childhood Insurance Coverage: Medicaid Implementation, Adult Health, and Labor Market Outcomes." *Working Paper*.

Hendren, Nathaniel. 2019. "Efficient Welfare Weights." *Working Paper*.

Kovak, Brian, Lindsay Oldenski, and Nicholas Sly. 2018. "The Labor Market Effects of Offshoring by U.S. Multinational Firms: Evidence from Changes in Global Tax Policies." *Working Paper*.

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.

Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." *Working Paper*.