Difference-in-Differences with Variation in Treatment Timing, Journal of Econometrics, 2021

Wenzhi Wang [*]

August 20, 2024

1. Introduction

A Difference-in-Differences estimate is the difference between the change in outcomes before and after a treatment (difference one) in a treatment versus control group (difference two):

$$\left(\overline{y}^{\text{post}}_{\text{treat}} - \overline{y}^{\text{pre}}_{\text{treat}}\right) - \left(\overline{y}^{\text{post}}_{\text{control}} - \overline{y}^{\text{post}}_{\text{control}}\right).$$

That simple quantity also equals the estimated coefficient on the interaction of a treatment group dummy and a post-treatment period dummy in the following regression:

$$y_{it} = \gamma + \gamma_{i\cdot}\text{Treat}_i + \gamma_{\cdot t}\text{Post}_t + \beta^{2\times 2}\text{Treat}_i \times \text{Post}_t + u_{it}. \tag{1}$$

Most DiD applications diverge from this 2×2 set up though because treatments usually occur at different times. In this case researchers estimate a regression with dummies for cross-sectional units, $\alpha_{i\cdot}$, time periods, $\alpha_{\cdot t}$, and a treatment dummy, D_{it}:

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

In contrast to our substantial understanding of canonical 2×2 DD, we know relatively little about the two-way fixed effects DD when treatment timing varies. *In particular, we do not know precisely how it compares mean outcomes across groups.*

This paper shows that the two-way fixed effects DiD estimator in (2) is a weighted average of all possible 2×2 DiD estimators that compare timing groups to each other (the DiD decomposition). Some use units treated at a particular time as the treatment group adn untreated units as the control group. Some compare units treated at two different times, using the later-treated group as a control before

[*] This note is written in my pre-doc period at the University of Chicago Booth School of Business.

its treatment begins and then the earlier-treated group as a control after its treatment begins. The weights on the 2×2 DDs are proportional to timing group sizes and the variance of the treatment dummy in each pair, which is highest for units treated in the middle of the panel.

First, I use this DD decomposition to show that TWFEDD estimates a variance-weighted average of treatment effect parameters sometimes with "negative weights". When treatment effects do not change over time, TWFEDD yields a variance-weighted average of cross-group treatment effects and all weights are positive. Negative weights only arise when average treatment effects vary over time. The DD decomposition shows why: when already-treated units act as controls, changes in their outcomes are subtracted and these changes may include time-varying treatment effects.

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

Finally, I develop simple tools to describe the TWFEDD estimator and evaluate why estimators change across specifications. Plotting the 2×2 DDs against their weights displays heterogeneity in the components of the weighted average and shows which terms and timing groups matter most. Summing the weights on the timing comparisons quantifies "how much" of the variation comes from timing (a common question in practice), and provide guidance on how well the TWFEDD estimator works compared to alternative estimators.

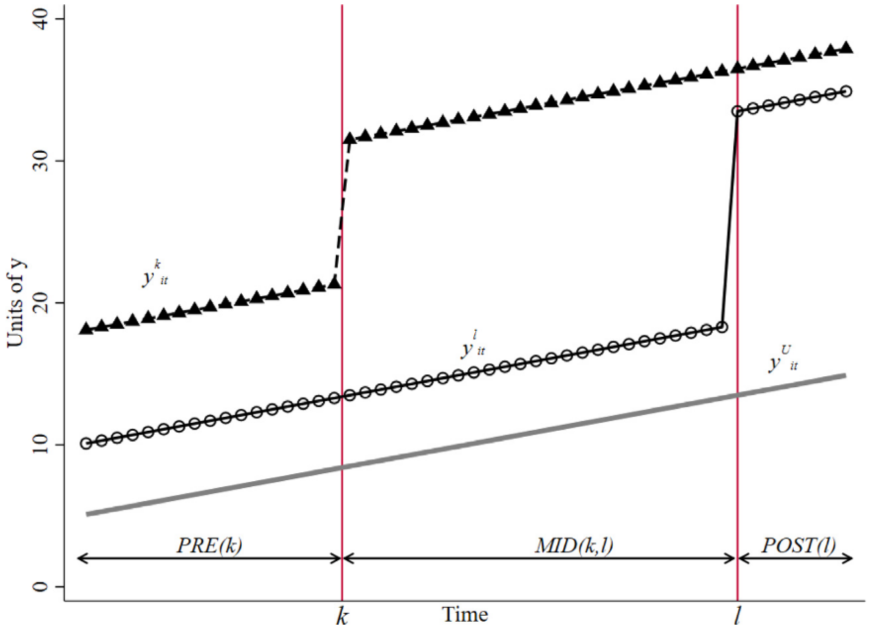
2. The Difference-in-Differences Decomposition Theorem

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

Each timing group's sample share is $n_j = \sum_i \mathbb{I}(t_i = j)/N, j \in \{k, \ell, U\}$ and the share of time it spends treated is $\overline{D}_j = \sum_t \mathbb{I}(t \geq j)/T$. I denote the sample mean of y_{it} for units treated at time t_b^* during the post period or treatment day t_a^* by

$$\overline{y}_b^{post(a)} \equiv \frac{1}{T-(a-1)} \sum_{b=a}^{T} \left[\frac{\sum_i y_{it}\, \mathbb{I}(t_i = b)}{\sum_i \mathbb{I}(t_i = b)} \right].$$

Figure 1: Difference-in-Differences with variation in treatment Timing: Three groups.



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

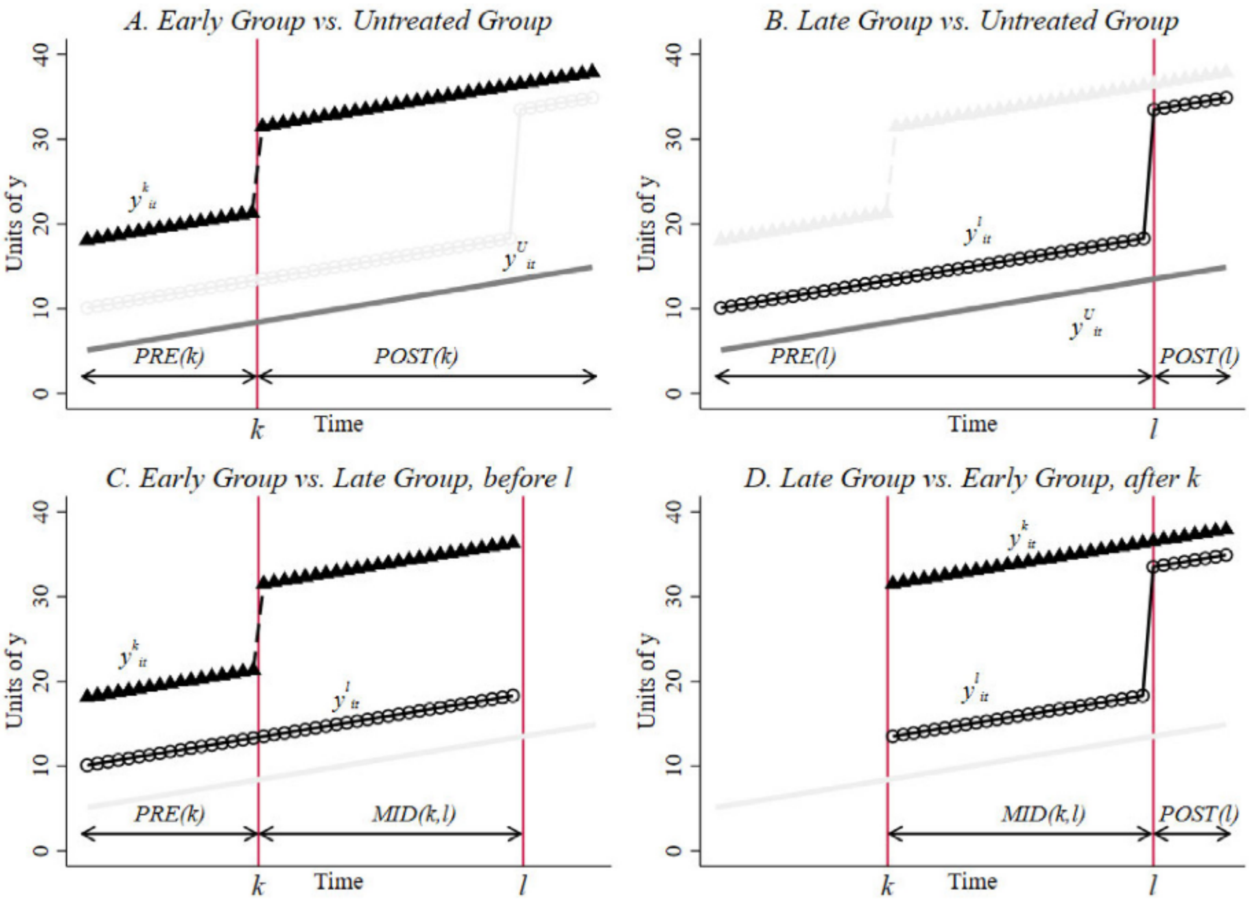
By the Frisch-Waugh-Lovell theorem, $\widehat{\beta}^{DD}$ equals the univariate regression coefficient between y_{it} and the treatment dummy with unit and time means removed:

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

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

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

Figure 2: The four simple (2 × 2) difference-in-differences estimates in the three group case.



A. Early Group vs. Untreated Group

B. Late Group vs. Untreated Group

C. Early Group vs. Late Group, before l

D. Late Group vs. Early Group, after k

Note: The figure plots outcomes for the subsamples that generate the four simple 2 × 2 difference-in-differences estimates in the three timing group case from Figure 1. Each panel plots the data structure for one 2 × 2 DD.

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

$$\widehat{\beta}^{2\times 2}_{jU} \equiv \left(\overline{y}^{post(i)}_{j} - \overline{y}^{pre(i)}_{j}\right) - \left(\overline{y}^{post(j)}_{U} - \overline{y}^{pre(j)}_{U}\right), \quad j = k, \ell. \tag{4}$$

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

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

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

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

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

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

Second, each 2×2 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:

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

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

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

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

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

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

where the 2×2 DD estimators are:

$$\widehat{\beta}^{2\times 2}_{kU} \equiv \left(\overline{y}^{post(k)}_k - \overline{y}^{pre(k)}_k\right) - \left(\overline{y}^{post(k)}_U - \overline{y}^{pre(k)}_U\right), \tag{10b}$$

$$\widehat{\beta}^{2\times 2,k}_{k\ell} \equiv \left(\overline{y}^{mid(k,\ell)}_k - \overline{y}^{pre(k)}_k\right) - \left(\overline{y}^{mid(k,\ell)}_\ell - \overline{y}^{pre(k)}_\ell\right), \tag{10c}$$

$$\widehat{\beta}^{2\times 2,\ell}_{k\ell} \equiv \left(\overline{y}^{post(\ell)}_\ell - \overline{y}^{mid(k,\ell)}_\ell\right) - \left(\overline{y}^{post(\ell)}_k - \overline{y}^{mid(k,\ell)}_k\right). \tag{10d}$$

The weights are:

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

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

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

And

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

Theorem 1 completely describes the sources of identifying variation in a TWFEDD estimator and their importance. With K timing groups, one could form $K^2 - K$ "timing-only" estimates that either compare an earlier- to a later-treated timing group or a later- to earlier-treated timing group. With an untreated group, one could form K 2×2 DDs that compare one timing group to the untreated group. Therefore, with K timing groups and one untreated group, the DD estimator comes from K^2 distinct 2×2 DDs.

The weights on each 2×2 DD combine the absolute size of the subsample and the variance of the fixed-effects-adjusted treatment variable in the subsample. The first part is the size of the subsample squared. The second part of each weight is the subsample variance from Equations (7)-(9), which comes from the relative size of the treatment and control groups and the timing of treatment. The variance is larger when the two timing groups are closer in size ($n_{kU} \approx 0.5$) and when treatment occurs closer to the middle of the time window.

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

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

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

Theorem 1 is not the only way to decompose the TWFEDD estimator. By grouping 2×2 terms according to the identifying variation (pre/post, treatment/control) that unites them, my "group-level" decomposition yields clear definitions for the weights and connects them to the features of the OLS estimation method.

The DD decomposition theorem is different from related results that decompose the TWFEDD estimand into a weighted average of treatment effect parameters with potentially negative weights. Theorem 1 expresses the TWFEDD estimator as a weighted average of simpler estimators with strictly positive weights that sum to 1.

3. Theory: What Parameter Does Dd Identify and Under What Assumptions

Theorem 1 relates the regression DD coefficient to sample averages, which makes it simple to analyze its statistical properties by writing $\widehat{\beta}^{DD}$ in terms of potential outcomes. Define $Y_{it}(k)$ as the outcome of unit i in period t when it is treated at $t_i = k$, and use $Y_{it}(t_i)$ to denote the treated potential outcomes under unit i's actual treatment date. $Y_{it}(0)$ is the untreated potential outcome. If $t < t_i$ then $Y_{it}(t_i) = Y_{it}(0)$. The observed outcome is $y_{it} = D_{it} Y_{it}(t_i) + (1 - D_{it}) Y_{it}(0)$. Define the ATT for timing group k at time $\tau \geq k$ (the "group-time average treatment effect"):

$$ATT_k(\tau) \equiv \mathbb{E}\left[Y_{i\tau}(t_k^*) - Y_{i\tau}(0) \mid t_i = k\right].$$

Because TWFEDD averages outcomes in pre- and post-treatment windows, I define the average $ATT_k(\tau)$ in a date range W (with T_W periods):

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

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

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

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

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

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

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

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

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

$$\text{plim}_{N\rightarrow\infty}\, \widehat{\beta}^{DD} = \beta^{DD} = VWATT + VWCT - \Delta ATT. \tag{15}$$

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

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

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

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

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

Like $VWATT$, $VWCT$ is an average of the difference in counterfactual trends between pairs of timing groups and different time periods using the weights from the decomposition theorem. It captures the way that differential trends map to bias in (10a). Note that one timing group's counterfactual trend affects many 2×2 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 2×2 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 within each timing group's before and after a later treatment time:

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

Because the 2×2 estimators in Equation (14c) already-treated groups as controls, they subtract average changes in their untreated outcomes and their treatment effects. ΔATT equals zero if average treatment effect is constant, but when they are not, Equation (15c) defines the resulting bias relative to $VWATT$ even when $VWCT = 0$. ΔATT is the source of the negative weights discussed in Borusyak et al. (2024) and de Chaisemartin and D'Haultfœuille (2020). This does not mean that the research design is invalid. In this case, specifications such as an event-study, "stacked DD", or reweighting estimators may be more appropriate.

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

3.1. Interpreting the TWFEDD Estimand

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

3.1.1. Effects that vary across units but not over time

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

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

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

In general, $w_k^T \neq n_k^*$, so $VWATT$ does not equal the sample ATT. Neither are the weights proportional to the share of time each unit spends under treatment, so $VWATT$ also does not equal the effect in the average treated period. The extent to which $VWATT$ differs from the ATT depends on the relationship between treatment effect heterogeneity and treatment timing in a given sample.

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

3.1.2. Effects that vary over time but not across units

Time-varying treatment effects shape Equation (15) in two ways. First, they generate heterogeneity across the 2×2 DDs that average over different post-treatment windows and up-weight short-run effects most likely to appear in the small windows between timing groups. These features typically make $VWATT$ different from the sample ATT. Second, time-varying effects bias estimates away from $VWATT$ because $\Delta ATT \neq 0$.

To illstrate this point, Figure 3 plots a case where counterfactual outcomes are identical, but the treatment effect is a linear trend-break, $Y_{it}(k) = Y_{it}(0) + \phi \cdot (t - t_i + 1)$. $\widehat{\beta}_{k\ell}^{2\times2,k}$ uses timing group ℓ as a control group during its pre-period and identifies the ATT during the middle window in which treatment status varies:

$$ATT(mid(k,\ell)) = \phi \cdot \frac{\ell - k + 1}{2}.$$

$\widehat{\beta}_{d\ell}^{2\times2,\ell}$, however, is biased for $ATT(post(\ell))$ because the control group (k) experiences a trend in outcomes due to it growing treatment effect:

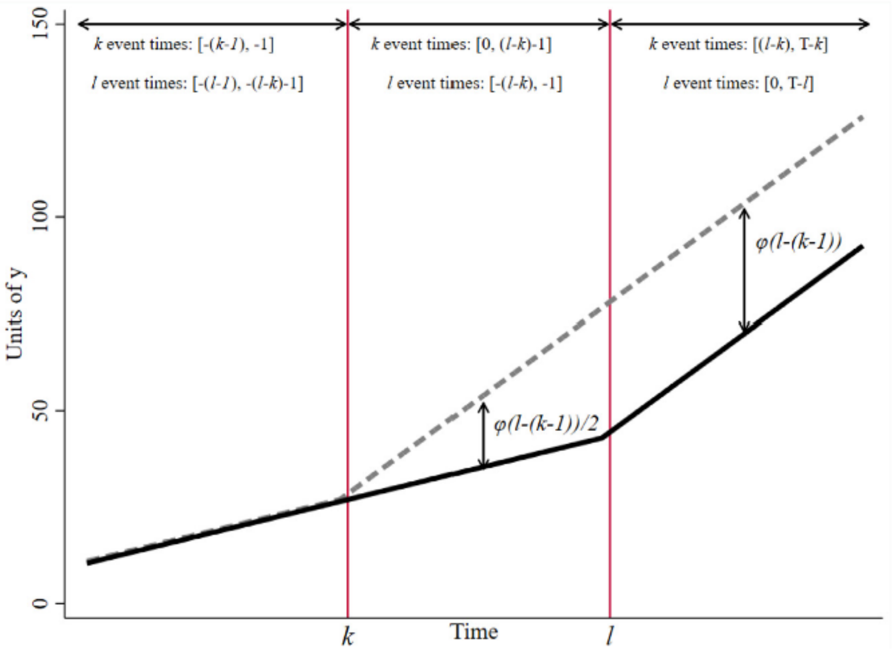
$$\widehat{\beta}_{k\ell}^{2\times2,\ell} = \overbrace{ATT_\ell(post(\ell))}^{\phi\frac{T-\ell+1}{2}} - \overbrace{\phi\frac{T-k+1}{2}}^{\Delta ATT/(1-\mu_{k\ell})} = \phi \cdot \frac{k-\ell}{2} \leq 0. \tag{17}$$

This bias feed though ot $\beta_{k\ell}^{2\times2}$ to the relative weight on the 2×2 terms:

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

The entire two-group timing estimate can be wrong signed if there is sufficiently more weights on $\widehat{\beta}_{k\ell}^{2\times2,\ell}$ than $\widehat{\beta}_{k\ell}^{2\times2,k}$. Summarizing time-varying effects using Regression (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.

Figure 3: Difference-in-Differences estimates with variation in timing are biased when treatment effects vary over time.



Note: 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. The trend-break effect equals $\phi \cdot (t - t_i + 1)$. The top of the figure notes which event-times lie in the $pre(k)$, $mid(k, \ell)$, and $post(\ell)$ periods for each unit. The figure also notes the average difference between timing groups in each of these periods. In the $mid(k, \ell)$ period, outcomes differ by $\frac{\phi}{2}\,(\ell - k + 1)$ on average. In the $post(\ell)$ period, however, outcomes had already been growing in the early group for $\ell - k$ periods, and so they differ by $\phi\,(\ell - k + 1)$ on average. The 2×2 DD that compares the later-treated group to the earlier-treated group is biased and, in the linear trend-break case, weakly negative despite a positive and growing treatment effect.

Note that this bias is specific to a single-coefficient specification. More flexible event-study specifications may not suffer from this problem.

3.2. What Is the Identifying Assumption for $VWATT$?

The preceding analysis maintained the assumption of *equal* counterfactual trends across timing groups, but (15) shows that when $\Delta ATT = 0$ identification of $VWATT$ only requires $VWCT = 0$. Assuming linear average untreated potential outcome trends

$$\overline{Y}_{kt}(0) - \overline{Y}_{k,t-1}(0) = \Delta Y_k^0$$

for all t, leads to a convenient and intuitive approximation to $VWCT$ as an average of each timing group's average trend in $Y(0)$ weighted by the difference between its weight as a treatment group w_k^T

from Equation (16) and a similar term measuring its weight as a control group:

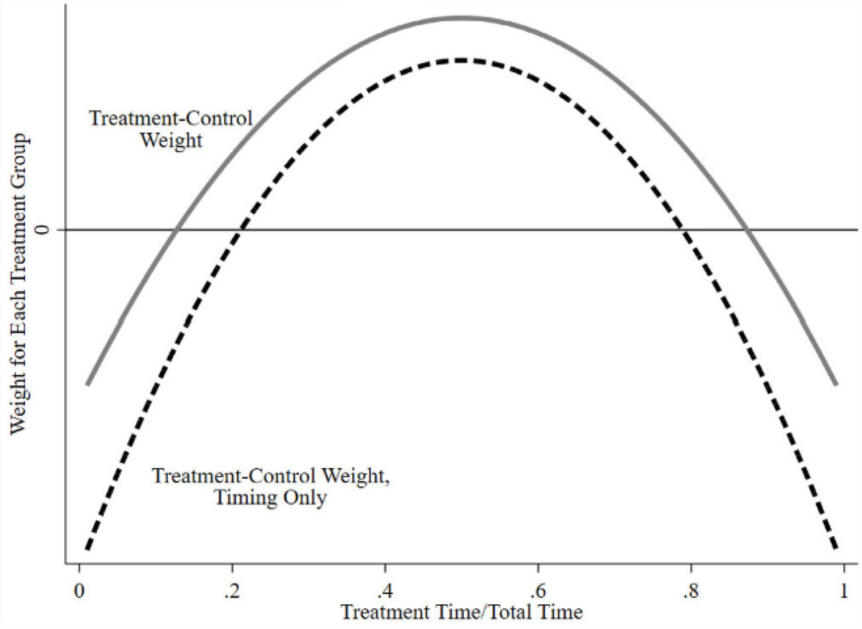
$$\begin{aligned} VWCT &\approx \sum_{k \neq U} \Delta Y_k^0 \left[\left(\sigma_{kU} + \sum_{j=1}^{k-1} \sigma_{jk}^k + \sum_{j=k+1}^{K} \sigma_{kj}^k \right) - \left(\sum_{j=1}^{k-1} \sigma_{jk}^j + \sum_{j=k+1}^{K} \sigma_{kj}^j \right) \right] - \Delta Y_U^0 \sum_{k \neq U} \sigma_{kU} \\ &= \sum_{k} \Delta Y_k^0 \left[w_k^T - w_k^C \right] \end{aligned} \tag{19}$$

Equation (19) generalizes the definition of common trends to the timing case and shows how a given timing group's counterfactual trend biases the overall estimate. To illustrate, assume there is a positive differential trend in timing group k only: $\Delta Y_k^0 > 0$. This will bias $\widehat{\beta}_{kU}^{2\times2}$ by ΔY_k^0 which gets a weight of σ_{kU} in the full estimation. In 2×2 DDs based on timing, however, biases offset each other. Take the comparisons to timing group 1, for example. When timing group k is the treatment group in $\widehat{\beta}_{1k}^{2\times2,k}$, the bias equals ΔY_k^0 and is weighted by σ_{1k}^k. When timing group k is the control group in $\widehat{\beta}_{1k}^{2\times2,1}$, the bias equals $-\Delta Y_k^0$ and is weighted by σ_{1k}^1. On net the bias in $\widehat{\beta}_{1k}^{2\times2}$ is ambiguous: $\Delta Y_k^0 \left(\sigma_{1k}^k - \sigma_{1k}^1 \right)$.

Similar expressions hold for the comparison of timing group k to every other group, and the total weight on each timing group's counterfactual trend equals the difference between the total weight it gets when it acts as a treatment group, w_k^T from Equation (16), minus the total weight it gets when it acts as a control group, w_k^C. This difference is a new result that maps (linear) differential trends to bias.

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

Figure 4: Weighted common trends: The treatment/control weights as a function of the share of time spent under treatment.



Note: The figure plots the weights that determine each timing group's importance in the weighted common trends expression in Equation (19).

References

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2024) "Revisiting Event-Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, rdae007, 10.1093/restud/rdae007.

de Chaisemartin, Clément and Xavier D'Haultfœuille (2020) "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110 (9), 2964–2996, 10.1257/aer.20181169.