NBER WORKING PAPER SERIES

HARVESTING DIFFERENCES-IN-DIFFERENCES AND EVENT-STUDY EVIDENCE

Alberto Abadie
Joshua Angrist
Brigham Frandsen
Jörn-Steffen Pischke

Working Paper 34550
http://www.nber.org/papers/w34550

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2025

This is a pre-publication draft chapter from *'Metrics Remastered: An Empiricist's Almanac,* Princeton University Press, 2026. We thank Kirill Borusyak for help with code, Nathan Nunn for help with data, and Ahmet Gulek, Merilin Martinson, Samuel McIntyre, and Grace Wang for exemplary research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w34550

Harvesting Differences-in-Differences and Event-Study Evidence
Alberto Abadie, Joshua Angrist, Brigham Frandsen, and Jörn-Steffen Pischke
NBER Working Paper No. 34550
December 2025
JEL No. C23, C5

ABSTRACT

This paper surveys econometric innovations related to differences-in-differences estimators and event-study models with time-varying treatment effects. Our discussion highlights tricky normalization issues, heterogeneous policy effects, the interpretation of exposure designs, pretrends pretesting, and the ever-bothersome question of logs versus levels. Key ideas are illustrated with applications.

Alberto Abadie
Massachusetts Institute of Technology
Department of Economics
and NBER
abadie@mit.edu

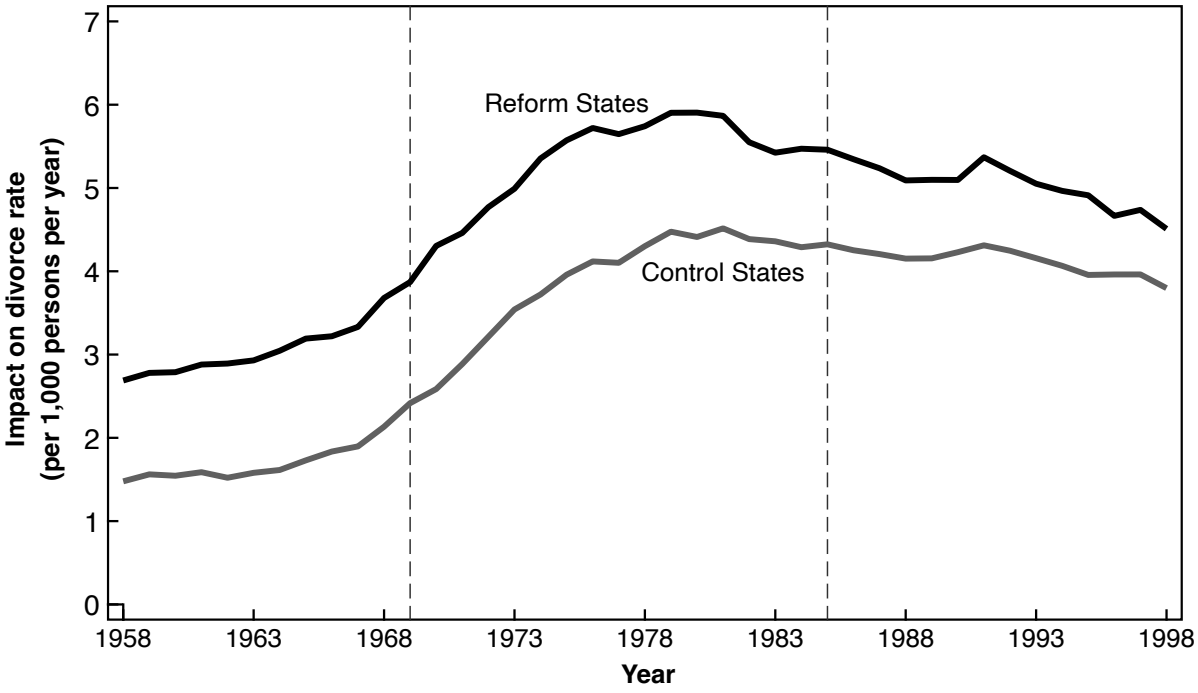
Joshua Angrist
Massachusetts Institute of Technology
Department of Economics
and NBER
angrist@mit.edu

Brigham Frandsen
Brigham Young University
Department of Economics
frandsen@byu.edu

Jörn-Steffen Pischke
London School of Economics and
Political Science (LSE)
Department of Economics
and NBER
s.pischke@lse.ac.uk

"Breaking up is hard to do" crooned Neil Sedaka in 1962—and American divorce law made sure of that. Until the 1970s, a stubborn spouse could veto a divorce. Then came the revolution: states began to adopt unilateral divorce laws, allowing one partner to dissolve a marriage without the other's consent.

Figure 1
Average divorce rates by reform status



Notes: This figure plots population-weighted averages of divorce rates for reform and control states. Reform states adopted unilateral divorce in the period between 1969-85 (marked by vertical lines). Control states left their divorce regime unchanged in this period. Data from Wolfers (2006).

Like weeds after a spring rain, divorce rates shot up across America in the 1970s, a pattern documented in Figure 1. Newly-enacted unilateral divorce laws are a likely culprit. How can we *know* whether unilateral divorce reforms are guilty of goosing divorce rates? Trends alone rarely offer definitive evidence of causal effects. But state legislatures have given us a series of legislative experiments that might answer this question. Some states adopted unilateral divorce sooner than others. This variation in timing is fertile ground for a differences-in-differences (DD) research design.

Friedberg (1998), Wolfers (2006), and others use within-state variation in the introduction of unilateral divorce laws to compute DD estimates of divorce law effects. Basic DD methods (detailed in Angrist and Pischke, 2008, 2015) estimate a single post-intervention treatment effect, a good starting point for any DD analysis. The modern event-study framework, however, goes beyond this by tracking treatment effects over time, both before and after the policy change of interest.

Our exploration of recent DD and related innovations starts with Wolfers's (2006) unilateral divorce data, cast in a conventional DD setup. Our divorce analysis uses data from 1958 to 1998, a period in which divorce rates are consistently available for most states (the panel is slightly unbalanced due to a few missing values). We exclude Louisiana, for which many years are missing, and Alaska and Oklahoma, which adopted unilateral divorce before 1958. States treated during the sample period reformed between 1969 and 1985. Dependent variable Y_{st} denotes the number of divorces per 1,000 persons in state s and year t; this is the way divorce rates are defined in US Vital Statistics. Divorce law has many dimensions. Unilateral adoption is defined as in Friedberg (1998).

After doing a little static DD, this paper focuses on event-study models with time-varying treatment effects, a major extension of the basic DD design. We highlight the fact that event-study identification relies on tricky normalization issues: depending on the data structure, identification requires one or more dynamic treatment effects to be set to zero. Better to choose these reference points deliberately than to let regression software make hidden—and potentially misleading—choices for you.

Contemporary DD and event-study frameworks (surveyed in Sun and Abraham, 2021, Goodman-Bacon, 2021, Roth et al., 2022) embrace heterogeneous policy effects across time and groups. Event-study models identify time-varying causal effects, but such effects can make simpler static DD estimates uninterpretable. And with unrestricted cross-sectional variation in impacts, regression-based DD estimates may fail to deliver the weighted average of state-specific effects you seek. Our analysis of two applications suggests that these

concerns, while theoretically important, are unlikely to derail DD or event-study designs in practice.

We also examine the key *parallel trends* assumption underpinning all DD-type models. Seasoned event-study researchers examine pretreatment data to scope out worrying signs of pretrends. This sort of pretrends pretest is usually a good idea—but not always. And, as readers of Angrist and Pischke (2008) will know, parallel trends in the log of an outcome variable precludes parallel trends in levels. The paper closes with an examination of a valiant recent attempt to wriggle out of this troubling functional-form constraint.

1 DD is for Divorce

Here's the simple DD setup for the state-year panel data set we use to estimate unilateral divorce effects. States are indexed by $s = 1, \ldots S$ and years by $t = 1, \ldots, T$. In discussions where specificity seems helpful, s takes on values from a list of state abbreviations ($s \in \{AL, ..., WY\}$) and t indexes calendar years. In both notational variations, dummy variable $D_{st} \in \{0, 1\}$ indicates states and years in which unilateral divorce is allowed. Each state s belongs to a *cohort*, stored in variable $c(s)$. This variable identifies the period when unilateral divorce came into effect in state s. That is, $D_{st} = 1$ for all years $t \geq c(s)$. In this *staggered-adoption design*, treatment, once switched on, stays on forever.

Potential outcome $Y_{st}(0)$ denotes the divorce rate in state s and year t that we'd see in the absence of a unilateral divorce law. $Y_{st}(1)$ is the potential divorce rate we'd see if such a law were on the books. Outcomes $Y_{st}(0)$ and $Y_{st}(1)$ are potential because only one or the other is revealed for a given state and year. Also, while individual treatment effects, $Y_{st}(1) - Y_{st}(0)$, might vary with state and year, we start with a simpler setup.

Our starter DD model posits a constant, additive treatment effect τ, so that $Y_{st}(1) = Y_{st}(0) + \tau$. Conventional DD analysis relies on a regression model in which expected untreated

potential outcomes depend on a state effect (γ_s) and a time effect (λ_t) according to

$$Y_{st}(0) = E[Y_{st}(0)] + \eta_{st} = \gamma_s + \lambda_t + \eta_{st}. \tag{1}$$

Residual η_{st} is a mean zero conditional expectation function (CEF) error term while γ_s (for $s = 1, \ldots, S$) and λ_t (for $t = 1, \ldots, T$) are parameters that restrict the behavior of this CEF.

Randomness arises in this setup from imagined alternative histories (admittedly a fanciful notion, though familiar to physicists and science fiction fans). One of these histories becomes reality. Parameters γ_s and λ_t are presumed to be constant, while η_{st} is drawn from a distribution of potential histories. $E[Y_{st}(0)]$ is the average of the resulting potential outcomes over all such possible draws.[1]

The assumption that state effects γ_s are time-invariant while time effects λ_t are common across states plays a major role in DD analysis. It's worth emphasizing that this model-based presentation of DD differs from the discussion of regression at the heart of Angrist and Pischke (2008). While the latter allows for the possibility that the regression function of interest approximates a more complicated CEF, DD models restrict the CEF from the get-go.

In a world where state legislators legislate by coin toss, family law is made independently of potential outcomes. In such a world, data on state divorce rates can be analyzed as if from a randomized trial. In the real world, legislators legislate for various and sundry reasons, some idiosyncratic and some systematic. Motivated by this, the state effects in (1) allow for cross-state differences in divorce rates in the absence of reform. In Figure 1, for instance, we see that reforming states had higher divorce rates in the pre-reform years. Likewise, the spread of unilateral divorce laws coincided with a nationwide trend towards increasing rates of marital dissolution. This trend, common to both reforming and non-reforming states, is captured by the time effects in (1).

[1]Divorce rates by state and year come from vital statistics and are measured for the relevant state populations. In other DD applications, aggregate variables (like average wages) come from sample surveys, in which case randomness in η_{st} reflects sampling variance as well as variation in realized potential outcomes.

By including state and time effects as parameters, the DD model allows for both cross-sectional and temporal variation in potential outcomes, variation we'd see even in the absence of a policy change. But DD requires these effects to be additive, a point this paper repeatedly returns to. The DD additivity requirement is stronger than the linearity assumption often invoked for regression controls. In conventional regression control strategies, linearity is a convenient and mostly harmless approximation. In DD applications, by contrast, functional form assumptions are consequential.

In addition to additive state and time effects, the basic DD model assumes that the random part of potential outcomes in each period is mean-independent of the sequence of treatments for state s:

$$E[\eta_{st}|D_{s1},\ldots,D_{sT}] = E[\eta_{st}|c(s)] = 0; \text{ for all } s = 1,\ldots,S \text{ and } t = 1,\ldots,T. \tag{2}$$

Again, the expectation here is computed over the distribution of CEF errors η_{st} for each state and year.

You might recognize (2) as a conditional independence assumption (CIA) of the sort used to imbue regression estimates with a causal interpretation (see e.g. Section 3.2 in Angrist and Pischke, 2008). The DD CIA embodies two key differences, however. First, the left-hand side conditions on the entire treatment sequence $(D_{s1},\ldots,D_{sT})$, not just contemporaneous treatment status. In the staggered-adoption design, this is the same as conditioning on treatment cohort $c(s)$. Second, as in Borusyak, Jaravel and Spiess (2024) and Callaway and Sant'Anna (2021), our DD CIA applies each period $t \in \{1,\ldots,T\}$, rather than averaged across periods.

Importantly, the panel regression setup embodied in (1) and (2) implies a strong *parallel trends* restriction. Parallel trends means that, in the absence of unilateral divorce laws, divorce rates evolve similarly in all states. This is formalized as:

Assumption 1 (Parallel Trends)**.**

$$E\left[Y_{st}\left(0\right) - Y_{st-1}\left(0\right)\middle| c\left(s\right)\right] = E\left[Y_{st}\left(0\right) - Y_{st-1}\left(0\right)\right]; \quad t = 2, \ldots, T.$$

To see why parallel trends is implied by (1) and (2), use (1) to write the change in potential outcomes as:

$$Y_{st}\left(0\right) - Y_{st-1}\left(0\right) = \lambda_t - \lambda_{t-1} + \eta_{st} - \eta_{st-1}. \tag{3}$$

By iterated expectations, $E\left[\eta_{st} - \eta_{st-1}\middle| c\left(s\right)\right] = 0$ implies $E\left[\eta_{st} - \eta_{st-1}\right] = 0$. Consequently, the expected trend in potential outcomes is $\lambda_t - \lambda_{t-1}$ for any and all cohorts. The parallel trends assumption rules out a world in which states, for instance, opt for unilateral laws in the wake of locally rising divorce rates. In principle, such idiosyncratic variation might happen at any time, so condition (2) disallows this for each period.

DD analysis runs on regression. In combination with constant treatment effects, (1) yields:

$$Y_{st} = \tau D_{st} + \gamma_s + \lambda_t + \eta_{st}. \tag{4}$$

Condition (2) ensures that this is a regression model.

State and year effects γ_s and λ_t in (4) may seem mysterious since they're defined as parameters while appearing to play the role of conditioning variables. Like τ, however, these are regression *coefficients* in a model that's as yet unstated. We state it here: each γ_k (where k indexes states) is the coefficient on a dummy variable, d_{ks}, that indicates whether observation Y_{st} is from state k. Each λ_l (where l indexes years) is the coefficient on a dummy, h_{lt}, that indicates observations from year l. Equation (4) is therefore shorthand for:

$$Y_{st} = \tau D_{st} + \sum_k \gamma_k d_{ks} + \sum_l \lambda_l h_{lt} + \eta_{st}. \tag{5}$$

This classic two-way fixed effects (TWFE) specification regresses Y_{st} on dummies for every state, dummies for every year, and a treatment dummy indicating states and years having a unilateral divorce regime in place.[2] The DD setup embodied in (5) is a TWFE model because it controls for two sets of dummy variables, also called *fixed effects*, with no interactions between them.

Note that (5) has no intercept. Because this model includes a dummy for every state, however, a constant is implicit since a full set of state or year dummies adds to one. In practice, therefore, most empirical DD analysis includes an intercept, while omitting one state dummy and one year dummy. This leaves a set of regressors with no linear dependencies—no news here. But normalization of this sort grows surprisingly complicated in the event-study framework to come.

A note on estimation: let $\bar{Y}_s$ denote mean divorce rates computed by averaging over t for a given s, with bars over other variables interpreted similarly. By the regression anatomy theorem, ordinary least squares (OLS) estimates of τ in (5) can be computed from a regression in which all variables are deviations from state means:

$$Y_{st} - \bar{Y}_s = \tau(D_{st} - \bar{D}_s) + \sum_l \lambda_l(h_{lt} - \bar{h}) + (\eta_{st} - \bar{\eta}_s), \tag{6}$$

where $\bar{h}$ is the average time dummy (a constant for all states). The deviations-from-means transformation eliminates state effects.[3]

The CIA described by (2) ensures that transformed treatments in (6) are uncorrelated with transformed residuals. The fact that the transformation involves treatment averaged

[2] In the sum $\sum_k \gamma_k d_{ks}$, subscript k loops over all states since every state gets a dummy, while subscript s keeps track of the state supplying the observations. So, $d_{ss} = 1$ and $d_{ks} = 0$; $k \neq s$. Likewise, in the sum $\sum_l \lambda_l h_{lt}$, subscript l loops over all years since every year gets a dummy, while subscript t keeps track of the year supplying the observations. So, $h_{tt} = 1$ and $h_{lt} = 0$; $l \neq t$.

[3] Regression anatomy (see e.g. Section 3.1 in Angrist and Pischke, 2008) says that multivariate coefficients on variables of interest can be obtained by regressing these variables on covariates and using the resulting residuals in a model without covariates. Here, the covariates are state dummies. Recall also that a regression on a full set of dummies is *saturated* and so recovers the CEF given regressors, in this case, a set of state means.

over time, $\bar{D}_s$, explains why restriction (2) conditions on all leads and lags of treatment. Assuming the regressor of interest varies over time, (2) implies that OLS estimation of (6) yields unbiased estimates of treatment effect τ (Wooldridge, 2016).

How many states are needed for fruitful DD? At least two: a treatment state (say, California, reforming in 1970) and a control (like New York, unreformed in the sample period). Parallel trends, formalized in (2), ensure that New York trends provide a valid counterfactual for California. But unbiasedness does not guarantee meaningful estimates. Both states are subject to idiosyncratic variation—captured by η_{st}—that is likely to make two-state comparisons misleading. We hope, therefore, that comparisons involving many states smooth such idiosyncrasies, painting a picture in which evidence for a treatment effect emerges clearly. Formally, OLS estimates are consistent in an asymptotic sequence in which the number of cross-sectional units grows while T is fixed (Chamberlain, 1984). Figure 1 looks promising since reform and control state averages move smoothly in tandem in pre-reform years.

In these data, basic DD delivers an estimate of τ equal to -0.22 divorces per thousand persons with a standard error of 0.16, a small effect that's not significantly different from zero. As can be seen in Figure 1, divorce rates rose country-wide from around three per thousand in the late 1960s to over five per thousand by 1980. Yet, TWFE regression estimates suggest unilateral divorce had little to do with this increase.

The Main Event

Modern event-study models extend simple DD by allowing for time-varying treatment effects. The payoff to this extension is a more nuanced picture of policy effects, such as those of unilateral divorce laws, and a framework that can be used to validate the key parallel trends assumption.[4] As it turns out, the dynamics of the unilateral divorce story are interesting indeed.

[4] Miller (2023) reviews contemporary event-study research methods.

Event-study DD requires additional notation. In a staggered adoption design, as in the unilateral divorce setting, treatment stays on once switched on, so

$$\Delta D_{st} \equiv D_{st} - D_{st-1}$$

equals one in the year that state s implements unilateral divorce and is zero otherwise. This makes ΔD_{st} a *treatment switch.*

Lagged treatment switches (*lags* for short), denoted ΔD_{st-j}, equal one in year t when state s adopted unilateral divorce j years ago. California, for instance, adopted unilateral divorce in 1970. $D_{CA,t}$ therefore equals 1 for California data in $t = 1970$ and later; $\Delta D_{CA,t}$ equals 1 only in 1970; and $\Delta D_{CA,t-2}$ equals 1 only in 1972. Leading treatment switches (*leads* for short), denoted ΔD_{st+j}, equal one in year t when state s switches to unilateral divorce j years from t.

This notation has the virtue that only one variable in the sequence of treatment switches $\{\ldots, \Delta D_{st+2}, \Delta D_{st+1}, \Delta D_{st}, \Delta D_{st-1}, \Delta D_{st-2}, \ldots\}$ is equal to one for a given state, s, and year, t. In California in $t = 1972$, for instance, $\Delta D_{CA,t-2}$ equals 1, while all other leads and lags for California defined at time t equal 0. At the same time, the sum of all lags equals the original treatment dummy D_{st}, the regressor of interest in a static-DD setup. These features facilitate the careful record-keeping needed to interpret time-varying treatment effects, which we call *dynamic effects*.

Event-study regression models retain the TWFEs used to mitigate omitted variable bias in simple, static DD regression, while also allowing for dynamic effects. Unilateral divorce might matter little, for instance, in the year it's first adopted, with a growing impact thereafter. Event-study models also look ahead, identifying leading policy effects. Leads might arise because people change their behavior in anticipation of a policy change. In the absence of anticipation effects, however, non-zero leads signal divergence from parallel trends.

Our event-study regression model specifies q lags and $m - 1$ leads (with $m \geq 2$). This

model can be written:

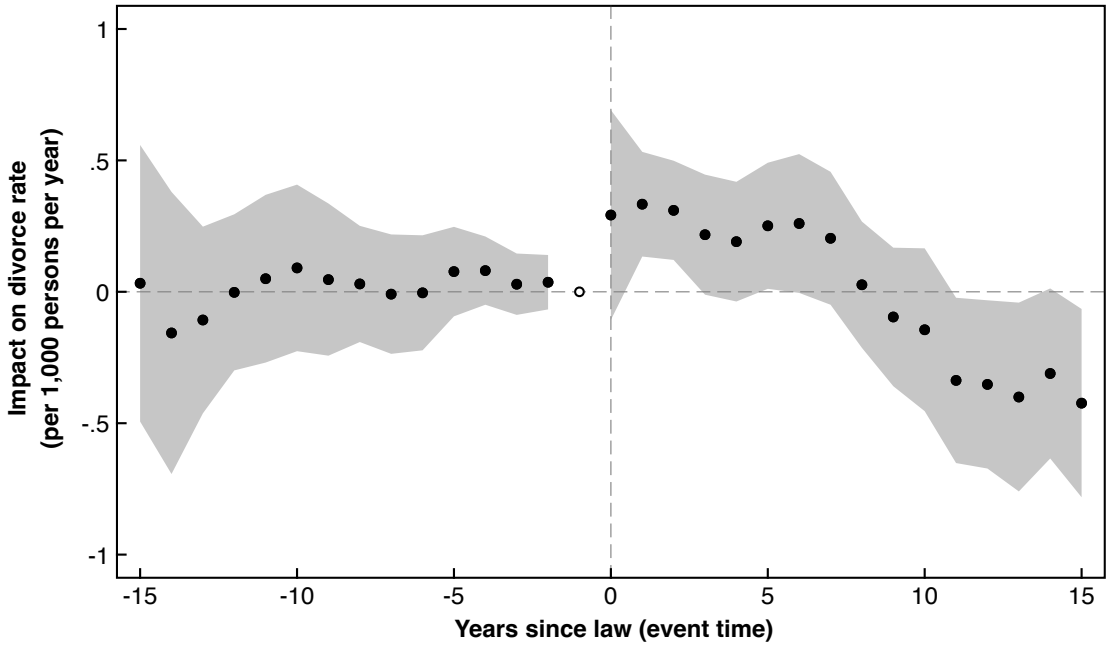
$$Y_{st} = \sum_{j=-m}^{-2} \tau_j \Delta D_{st-j} + \sum_{j=0}^{q} \tau_j \Delta D_{st-j} + \gamma_s + \lambda_t + \eta_{st}, \tag{7}$$

where parameters τ_j are the dynamic treatment effects of interest. These are indexed by years since or until the year of adoption, called *event time*, and denoted by j in the summations above.

The notation here looks daunting, so let's spell things out. Recall that $\Delta D_{st} = 1$ in the year t in which unilateral divorce is adopted in state s. The coefficient associated with the adoption year is τ_0, a term that appears in the sum $\sum_{j=0}^{q} \tau_j \Delta D_{st-j}$ when $j = 0$. The treatment effect two years after adoption is τ_2; this is the period in which $\Delta D_{st-2} = 1$. In general, coefficients on D_{st-j} capture evolving impacts after a policy change, revealing whether effects increase, stabilize, or fade. Leads are a little trickier than lags. The former are captured by the term $\sum_{j=-m}^{-2} \tau_j \Delta D_{st-j}$ in (7). For California, the term inside this sum that looks two years ahead switches on in 1968 (this is $\Delta D_{CA,t+2}$). Thus, leads allow for pre-treatment treatment effects.

When using dummies to indicate values of any categorical regressor, we omit a reference category to avoid collinearity with the constant. Similarly, at least one treatment switch must be omitted in an event-study model. To see why, note that for a reform state, the sum of treatment switches equals one. Consequently, the sum of the full set of switches, including all leads and lags, is equal to a dummy variable indicating all reform states. The latter, of course, is a linear combination of state dummies. The parameterization described by equation (7) omits $\tau_{-1} \Delta D_{st+1}$. Treatment effect estimates in this model are therefore measured relative to divorce rates in the year ahead of each reforming state's reform year (which defines a state's cohort). Estimates of τ_0, for instance, measure the extent to which divorce rose in the year unilateral divorce was introduced, relative to the year before, while τ_2 contrasts divorce rates two years post reform with the same pre-reform benchmark.

Figure 2
Event-study estimates of the effect of unilateral divorce on divorce rates



Notes: This figure plots event-study estimates of the effects of unilateral divorce reform on divorce rates, with the year before reform set as the reference year. Treatment is staggered; once treated, states remain unilateral thereafter. The shaded area marks confidence bands based on standard errors clustered by state. Estimates are computed using data from 1958-1998.

Event-study estimates, plotted in Figure 2, reveal interesting dynamic effects masked by simple DD results. Unilateral divorce appears to boost divorce rates by around 0.3 per thousand points in the first seven years after adoption. But the estimates then plummet, and, nine years out, turn negative. The divorce rate ultimately falls to roughly 0.4 per thousand points below what it would have been absent reform. This pattern may be explained by pent-up demand for marital dissolution in pre-reform years. Once couples held together by the old regime have separated, divorce rates settle at a new, lower level.

The estimates in Figure 2 suggest that the shift to unilateral divorce may indeed have contributed to the rise in divorce rates in the 1970s. But this effect appears to have been modest as well as short-lived, particularly since only about 60 percent of states introduced a unilateral regime in the sample window. At the same time, the long-run impact of reforms

of around -0.4 per thousand potentially explains almost half of the 1 per thousand point *decline* in divorce rates seen during the 1980s and 1990s.

The long-run negative impact of unilateral divorce on marital dissolution is a striking finding. Should the treatment effect estimates in Figure 2 be seen as causal? Evidence of causal validity comes from the pre-treatment coefficients plotted in the figure. These hover around zero, rarely exceed 0.1 in absolute value, and are statistically indistinguishable from zero. Divorce trends in reforming states and years do not appear to have been diverging ahead of the advent of a unilateral regime. In other words, the estimated leads are consistent with a presumption of parallel trends, as we explain in Section 3.[5]

Our event-study model includes 27 leads and 29 lags, values derived below (Figure 2 omits leads and lags beyond 15 to avoid clutter). Standard errors are higher for treatment effects at longer leads and lags because fewer states contribute to estimates of effects at more distant horizons. This motivates DD estimators that pool—or *bin*—effects distant from the adoption date. Suppose we bin leads and lags for values of $j \geq 15$ and $j \leq -15$, with the associated pooled treatment effects labeled $\tau_{\geq 15}$ and $\tau_{\leq -15}$, respectively. As before, the reference group for treatment effects is one year before the advent of unilateral divorce. Our event-study regression model with binned leads and lags looks like this:

$$\begin{aligned} Y_{st} = &\sum_{j=-14}^{-2} \tau_j \Delta D_{st-j} + \sum_{j=0}^{14} \tau_j \Delta D_{st-j} \\ &+ \tau_{\leq -15} \left(\sum_{j \leq -15} \Delta D_{st-j} \right) + \tau_{\geq 15} \left(\sum_{j \geq 15} \Delta D_{st-j} \right) \\ &+ \gamma_s + \lambda_t + \eta_{st}. \end{aligned} \tag{8}$$

The terms in parentheses collapse treatment indicators beyond 14 years pre-treatment and 14 years post-treatment into single dummies. Estimates from the binned model differ little

[5]The evidence here is not seamless. Lee and Solon (2011) note that unweighted estimates in this context differ from those weighted by state population. As in Wolfers (2006), the estimates here use state population weights, thereby giving larger states more weight. Population weighting is defensible—divorce rates may be noisier in smaller states—but not obviously essential in a group-level analysis of administrative data.

from those plotted in Figure 2 and so are omitted.

Into the Weeds: Collinearity Complications

Event-study analysis demands careful model specification; the devil is in the DD details. We've seen that simple DD models omit a reference category when specifying a full set of state and year effects. A coherent event-study regression specification requires more than this, however. For starters, event-study models specify the number of treatment leads (m) and the number of treatment lags (q).

How many leads and lags can be estimated? A lag of length q requires at least one treated state with data up to q years after adoption. The longest possible lag is therefore $q = T - \min_s(c(s))$, where T is the panel length and $c(s)$ is the treatment cohort. Similarly, a lead of length m requires at least one treated state with observations up to m years before adoption. The longest lead is therefore $m = \max_s(c(s)) - 1$.

Our divorce analysis uses data from 1958 (indexed by $t = 1$) to 1998 ($T = 41$). Kansas is the first state in this sample to introduce unilateral divorce, in 1969 ($t = 12$). This determines the number of allowable lags as $q = T - \min(c(s)) = 41 - 12 = 29$. South Dakota, the last adopter in our sample, adopted in 1985 ($t = 28$). This allows for up to $m = \max(c(s)) - 1 = 28 - 1 = 27$ leads.

Even with lead and lag lengths specified, the event-study setup is not yet good to go. Because treatment dummies sum to an indicator of reform states, it's customary to omit the dummy for the period before adoption, $j = -1$. The remaining τ_j terms then reflect treatment effects relative to the period just before treatment. The estimates in Figure 2 were computed in a sample that includes some states (like New York) that are untreated in the sample period (New York adopted unilateral divorce in 2010, while our sample runs until 1998). When some states are never treated, omission of a reference period dummy is enough to identify event-study leads and lags.

With no never-treated states, model specification grows messier. Econometricians an-

alyzing panel data have long grappled with the fact that cohort (year of birth), age, and calendar time are linearly dependent (your age in year t equals t minus your birth year). We cannot, therefore, use dummies to control for all three of these variables, conceptually distinct though they may be. When all states are eventually treated, the same reasoning applies to models with state effects, year effects, and a full set of treatment leads and lags. For treated states, event time, like age, equals calendar year minus treatment cohort, that is, $j = t - c(s)$. A full set of state and time dummies is therefore collinear with dynamic treatment effects that vary linearly with event time. The presence of never-treated states breaks this linear dependence because event-study coefficients equal zero for such states.[6]

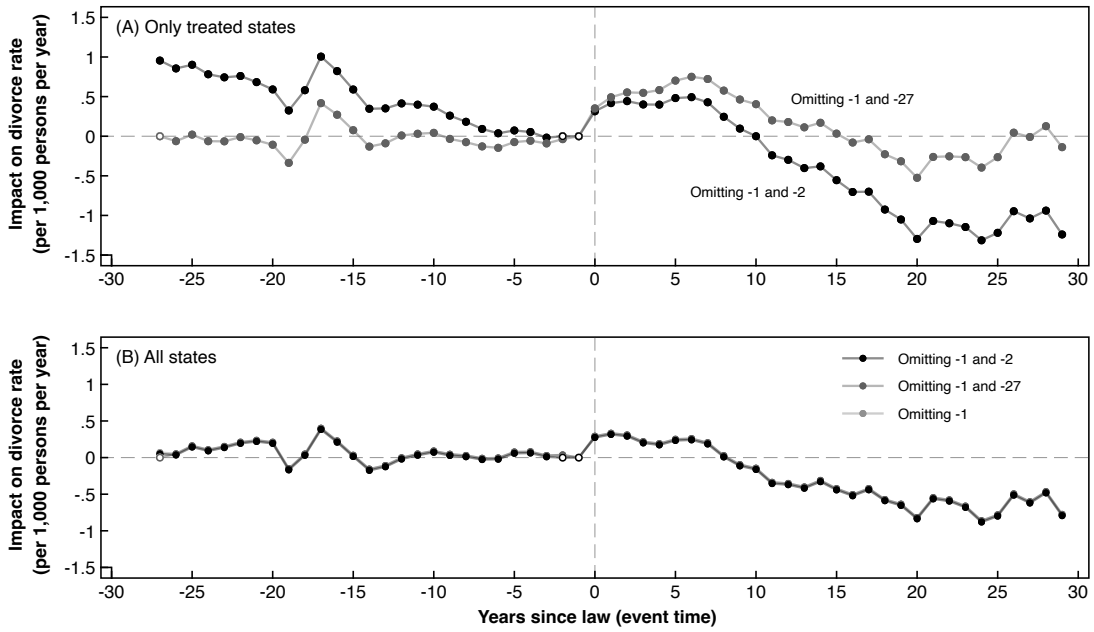
Omission of at least one *additional* lead or lag kills collinearity when there are no never-treated states. This omission (setting an additional treatment switch to zero besides that for $j = -1$) makes treatment effects a non-linear function of calendar time and cohort. Problem solved—perhaps. After all, it seems reasonable to omit far-ahead leads since, under parallel trends, these should be zero anyway. Does the additional normalization needed here matter?

The question of how such specification details affect event-study results is explored in Figure 3. The top panel plots two sets of estimates computed in a subsample that drops never-treated states. Estimates plotted in grey in this panel are from a model that omits the treatment switch for $j = -27$ (in addition to the -1 lead omitted as a reference group), while the line plotted in black is from a model that omits the second lead ($j = -2$) instead of $j = -27$. The top panel shows how dynamic treatment effects are substantially rotated around the common intercept at $j = -1$ by the choice of additional normalization—the linear portion of the entire path of treatment effects is no longer identified. The lower panel of Figure 3 displays estimates in the full sample omitting the same switches as before. Because the full sample only requires one normalization, that panel also includes estimates only omitting the dummy for event time -1. All three estimates are indistinguishable.[7]

[6] Borusyak, Jaravel and Spiess (2024) appears to be the first to make this point.

[7] Not for nothing does the extra normalization in this context appear to *rotate* the set of dynamic effects. As Borusyak, Jaravel and Spiess (2024) explains, it's the slope of τ_j viewed as a function of j that's unidentified in an analysis with no never-treated states. To see this, suppose $\tau_j = 1 + j$. This preserves our

Figure 3
Event-study normalizations with and without never-treated states



Notes: This figure plots event-study estimates of unilateral divorce effects on divorce rates, computed in a sample limited to ever-adopting states (in the top panel) and in the full sample (in the bottom panel). Plotted lines show effects under alternative normalizations, omitting either the coefficient for -2 or -27 in addition to the reference year of -1. Estimates are computed using data from 1958-1998.

Seemingly-innocuous normalizations are consequential for event-study estimates in state panels with no never-treated states. This problem arises from the fact that without never-treated states, identification of event-study treatment effects hinges on an additional restriction. The identification problem here mirrors that arising in DD models that control for state-specific linear trends, such as the models detailed in Angrist and Pischke (2008, 2015). As a reminder, TWFE models with state-specific trends add a term like $\sum_k \psi_k d_{ks} t$, where ψ_s is the trend for state s, to DD and event-study regressions.

reference-period normalization, with $\tau_{-1} = 0$. Substituting for τ_j and combining leads and lags in a single sum in (7), we have:

$$Y_{st} = \sum_{j=-m}^{q} (1+j)\Delta D_{st-j} + \gamma_s + \lambda_t + \eta_{st}.$$

The first term above equals $1+j$ when ΔD_{st-j} switches on, and is therefore equal to $1+t-c(s)$, which is a linear combination of time and state effects.

Models with state-specific trends weaken the parallel trends assumption by allowing for diverging *linear* trends. With no trend break at the time treatment switches on, however, we can't say whether trend divergence is due to slowly rising dynamic treatment effects coupled with anticipation of future treatment or a reflection of omitted variables bias due to trending $Y_{st}(0)$. As in event-studies with no never-treated states, TWFE with state-specific trends requires a second normalization to identify a full set of event-study coefficients. If you're willing to commit to zero pre-treatment effects, models with state-specific linear trends are identified by discontinuities or jumps in outcomes at the time treatment switches on.

2 DD With Heterogeneous Effects

As Mark Twain might have said had he been an economics Ph.D. student, everybody talks about DD treatment effect heterogeneity, but nobody does anything about it. Well, not anymore. Event-study models allowing for effect variation over time, for instance, show that the effects of divorce laws change markedly in the years following reform. As it turns out, in the divorce example, the classic one-parameter TWFE regression estimator delivers something close to an average of the corresponding event-study estimates. But this result is not a theorem; de Chaisemartin and d'Haultfoeille (2020), Borusyak, Jaravel and Spiess (2024), and Goodman-Bacon (2021), among others, show that DD estimates need not average time-varying effects. Of course, you can always opt for event-study estimates and average these yourself.

Here's some good news from the cross-sectional heterogeneity front: with a single dummy treatment and two periods, old-school TWFE estimation (as in many first-generation TWFE studies like Card, 1992) recovers a cross-state average causal effect on the treated. This and other key points are made here in a simplified setup inspired by Sun and Shapiro (2022), which considers effects of state health insurance mandates. The aptly-named "Appendix B" to Goldsmith-Pinkham, Hull and Kolesár (2021) discusses DD models with heterogeneity

more generally, as does Sun and Abraham (2021).[8]

Two-period DD With Heterogeneity Too

Consider a unilateral divorce reform scenario evolving over two periods, $t \in \{0, 1\}$, and two groups of states, reform and non-reform, with reformers indicated by the dummy variable M_s. Reform is implemented in $t = 1$. The causal model of interest is:

$$Y_{st} = \gamma_s + \lambda t + \tau_s M_s t + \eta_{st}, \tag{9}$$

where γ_s and λ are state and time effects, respectively (λ is the coefficient on dummy variable t, indicating the reform year), and τ_s is a state-dependent causal effect of reform. Interacting M_s with time defines the DD treatment dummy:

$$D_{st} \equiv M_s t.$$

As in the general DD regression model (5), identification comes from parallel trends, expressed in this case by:

$$E[\eta_{st}|M_s] = 0; \ t \in \{0, 1\}.$$

With two periods and treatment switched on in only one, there's no scope for time-varying effects. We might, however, allow for variation in reform effects across states, a possibility captured by writing τ_s for the causal effect of interest. In particular, treatment effects might differ in reform and non-reform states. As in Sun and Shapiro (2022), this dependence can be modeled as:

$$\tau_s = \kappa + \phi M_s, \tag{10}$$

where Greek letters denote parameters. Since M_s is a dummy, this is an unrestricted model

[8] Aptly named because, as 'metrics masters will know, IV was invented in Appendix B of the pamphlet by Wright (1928).

for the relationship between causal effects and reform status. In particular, $\kappa+\phi$ is the effect of reform in states that reform, while κ is the effect of reform elsewhere.

To see what this sort of heterogeneity implies for regression DD estimates, substitute (10) into (9):

$$Y_{st} = \gamma_s + \lambda t + (\kappa + \phi M_s)M_s t + \eta_{st} \tag{11}$$

$$= \gamma_s + \lambda t + (\kappa + \phi)M_s t + \eta_{st}, \tag{12}$$

where the second line uses the fact that $M_s^2 = M_s$ for Bernoulli (dummy) M_s. From this, we see that for simple DD with heterogeneous effects, TWFE estimation recovers the effect of reform in reform states.

This DD averaging story carries over to models with covariates beyond time and state effects. Suppose we add $\lambda_t' X_s$, where λ_t is a time-dependent coefficient vector conformable to a vector of discrete controls, X_s. Assume these are dummies that saturate (i.e., indicate all values of) an underlying set of discrete controls, so that $E[M_s|X_s]$ is linear.[9] Covariates are assumed to have time-varying effects (otherwise they're absorbed by state dummies). The covariate vector includes a constant, so this model nests the simpler TWFE specification embodied in (9). Replacing λt with $\lambda_t' X_s$ in (9), and differencing to eliminate state effects, we have:

$$\Delta Y_s = (\lambda_1 - \lambda_0)' X_s + \tau_s M_s + \nu_s, \tag{13}$$

where $\Delta Y_s \equiv Y_{s1} - Y_{s0}$ and ν_s is the differenced residual.

Now, suppose we allow covariate interactions in heterogeneous treatment effects as well as variation that depends on M_s. This can be expressed by writing

$$\tau_s = X_s'\kappa + M_s(X_s'\phi), \tag{14}$$

[9] If the control is census region, for instance, X_s includes dummies for 3 of 4 regions.

where κ and ϕ are now vectors conformable with X_s. We can use regression anatomy and (13) to show that in this case the coefficient on M_s in a regression of ΔY_s on X_s and M_s can be written as:

$$\begin{aligned}\tau^{\text{OLS}} &\equiv \frac{E[\widetilde{M}_s \Delta Y_s]}{E[\widetilde{M}_s M_s]} = \frac{E[\widetilde{M}_s M_s \tau_s]}{E[\widetilde{M}_s M_s]} \\ &= \frac{E[\widetilde{M}_s M_s X_s'(\kappa + \phi)]}{E[\widetilde{M}_s M_s]}\end{aligned} \tag{15}$$

where $\widetilde{M}_s = M_s - E[M_s|X_s]$.[10] Finally, iterating over X_s gives:

$$\tau^{\text{OLS}} = \frac{E[\sigma_Z^2(X_s) X_s' \tau]}{E[\sigma_Z^2(X_s)]},$$

where

$$\sigma_Z^2(X_s) \equiv E[\widetilde{M}_s^2 | X_s],$$

$\tau \equiv \kappa + \phi$, and $\tau' X_s$ is the reform effect on reform states at covariate value X_s; the parameter denoted τ^{OLS} is a variance-weighted average of these.

This result aligns with formulas in Angrist (1998) and Angrist and Krueger (1999) showing that regression on a dummy treatment with saturated covariate controls recovers a variance-weighted average of covariate-specific treatment effects. The weights in this context are given by the variance of M_s given X_s. Abadie (2005) offers a semiparametric take on this, deriving a weighted least squares estimator that estimates average treatment effects on the treated in simple DD models with covariates. Happily, the DD angle requires no new thinking in this two-period, dummy-treatment case.

Potato, Potahto: Heterogeneity in Exposure Designs

Potatoes have been welcome at European dinner tables for centuries. Hungarians hunger for a savory goulash of stewed beef and potatoes. Ashkenazi Jews pine for potatoes shaped

[10] This derivation uses the facts that $\widetilde{M}_s$ is a CEF residual because $E[M_s|X_s]$ is linear and that $M_s(X_s'\kappa + M_s(X_s'\phi)) = M_s X_s'(\kappa + \phi)$ because M_s is a dummy.

into latkes and for dense potato kugel. Josh Angrist's Eastern European grandparents knew these satisfying dishes well.

Potato products took off in 18th-century Europe, quickly becoming a staple. By offering a new and inexpensive source of nourishment, the humble potato may have improved the lives of all those who eat for a living. Nunn and Qian (2011) assess this intriguing claim in a DD analysis of a country-by-century panel. The potato treatment here is the share of a country's arable land suitable for potato cultivation, interacted with the approximate date of potato arrival from the Americas, set at 1700 worldwide. The dependent variable is log population size (by country), a crude measure of human welfare. A positive coefficient indicates that countries with more potato-positive arable land saw a greater increase in population growth after 1700 relative to before 1700. This difference is estimated in centennial data from the second millennium.

Treatment in the potato problem comes from differences in cross-sectional exposure (specifically, land suitability) rather than differences in timing. Such identification strategies are nowadays known as *exposure designs* (Sun and Shapiro, 2022). Pioneering exposure designs include Card (1992), which estimates the effect of an increase in the federal minimum wage in 1990 on the employment of teenagers. The Card study exploits the fact that high-wage states (defined as such pre-increase) are less affected by a federal minimum wage hike than are low-wage states. Similarly, Finkelstein (2007) estimates the effect of the advent of Medicare in 1966 on the hospital industry. This exposure design exploits the fact that Medicare (America's government insurance program for the elderly) was less important insurance-wise in states with high pre-Medicare private insurance rates. Recently, Figlio and Özek (2025) uses variation in pre-ban use to identify the effects of cell-phone bans on children's learning across American school districts.

In these applications, the key exposure variable, defined in a fixed pre-treatment or baseline period, is a fraction like the share of arable land suited for potato farming, rather than a dummy. Typically, the sample used in an exposure design contains no never-treated

units. Rather, treatment varies more or less continuously.[11]

How well do exposure designs accommodate heterogeneous effects? Sticking with the notation used to analyze dummy DD heterogeneity at the beginning of this section, here we assume that M_s denotes fractional baseline exposure. The potato exposure design with heterogeneous effects can be implemented by estimating (11), rewritten as:

$$Y_{st} = \gamma_s + \lambda t + \kappa M_s t + \phi M_s^2 t + \eta_{st}. \tag{16}$$

The squared term above is generated by substituting for τ_s using (10). Since M_s is now a fraction rather than a dummy, $M_s \neq M_s^2$ and the simplification in (12) doesn't apply. Treatment-effect parameters κ and ϕ are the coefficients on exposure and exposure squared, both interacted with time.

While this model is easily estimated, it's interesting to examine the exposure-design *estimand* when the potential-outcome CEF given s and t follows (16), while exposure effects are estimated using a regression that omits $M_s^2 t$. [12] After all, the typical exposure design study reports estimates of models with a linear interaction only. It seems reasonable to expect the bivariate slope coefficient in question to be the average of state-specific τ_s,

$$E[\tau_s] = \kappa + \phi E[M_s],$$

or at least some weighted average of τ_s. In this matter, however, we are destined for disappointment. Sun and Shapiro (2022) note that the population regression of ΔY_s on M_s does not, in general, generate an average of underlying heterogeneous coefficients like those

[11] When exposure variables are used as instrumental variables, as in Autor, Dorn and Hanson (2013), the resulting exposure design is said to apply *shift-share IV*.

[12] In econometrics and statistics, an estimand is a population quantity to be estimated. This is distinct from an *estimator*, a function of the data used to construct an estimate, and the *estimate* itself, which is the numerical value generated by an estimator applied to data. In the context of a population mean of a random variable, X_i, for instance, the estimand is $E[X_i]$, an estimator of this is the sample mean in a sample of size N, $\frac{\sum_{i=1}^{N} X_i}{N}$, and an estimate of this is the sample mean computed in a particular data set, which takes a numerical value like 42.

described by (10).

To see what the usual exposure design estimates in this context, it's again helpful to difference (16) so as to eliminate state (or country) effects. This yields:

$$\Delta Y_s = \lambda + \kappa M_s + \phi M_s^2 + \nu_s, \tag{17}$$

where ν_s is the differenced residual. Nonlinear models like this generate *marginal effects*, defined as the derivative of the CEF with respect to the regressor of interest. In this case, the marginal effect of exposure is

$$\mu(M_s) = \kappa + 2\phi M_s.$$

Importantly, the average marginal effect of exposure, $E[\mu(M_s)]$, differs from the average exposure effect, $E[\tau_s]$, implied by (10).

The distinction between average marginal effects and average exposure effects suggests a reassuring explanation for the divergence between linear regression estimates and average τ_s. The bivariate slope coefficient associated with a nonlinear CEF averages marginal effects of M_s, rather than random coefficients like τ_s. In other words, a regression of ΔY_s on M_s estimates something like

$$E[\mu(M_s)] = \kappa + 2\phi E[M_s], \tag{18}$$

which, using (10), equals $E[\tau_s] + \phi E[M_s]$. OLS does not quite estimate the average in (18); the OLS estimand is a weighted average marginal effect. However, Angrist and Pischke (2008) note that OLS estimates are typically close to the corresponding average marginal effects from a nonlinear CEF (the formula for OLS weights is repeated here in a footnote).[13]

[13]Ignoring subscripts, denote the dependent variable by Y and a scalar continuously-distributed mean-zero regressor by x; the associated CEF is $E[Y|x] \equiv h(x)$, assumed to be a differentiable function of x. The OLS estimand in this scenario can be written

$$\frac{E[xY]}{E[x^2]} = \frac{\int h'(u)\omega(u)du}{\int \omega(u)du},$$

The upshot of this analysis is that, regardless of whether $E[\mu(M_s)]$ is computed as a weighted or unweighted average, average marginal effects analogous to (18) are twice as sensitive to changes in $E[M_s]$ as the model for τ_s might lead you to expect. Equation (18)—and the analogous OLS estimand—both reflect the fact that when M_s changes, outcomes change by τ_s while τ_s itself also changes. In combination, these changes induce nonlinear exposure effects on outcomes. Given sufficient variation in M_s, the parameters determining μ_s are identified by models like (16). But the divergence between $E[\tau_s]$ and $E[\mu(M_s)]$ does not necessarily make the latter a misleading guide to the population consequences of changes in M_s.

The Nunn and Qian (2011) potato study offers an empirical testbed for this view of heterogeneous effects in an exposure design. As we've noted, the exposure variable in this context (M_s, where s indexes countries) is the share of a country's arable land suitable for potato cultivation, while treatment is this share interacted with a post-1700 dummy. The estimate generated by a constant-effects version of model (9) is 0.81 (with a standard error of 0.10), indicating that a 10 percentage point increase in the share of land suitable for growing potatoes increases population after 1700 by about 8%.[14]

The sample analyzed in Nunn and Qian (2011) includes a diverse set of 130 countries in Europe, Asia, and Africa. The continent in which a country sits is an important source of heterogeneity. In particular, the estimated impact of the arrival of the potato is larger for Europe (0.90) than for Asia and Africa (0.42). And European soils and climate are much more suitable for potatoes: about 52% of European farm land lends itself to potato cultivation compared with only 7% outside Europe. In other words, M_s is much higher in Europe than elsewhere.

What does this exposure-effect heterogeneity imply for the contrast between average

where the limits of integration range over the support of x and weighting function $\omega(u)$ is non-negative. This weighting function is $\omega(u) \equiv (E[x|x > u] - E[x|x < u])P[x < u](1 - P[x < u])$.

[14] Nunn and Qian (2011) uses the log of total land area suitable for potato cultivation as regressor. In keeping with the exposition in this section, exposure is defined here as the land area suitable for potato cultivation as a share of the total land suitable for the cultivation of any foodstuff.

τ_s and average marginal effects? The average of the Europe and non-Europe estimates, weighted by the number of countries in each group, is 0.55, much less than the pooled OLS estimate of 0.81. Interpolating between the values for our two country groups, ϕ is roughly $\frac{.90-.42}{.52-.07} \approx 1.07$. In other words, the effect of potato-positive land on log population increases by 1.07 as the share of potato-positive land rises. The mean share potato-positive in the sample is around 0.19. The approximate average marginal effect of potato suitability in a model allowing an interaction with baseline exposure, therefore, roughly equals

$$E[\tau_s] + \phi \times E[M_s] = 0.55 + 1.07 \times 0.19 \approx 0.75.$$

The pooled OLS estimate of 0.81 is remarkably close to this.

Which is a better guide to the causal effects of potato cultivation on population growth, $E[\tau_s] = 0.55$ or the larger pooled OLS estimate? In a world where τ_s is a fixed attribute of countries unchanged by changing their M_s, the average of τ_s is probably the effect of primary interest. If the fact that τ_s increases with M_s is itself a causal effect, however, then it should be the average marginal effect, $E[\mu(M_s)]$, that you seek. It's especially interesting to juxtapose the two; like the potato, neither effect is fairly said to be unsatisfying.[15]

Doing DD in an Eventful World

Event-study estimates offer a straightforward approach to time-varying treatment effects. But parsimonious DD analysis with a single post-treatment estimate remains appealing, especially when event-study lags proliferate. We might be interested, for instance, in the impact of unilateral divorce reform on divorce rates averaged over all post-treatment years. This offers a compact summary measure of reform impact.

Alas, elegant regression averaging of the sort seen in static DD analysis of exposure

[15] Because the conditions for potato cultivation are likely correlated with other conditions favorable for 18th century growth, Nunn and Qian (2011) focuses on exposure design models with additional controls that yield somewhat smaller potato effects. The potato estimates, it would seem, depend on what else is in the pot when you cook 'em.

designs with cross-sectional heterogeneity needn't carry over to DD scenarios with time-varying treatment effects. A particular concern here, first highlighted by de Chaisemartin and d'Haultfoeille (2020) and Borusyak, Jaravel and Spiess (2024), arises from the fact that simple DD implicitly uses already-treated units as controls.

The challenge presented by dynamic treatment effects is illuminated by a numerical example drawn from de Chaisemartin and d'Haultfoeille (2020), reproduced here as Table 1. This scenario has two states, three periods, and two event-study coefficients. State 1 is treated in period 2, while state 2 is treated in period 3. Untreated potential outcomes $Y_{st}(0)$ are fixed at zero, while the treatment effect equals one in the initial period of treatment and four thereafter. Causal effects are the same in each state (a subsequent example relaxes this), but the impact of the first post-treatment lag (indicated by ΔD_{st-1}) is observed only in period 3 in state 1.

Table 1
Time-varying treatment effects

	State 1			State 2		
Period:	1	2	3	1	2	3
treatment	0	1	1	0	0	1
$Y_{st}(0)$	0	0	0	0	0	0
$Y_{st}(1)$	-	1	4	-	-	1
outcome	0	1	4	0	0	1

Although causal effects equal either 1 or 4 in this example, static DD regression applied to the data in Table 1 yields an estimate of -0.5. This estimate averages two partial DD estimates, one for the first two periods and one for the last two periods. In the first two periods, state 1 is treated and state 2 is control. The changes in outcomes from period 1 to period 2 are $1 - 0 = 1$ for state 1 and $0 - 0 = 0$ for state 2, generating a sensible DD estimate of 1. In the last two periods, state 2 is treated and state 1 is control (since the latter's treatment status is unchanged in periods 2 and 3). The resulting DD estimate is $(1 - 0) - (4 - 1) = -2$. DD regression averages these two to generate a pooled effect of -0.5.

The problem here is that state 1 fails as a control group in periods 2 and 3. Since

treatment impact increases over time, state 1 outcomes increase even though this state's treatment status is fixed in periods 2 and 3. Thus, the putative control trend generated by state 1 diverges from that for $Y_{2t}(0)$, which is fixed at zero in all three periods. This example highlights the value of never-treated states, which are more likely than treated states to capture counterfactual trends in the absence of treatment.

The absence of never-treated states notwithstanding, an event-study model allowing for distinct τ_0 and τ_1 resolves the averaging failure illustrated by Table 1. The first partial DD estimate comparing outcome growth from period 1 to 2 in state 1 (treated) with contemporaneous growth in state 2 gives an estimate of τ_0, the treatment effect at event time 0 (i.e., the year when treatment turns on).

Identification of the first lag, τ_1, is a little harder to see. Counting parameters and observations suggests this effect is identified: we have 6 observations to identify a constant, a state effect, two year effects, as well as τ_0 and τ_1. Borusyak, Jaravel and Spiess (2024) shows that in an example with this structure, the estimator for τ_1 can be written

$$\tau_1 = \{(Y_{13} - Y_{11}) - (Y_{23} - Y_{21})\} + \{(Y_{12} - Y_{11}) - (Y_{22} - Y_{21})\}. \tag{19}$$

To understand this formula, start with the first term in braces, which subtracts state 2 growth in periods 1 to 3 from state 1 growth in periods 1 to 3. This difference-in-differences (equal to 3) eliminates any common trends but also subtracts the treatment effect in state 2 (treated in period 3), leaving us with $\tau_1 - \tau_0$. We therefore obtain τ_1 by adding τ_0 to this. This is obtained from the period 2 versus period 1 DD contrast, which isolates τ_0 for state 1. This effect is 1, which leads us to the correct τ_1 value of 4. Note, however, that we extrapolate τ_0 across states to make this work.

The event-study estimates of unilateral divorce effects plotted in Figure 2 show interesting dynamics. How far off is static DD from an average of these? The average post-reform effect is roughly -0.28, while the corresponding static DD estimate of -0.22 is reasonably close

to this. The fact that the divorce panel includes a fair number of never-treated states likely contributes to this coincidence of findings.

What of cross-sectional heterogeneity in a staggered adoption design? States with higher divorce rates in the 1960s or younger populations, for instance, may have experienced larger effects of unilateral divorce. If so, event-study estimates such as those based on equation (7) need not be a weighted average of underlying state-specific time-varying effects.

Table 2 illustrates a scenario with cross-sectional heterogeneity. In this case, the treatment effect in state 1 is fixed at one, while the effect in state 2 is fixed at four. In contrast to Table 1, here the static estimate computed using two states is more appealing: static DD generates a sensible estimate of 2.5, midway between one and four. Event-study lags, however, are $\tau_0 = 1$ on impact and, applying (19), $\tau_1 = -2$ one period later.

Table 2
Cross-section variation in dynamic treatment effects

	State 1			State 2		
Period:	1	2	3	1	2	3
treatment	0	1	1	0	0	1
$Y_{st}(0)$	0	0	0	0	0	0
$Y_{st}(1)$	-	1	1	-	-	4
outcome	0	1	1	0	0	4

Why does the event-study model fail to yield something sensible in this case? In contrast with the Table 1 example, extrapolation of treatment effects across states is unwarranted in Table 2. Specifically, $\tau_0 = 1$ in state 1 and $\tau_0 = 4$ in state 2. Thus, the first DD term in (19) does not estimate $\tau_1 - \tau_0$ for either state. Rather it contrasts τ_1 in state 1 with τ_0 in state 2, a difference of -3. Adding the correct estimate of $\tau_0 = 1$ for state 1 doesn't fix the problem generated by cross-state differences in impact.

This example shows how cross-sectional heterogeneity invalidates the extrapolation across states inherent in event-study identification strategies. Sun and Abraham (2021), the first to discuss this problem, notes also that the bias from cross-sectional heterogeneity may likewise compromise event-study tests for pretrends.

Luckily, Borusyak, Jaravel and Spiess (2024) offers a home remedy for staggered models with both dynamic and cross-sectional heterogeneity in treatment effects. The BJS estimator proceeds in two steps. The first estimates time and state effects using the sample of *untreated* observations only. These estimates are then applied to treated observations to impute the counterfactual means implied by the DD parallel trends assumption. The difference between treated outcomes and the imputed counterfactual gives an estimated effect for each treated observation. A final step averages these.

The BJS estimator is undefined for periods when all states are treated since no counterfactual time effects can then be constructed (BJS does not produce an estimate for period 3 in Table 1). In other words, BJS omits effects for states and periods where clean comparisons are unavailable.[16]

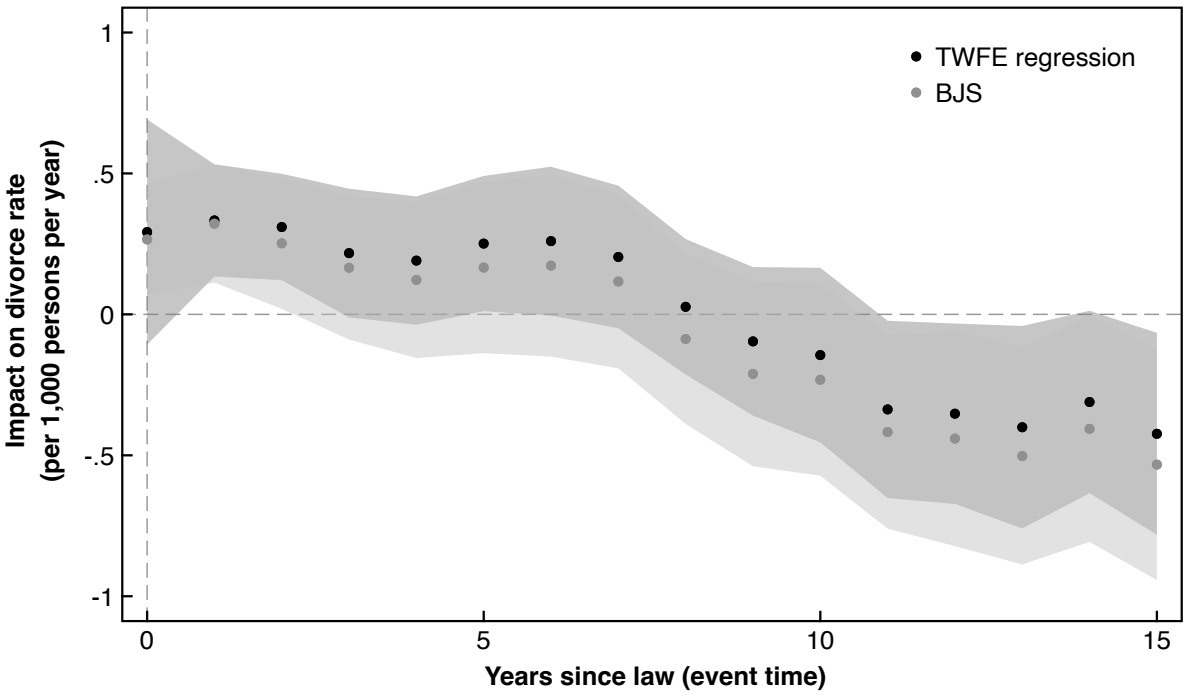
Figure 4 compares BJS imputation estimates of unilateral divorce effects with regression-based event-study estimates. Perhaps surprisingly, the two estimators generate similar results. Both show, for instance, marginally significant positive effects of unilateral divorce early on, followed by marginally significant negative effects farther out. There doesn't seem to be enough cross-sectional heterogeneity in the impact of unilateral divorce for the extrapolation problem highlighted in our numerical example to matter.

Importantly, however, cross-state heterogeneity can matter for reasons besides the implicit extrapolation highlighted in our examples. In a staggered adoption design, the set of states contributing to event-study estimates differs for different leads and lags. As the set of available states changes, cross-state differences in impact may generate illusory dynamics.

Table 2 illustrates this. Treatment effects are fixed over time in this example. In state 1, $\tau_0 = \tau_1 = 1$ and in state 2 $\tau_0 = 4$. Average $\tau_0 = 2.5$, while only state 1 contributes to τ_1, so $\tau_1 = 1$. Although impacts are time-invariant, changing state composition generates what looks like declining effects. This is easily remedied, however, by focusing on dynamic effects for a fixed set of states (state 1 in the numerical example) or by limiting analysis to event

[16] In the same spirit, a procedure introduced in de Chaisemartin and d'Haultfoeille (2020) computes a set of valid 2x2 DD contrasts for the effect of interest and aggregates these.

Figure 4
Regression and BJS estimates of unilateral divorce effects



Notes: This figure compares TWFE regression estimates with event-study estimates computed as described in BJS. Estimates are computed using data from 1958-1998.

times where all states contribute ($j = 0$ in the example).

Time and state heterogeneity may indeed matter for event-study estimates. But heterogeneity need not be fatal. Although one empirical example does not a theorem make, our results highlight the remarkable robustness of event-study regression estimates in an example with strong dynamics and a good ex ante case for cross-sectional differences in impact. Event-study regression estimates allow for dynamic effects while imposing constant effects across states. The BJS estimator offers a simple remedy for analysts worried about cross-sectional heterogeneity in models for dynamic effects. In the divorce example, BJS estimates differ little from event-study regression estimates of the same dynamic effects.

3 Pondering Pretrends

Event-study lags provide a rich picture of treatment effect dynamics, while event-study leads help us assess the parallel trends assumption. Such assessments are increasingly seen as essential to any event-study story. Yet, examination of the size and significance of estimated event-study leads is a pretest, and pretesting is statistically perilous. Pretesting problems often manifest in misleading statistical inference for the post-test estimates of interest (Leeb and Pötscher, 2005). Most importantly for our event-study agenda, pretests for parallel trends may exacerbate rather than mitigate bias (Roth, 2022).

This section evaluates pretesting pros and cons in the context of event-study regression models like (7). We conclude that, by flagging at least some models with divergent trends, the bias-mitigation benefits of pretesting are likely to outweigh the risks.

Tough Talk about Standard Errors

A prerequisite for our pretesting pitch is a discussion of the thorny matter of DD inference, i.e., standard errors. The thorns here grow out of the fact that DD often involves data with a time dimension. One observation in a time series, whether for states, countries, people, or firms, is likely to be highly correlated with the next. Bertrand, Duflo and Mullainathan (2004) shows that standard errors for DD treatment effects computed without accounting for such serial correlation are likely to be too small. Empiricists working with state-year panels therefore cluster standard errors on state (or on whatever cross-sectional unit is relevant for a given DD design).

Bertrand, Duflo and Mullainathan (2004) focuses on a single treatment effect while event-study models estimate many coefficients. This difference in modeling strategies matters for inference. Clustered standard errors are typically biased downwards in small samples, often badly so. Our divorce analysis uses 48 states, which may seem like enough clusters for clustered inference to work. This is indeed the case for DD estimates of a static treatment effect.

But many event-study regression estimates rely on treatment switches that are switched on for only a few observations at a time.

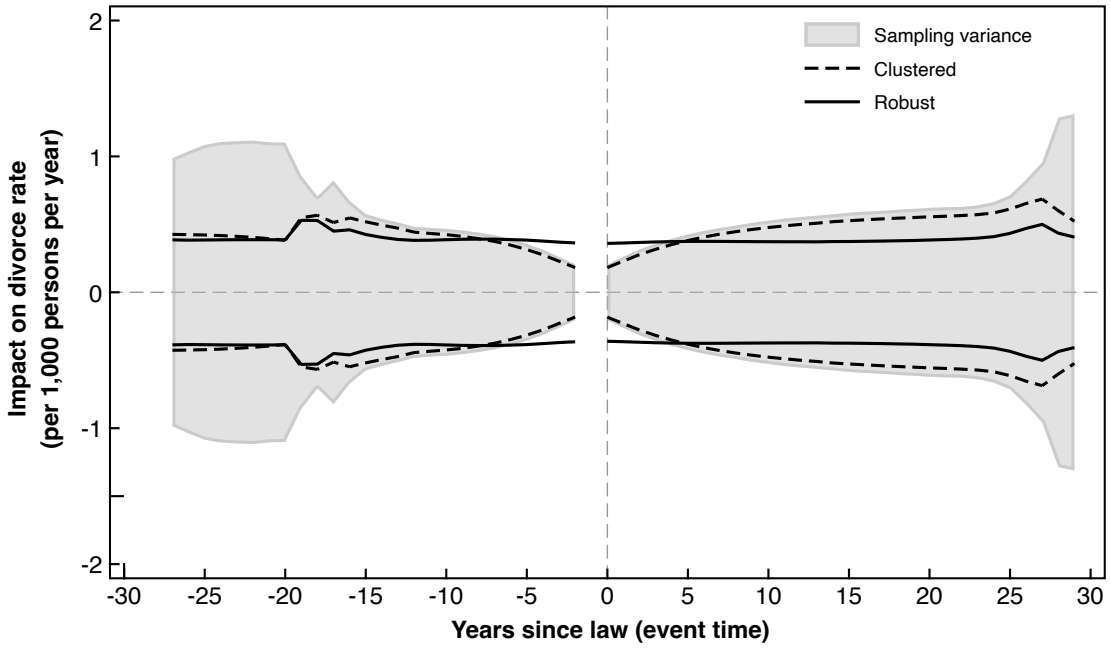
The inference problem here is that, in a staggered design, only a few states are observed for extended pre-treatment or post-treatment periods. This gives these states high *leverage* for estimates of long leads and lags, meaning their data matters greatly for the corresponding estimates. Although not obvious, it's possible to show that sample residuals generally have less variance than the corresponding population residuals and that this variance shortfall increases as leverage increases.[17] This imparts a downward bias in clustered standard error estimates that rely on few states. The upshot is that staggered event-study designs pose a particular problem for clustered standard errors, and especially so for estimates of effects at long leads and lags.

A simulation based on the divorce dataset captures the bias in robust and clustered standard errors. States, time periods, and treatment dates in the simulation design mirror those seen in the original divorce data. Simulated divorce rates follow an AR(2) process with residuals drawn from a normal distribution added to the state and time effects estimated in the original data. AR(2) parameters are set to the estimates from an AR(2) fit to original-sample residuals. Unilateral divorce has no effect in the simulations, so the simulation produces the sampling distribution under the null hypothesis of zero treatment effects (and no pretrends). The appendix gives other simulation details.

Figure 5 plots confidence intervals computed using the average over simulation draws of robust and clustered standard errors, as well as the Monte Carlo standard deviation of the simulation estimates. For leads up to about -15, confidence intervals based on clustered standard errors track the actual sampling distribution reasonably well. Intervals based on clustered standard errors also do better than intervals using unclustered robust standard

[17] See, e.g., Section 8.1 in Angrist and Pischke (2008). In the context of an OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$, where X is a an $N \times K$ regressor matrix with rows x_i', the difference between $\hat{\beta}$ and the OLS estimator omitting row i is $\frac{(X'X)^{-1}x_i e_i}{1-h_{ii}}$, where e_i is the ith residual and $h_{ii} = x_i'(X'X)^{-1}x_i$, is the leverage of observation i. In a saturated model, the leverage of observation i is the reciprocal of the size of the regressor-defined group to which i belongs. The highest leverage values here are for South Dakota and Wyoming, the last two adopters, in years with long leads switched on.

Figure 5
Simulated event-study confidence intervals
in the unilateral divorce design



Notes: This figure plots event-study estimates and 95% confidence intervals for estimated unilateral divorce effects in simulated data constructed to mimic the 1958-98 unilateral divorce sample. The simulation is detailed in the appendix. Grey bands mark intervals constructed using simulation standard deviations; other intervals use average-across-simulations clustered and robust standard errors. Simulations use 25,000 replications.

errors. Both types of standard errors are, for the most part, too small; but for moderate lags and leads, the clustered standard errors are not off by much. Paralleling the pattern of bias in estimated standard errors, confidence interval coverage declines steeply at longer leads and lags.

To make this problem concrete, observe that the last state to adopt unilateral divorce in our data is South Dakota, which went unilateral in 1985. The previous change was in Wyoming in 1977. The eight farthest-ahead lead coefficients are therefore estimated from South Dakota alone. The resulting imprecision can be seen in Figure 5 in the fanning out of sampling variance at leads longer than 15 and, even more so, for leads longer than 20. Yet, clustered standard errors fail to reflect declining precision as a function of lead length.

Binning long leads and lags reduces the risk that an analyst makes too much of imprecisely estimated coefficients with misleadingly small estimated standard errors (simulations of estimates based on (8) bear this out). This consideration leads us to bin estimates at +/-15 in the pretesting analysis below. Happily, this leaves us with a horizon long enough for interesting dynamics to emerge.

Clustered standard errors are sometimes misleading, even for shorter leads. Cameron, Gelbach and Miller (2008) shows that inference using a wild block bootstrap improves on conventional clustering. A wild bootstrap for regression estimates with regressor vector X_i starts by computing estimates, $\hat{\beta}$, and estimated residuals, $\hat{e}_i$. Each replication r of the wild bootstrap retains the original values of the X_i but creates a new dependent variable according to

$$Y_i^r = X_i'\hat{\beta} + w_i\hat{e}_i,$$

where w_i equals $+1$ or -1 with probability 0.5. Each wild bootstrap replication regresses Y_i^r on X_i to obtain $\hat{\beta}^r$. This estimate is then used to compute the test statistic of interest, most often a t-statistic. This is constructed by dividing $\hat{\beta}^r - \hat{\beta}$ by the clustered standard error obtained in replication r. The bootstrap then uses R replications of the t-statistic constructed in this manner to generate an empirically-grounded null distribution rather than relying on asymptotic theory. Specifically, the null distribution delivers a set of critical values that can be used for testing, but it does not deliver confidence intervals (though they can sometimes be re-engineered). A wild *block* bootstrap fixes w_i within each cluster (such clustered resampling makes this bootstrap "block"). Our pretesting simulation includes tests computed using wild block bootstrap critical values as an alternative to clustered standard errors with the usual normal approximation.[18]

We conclude this digression with a reminder that when testing event-study leads for statistically significant pretrends one at a time, the tester risks over-rejection. As in any multiple-testing problem, the odds of a Type I error (rejecting a true null hypothesis) increase

[18]Chapter 8 in Angrist and Pischke (2008) discusses bootstrap inference in more detail.

with the number of tests under consideration. A straightforward solution here is to test all pretrend coefficients jointly using an F or chi-square test. But this is tricky in practice since clustered joint test statistics tend to exacerbate the downward bias in clustered standard errors. In particular, Pustejovsky and Tipton (2018) and MacKinnon, Nielsen and Webb (2023) warn that the downward bias in F-statistics based on a clustered covariance matrix grows with the number of restrictions tested.[19]

Rather than trying to wrangle a reliable clustered F, we can stick with individual t-tests while adjusting critical values to control the family-wise error rate (FWER). This is the probability that you make at least one Type I error in a set of tests. *Sup-t critical values* offer such an adjustment. These are obtained as follows:

- Estimate a model with the $m-1$ leads you'd like to test; save the estimated covariance matrix for these estimates.
- Draw from a multivariate $(m-1)$-dimensional normal distribution with mean zero and correlation matrix corresponding to the covariance matrix for these estimated leads; this simulates the distribution of estimates under the null hypothesis that all leads are zero.
- For each replication, indexed by r, calculate t^r_{max}, the maximum of the absolute values of the normal variates from draw r.
- The sup-t critical value for a test with size α is the $1-\alpha$ quantile of the distribution of t^r_{max}.

Sup-t critical values are also used to construct *uniform* (also called *simultaneous*) confidence bands that cover an entire set of parameters (like the set of leads) with a minimum probability, say 95%. A uniform confidence band of level α guarantees that the coverage probability

[19]Intuitively, this is because a test of multiple restrictions involves the inverse of the relevant estimated covariance matrix. Cluster bias affects not just the diagonal but also the many off-diagonal elements of this matrix, and so the overall bias in a joint test quickly adds up.

for (say) a set of leads is $1-\alpha$. This is equivalent to ensuring that the FWER for the entire set is controlled at level α.

Uniform sup-t confidence bands have the virtue that, like the usual pointwise intervals, they're easily plotted, while F-based confidence regions are hard to visualize with more than two parameters (with two parameters, the relevant region is an ellipse). For correlated coefficient estimates, sup-t generates the narrowest uniform confidence bands with correct asymptotic coverage relative to Bonferroni and other widely-used approaches to multiple-testing problems (Montiel Olea and Plagborg-Møller, 2019).[20] The bootstrap, wild or otherwise, delivers sup-t p-values by creating the maximum of the t-statistics, t_{max}, (e.g., for all the leads) in each replication.

Pretrends Pretests: The Gripes of Roth

A pretrends pretest sensibly asks whether estimated leads are statistically significantly different from zero. This offers a check on the parallel trends assumption. When treatment and control outcomes indeed follow parallel trends, such pretesting imparts no bias in downstream estimates of treatment effects. It's worrying, however, that when treatment and control trends diverge for reasons unrelated to treatment, event-study estimates that pass a pretrends pretest might be *especially* misleading. Roth (2022) describes both of these features of the pretrends pretesting problem.[21]

The pretesting payoff is highest when a test correctly identifies DD research designs that fail parallel trends. As Roth (2022) explains, pretrends pretests often have low power, missing many violations. This is perhaps unsurprising given the inference challenges detailed above. Less intuitively, however, studies with truly divergent trends that nevertheless slip through

[20] When pretrend coefficients are uncorrelated, the sup-t procedure results in critical values similar to those derived using a Bonferroni correction (Montiel Olea and Plagborg-Møller, 2019). The Bonferroni correction obtains a p-value of α for a set of k tests by requiring that at least one test reject with a p-value less than $\frac{\alpha}{k}$. For instance, to test whether either of two leads is significant using a procedure with a 10% Type I error rate, Bonferroni requires that at least one of the associated test statistics exceed the critical value for a 5% test.

[21] The first result presumes the pretest in question is two-sided, like a two-sided t-test or an F test.

a pretrends pretest might generate the most misleading estimated treatment effects. This bias arises as a result of the fact that the TWFE estimates used for pretrends pretesting are correlated with the treatment effect estimates of primary interest.

In the big picture, the question of whether pretesting benefits outweigh the corresponding costs in terms of bias depends on the incidence of divergent trends, the bias of screened studies that pass a pretest, and pretest power. To characterize this dependence, we imagine a pool of possible event studies you might undertake, some of which satisfy parallel trends and some that don't.

Let θ denote the fraction of potential studies with divergent trends (indicated by dummy variable B) and assume this divergence results in estimated treatment effects that are biased by the amount $\delta > 0$. The unconditional bias of an estimator that ignores pretrends can then be written:

$$E[b] = \delta P[B = 1] = \delta\theta, \tag{20}$$

where b is the bias in estimated treatment effects and expectations are taken over studies chosen at random from the pool of possible studies.

In what sense is the bias of a study random? To be concrete, for our purposes, a study is defined by a research question, a sample, and an event-study research design. We usually think of bias as the expected difference between a parameter estimate and the corresponding target parameter. In this view, bias is a parameter, and, like other parameters, it is non-stochastic. Here, however, we imagine the pool of possible studies as defining a sampling frame that we draw from at random. Fraction θ of these studies yield biased estimates due to divergent trends, but we don't know which studies are biased.

Imagine pretesting to sort biased weeds from unbiased flowers, harvesting only those studies that pass the pretest. Studies that pass a pretest are either unbiased or biased. For any given study, test results are random, where randomness is determined by the usual within-study sampling distribution of estimates. Tests that fail to reject the null hypothesis of no pretrends when trends indeed diverge result in a Type II error. Call the bias of studies

mislabeled as unbiased δ_s ("s" for "screened"). This can be above or below the bias due to pretrends in the absence of pretesting, δ, though Roth (2022) argues we should presume $\delta_s > \delta$. One minus the probability of a Type II error is the power of the pretest. Call this π, a number between 0 and 1. Let α be the corresponding test size, i.e., the pretest's probability of Type I error. The costs and benefits of pretesting are determined by π, α, θ, and the bias of screened and unscreened studies.

Let $T_P = 1$ indicate pretest rejection, with $T_P = 0$ otherwise. We aim to compare the magnitude of screened bias $E[b|T_P = 0]$ with unscreened bias, $E[b]$ (which is positive by construction). As a first step, note that

$$E[b|T_P = 0] = \delta_s \times P[B = 1|T_P = 0],$$

since, as Roth (2022) shows, pretesting imparts no bias in estimates for studies without pretrends, i.e., studies for which $B = 0$. Applying Bayes' Rule, we have that

$$\begin{aligned} P[B = 1|T_P = 0] &= \frac{P[T_P = 0|B = 1]P[B = 1]}{P[T_P = 0]} \\ &= \frac{(1 - \pi)\theta}{(1 - \pi)\theta + (1 - \alpha)(1 - \theta)} \end{aligned}$$

Using this to contrast screened bias with (20) shows that pretesting reduces bias magnitude when

$$|\delta_s| \left[\frac{1 - \pi}{(1 - \pi)\theta + (1 - \alpha)(1 - \theta)} \right] < \delta \tag{21}$$

Pretesting is always bias-reducing when $\pi = 1$ and generally more likely to be beneficial as power increases. Suppose power is 0.5 (the Roth (2022) benchmark) and size is nominal, a small number like .01 or .05. Then pretesting is likely to pay when

$$\frac{|\delta_s|}{2 - \theta} < \delta. \tag{22}$$

For a given θ, this ratio depends on the relative bias of screened and unscreened studies, which can go either way. If $|\delta_s| \leq \delta$, screening always pays since $2-\theta$ is at least 1. Otherwise, the payoff to screening depends on biased study prevalence, θ, as well as the relative bias of screened and unscreened studies. Paradoxically, when divergent trends proliferate, the pretesting payoff is less certain. This reflects the fact that the probability an unscreened study is biased increases one-for-one with θ while the probability a screened study is biased is a convex function of θ (as can be seen by differentiating the left-hand side of (22)).

Pretest performance might be improved by splitting the sample of interest, using half for testing and half for estimation. In this spirit, Borusyak, Jaravel and Spiess (2024) suggests pre-treatment leads be estimated in a sample limited to untreated observations only. When the test passes, the BJS split-sample imputation estimator computes event-study lags as described above, implicitly setting all leads to zero. Recall that BJS is an imputation estimator because state and time effects are estimated using pretreatment data and then used to impute counterfactuals for the post-treatment period. BJS shows that lead and lag estimates computed in this manner are uncorrelated, removing pretest bias in screened studies.

The BJS imputation estimator has its virtues, especially in a world of heterogeneous treatment effects (since BJS ensures convex weighting of these). But the case for fancy event-study estimation footwork is decidedly mixed. Sample splitting à la BJS is meant to break the correlation between test statistics and estimators, but some of the studies that pass a pretest are still biased. In such studies, the BJS estimator can have more bias from pretrends than event-study estimates based on regression models like (7). To see why, suppose biased studies are biased by virtue of linear trends in untreated potential outcomes that arise only in treated states. BJS omits leads when the pretest passes and therefore compares post-treatment outcomes to average outcome levels for the entire pre-treatment period. An event-study regression estimator, by contrast, compares post-treatment outcomes to outcomes at event time -1. Given linear trend divergence, more distant pre-treatment

outcomes make for a more misleading baseline than does the outcome at event time -1 (Roth, 2024, details this concern).

A simulation shows how these things shake out for pretesting in a relevant empirical context. The simulation is built from the divorce-inspired setup deployed for the exploration of standard errors in Figure 5. Treatment effects are set to zero as before, but half of the simulated studies have divergent trends (i.e., $\theta = P[B = 1] = 0.5$). These are specified as a common linear trend that characterizes untreated potential outcomes in treated states. As in Roth (2022), the trend parameter is set so that power equals 0.5 when $B = 1$ (for a pretest using standard TWFE regression with clustered standard errors and sup-t critical values).

Table 3
Pretrend pretests in the simulated unilateral divorce design: power and size

	Size (1)	Power (2)	Power–Size (3)
Baseline: Regression clustered sup-t	0.146	0.500	0.353
Regression robust sup-t	0.155	0.607	0.452
Regression clustered	0.429	0.764	0.334
Regression wild block bootstrap sup-t	0.049	0.310	0.261
BJS clustered sup-t	0.156	0.495	0.339

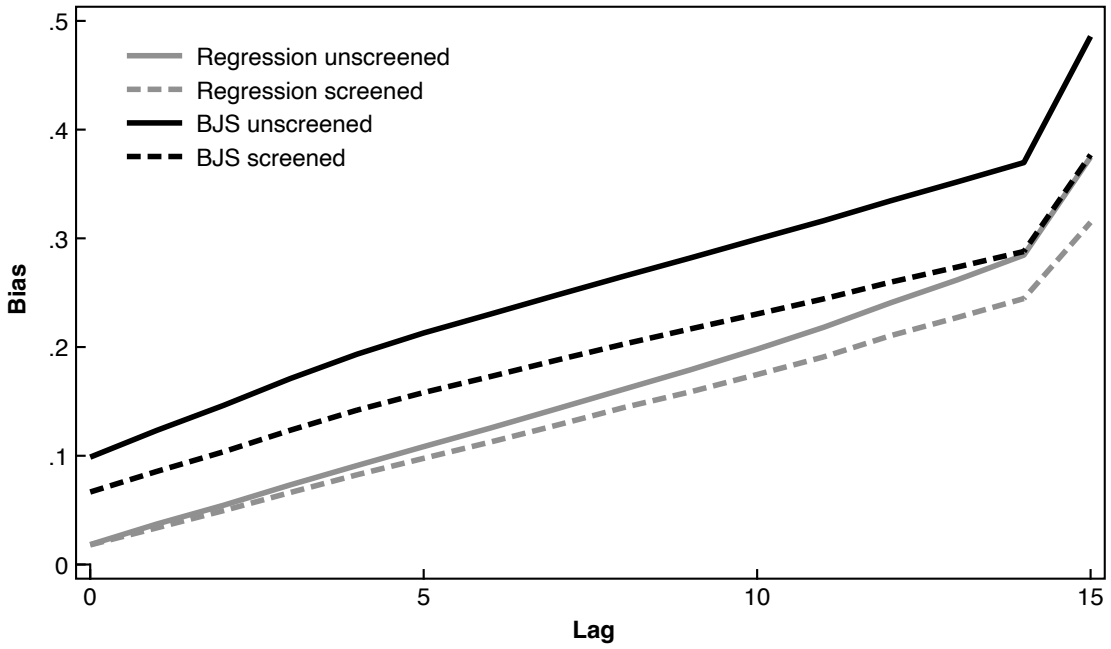
Notes: This table reports power and size for a nominal 5% pretrends pretest using simulated data. Simulations draw from the unilateral divorce design detailed in the appendix, with 25,000 replications. Treated states follow divergent linear trends in half of the studies from which simulated data are sampled.

Our empirical analysis of pretesting problems starts with test size and power. Pretests are applied to the 14 event-study leads generated by (8), binned at +/-15 (recall that the -1 lead is omitted). Reflecting the inference challenges discussed at the beginning of this section, only one of the test statistics in Table 3 has the correct size. Results in the first row of the table show that a clustered t-test with sup-t critical values has a Type-I error rate around 15% (this is the probability the pretest rejects when $B = 0$). Tests using robust standard errors without clustering have similar size and greater power (the probability of rejection when $B = 1$), but a clustered t-test without allowance for multiple testing rejects

more than 40% of the time under the null. The wild block bootstrap generates a test with near-nominal size, but power here is low. Performance of the BJS split-sample pretest using sup-t differs little from that of the corresponding TWFE regression-based test.

Power minus size provides a useful summary of test quality, at least as far as bias detection goes. This ranges from around 0.26 to 0.45.[22] Sup-t without clustering (i.e., robust but unclustered standard errors) leads the pack at 0.45. In this day and age, however, it may be hard to convince referees and other readers that DD standard errors needn't be clustered. Clustered standard errors improve size moderately, reducing power in roughly equal measure, so power minus size falls to around 0.35.

Figure 6
Bias in event-study estimates with and without pretesting
in the unilateral divorce design



Notes: This figure plots estimated event-study lag coefficients averaged over 25,000 simulation draws of the unilateral divorce design detailed in the appendix. In the simulated model, all lags are set to zero.

Figure 6 compares the bias of screened and unscreened estimates of individual lag coef-

[22] Tests that maximize power minus size minimize the sum of Type I and Type II errors. This criterion for test quality appears to originate in the classic DeGroot (1986) statistics text; see Gannon, de Bragança Pereira and Polpo (2019) for details.

ficients computed using event-study regression models and the BJS estimator. Pretests use t-statistics with clustered standard errors and sup-t critical values. In this model, the bias from divergent trends increases with lag length. Pretesting mitigates the bias of event-study regression estimates. While gains from screening are larger for BJS estimates, these are still mostly more biased than screened regression estimates, especially at short lags. This reflects omitted variable bias in flawed studies that nevertheless pass the pretest.

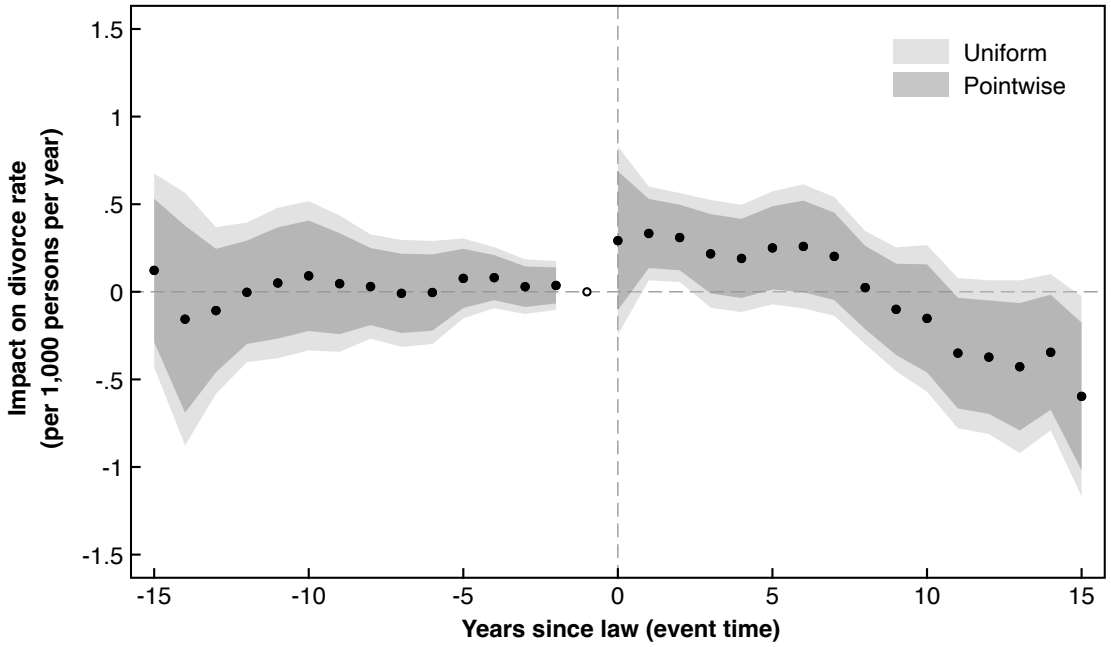
Fruitful Pretests

The results in this section convince us that, in the messy world of event-study estimation, it's worth pretesting estimated leads. The pretesting implicit in plots like Figure 2 supports—but cannot ensure—valid causal inference. Our analysis also suggests that in both estimation and testing, binning poorly-estimated long-horizon leads and lags helps avoid spurious statistical results. Finally, uniform confidence bands based on a sup-t adjustment offer a simple fix for multiple testing problems.

Figure 7 implements these principles for estimates of unilateral divorce effects, binning at +/-15. Confidence bands use clustered standard errors, showing both pointwise (one at a time) and uniform (sup-t) critical values (the latter use critical values computed separately for leads and lags). Estimated leads are close to zero and well inside the tighter pointwise bands. Estimated leads are also much smaller than estimated post-treatment lags. The latter paint a picture of plausible dynamic effects in the treatment period.

Overall, the pretrends picture here seems salutary. It's noteworthy that many estimated lags fall outside of the range covered by uniform bands for the estimated leads. This is especially true for short leads and lags, unsurprisingly, since these are estimated with the greatest precision. Pretesting is therefore powerful enough to reject the null hypothesis of leads as small as the short lags. At the same time, even in this relatively well-provisioned empirical setting with 48 states and 41 periods, estimated long-run event-study lags are noisy and marginally significant at best.

Figure 7
Estimated unilateral divorce effects with pointwise and uniform confidence intervals



Notes: The figure plots event-study estimates of unilateral divorce effects, with the year before adoption set as the reference year. Leads and lags are binned at 15 periods. Confidence intervals use clustered standard errors and critical values of 1.96 (for pointwise intervals) or sup-t critical values (for uniform intervals). The sup-t critical value adjustment is done separately for leads and lags. Estimates are computed using data from 1958-1998.

Some pretesting paths lead in directions other than those we've taken here. Instead of abandoning a study that fails a pretrends pretest, you might forge ahead with a model that includes linear state-specific trends. In our simulations, divergent pretrends are indeed linear, so DD estimates computed with state-specific trends are unbiased. Of course, this needn't work out in real life. Moreover, event-study models with state-specific trends raise further vexing collinearity complications of the sort discussed at the end of Section 1. In such an event-study specification, you might miss this and report estimates of parameters that aren't identified. Synthetic control methods offer an alternative that sidesteps collinearity concerns while allowing for divergent nonlinear trends in event-study designs; see Abadie (2021) for an overview.

4 Dysfunctional Forms

Outside models with parametric trends, DD identification strategies stipulate that changes in average outcomes in the absence of treatment are the same for treated and control units. This parallel trends assumption requires commitment to a specific transformation of the outcome: when trends run parallel for logs, they diverge for levels, and vice versa. Our DD discussion concludes with an examination of efforts to weaken commitment to functional form.

Consider the question of how minimum wage laws affect workers' incomes. Economists have long argued over the distributional consequences of a wage floor.[23] The fact that U.S. states set minimum wages above the federal minimum makes this a classic DD question. A static DD analysis of minimum wage effects on income might examine the effect of the prevailing minimum wage (denoted W_{st}) on the *log* income of worker i in state s at time t, denoted $\ln Y_{ist}$.

With many states and years, minimum wage effects on log income can be estimated using a TWFE model like

$$\ln Y_{ist} = \tau \ln W_{st} + \gamma_s + \lambda_t + \eta_{ist}, \tag{23}$$

where γ_s and λ_t are state and year effects and η_{ist} is a residual that's mean-independent of state and year. A causal interpretation of the resulting estimates is predicated on parallel trends for log income. But parallel trends for log income implies trends for income levels that are proportional rather than parallel. That is, this model implies

$$Y_{ist} = W_{st}^{\tau} e^{\gamma_s} e^{\lambda_t} \nu_{ist},$$

where ν_{ist} is a proportional error term whose log is mean-independent of state and year.

Functional-form dependence of this sort is troubling for two reasons. First, DD analysts rarely have clear institutional or theoretical reasons to prefer a particular dependent variable

[23] Research on this question dates back at least to Gramlich (1976).

transformation. Models for log income are appealing because τ can be interpreted as the percent change in wages due to a higher minimum wage. Effects in percentage terms are unaffected by inflation, and proportional models seem a good fit for non-negative dependent variables like earnings. But these dependent variables can also be zero, and we have not yet learned how to log that.[24] Second, some research questions inherently involve a variety of transformations.

The latter concern is highlighted by research using state minimum wage changes to estimate minimum wage effects on the *distribution* of income as well as on mean income. Specifically, Dube (2019) asks how minimum wage hikes affect the family income distribution. This question is answered by making the dependent variables in DD analysis a set of indicators of whether family income falls below multiples of the official federal poverty threshold (FPT). Dube (2019) also examines minimum wage effects on family income quantiles like the median and lower quartile.

An analysis of this sort raises the question of when, if ever, minimum wage mavens can reasonably invoke parallel trends for more than one transformation of a given underlying dependent variable. Can parallel trends hold jointly for income quantiles and for a set of indicators for income below cutoffs like the FPT? Meyer (1995) and Athey and Imbens (2006) are among the first to explore possible resolutions of this enduring DD conundrum. More recently, Roth and Sant'Anna (2023) establishes a necessary and sufficient condition—called *parallel distributions*—under which parallel trends applies to all monotonic transformations of a dependent variable.

Parallel distributions, like parallel trends, restricts the evolution of potential outcomes in the absence of treatment over time and across groups. For the minimum wage, a continuous treatment, we assume that untreated means the state minimum wage equals the federal minimum. Potential outcome $Y_{ist}(0)$ denotes the outcome for household i in state s and period t where the minimum wage is set at the federally mandated floor. Note that outcomes

[24] Chen and Roth (2024) critiques recent efforts to deal with this quotidian problem.

here are measured at the level of individual households rather than as state-year averages. This allows us to model the distribution of income within states and years.

For all periods t, and for all non-negative values y, parallel distributions requires:

$$E\left[1\{Y_{ist}(0) \leq y\} - 1\{Y_{ist-1}(0) \leq y\}\right] = E\left[1\{Y_{ict}(0) \leq y\} - 1\{Y_{ict-1}(0) \leq y\}\right] \tag{24}$$

for all states indexed by s and c. Because the mean of an indicator function is a cumulative distribution function (CDF), parallel distributions says that the change in the CDF of potential outcomes from one year to the next is the same in all states. The appendix shows that parallel distributions imposed on microdata implies a TWFE model for state-year averages of any transformation of the outcome.[25]

As a test-bed for the parallel distributions idea, Figure 8 plots event-study estimates of the effects of minimum wage increases on transformations of household income. These transformations, denoted $\tilde{Y}_{ist}$, include indicators for crossing two FPT cutoffs, household income itself (measured in multiples of FPT), and log income. These estimates were computed using the cross-sectional household data analyzed by Dube (2019).[26]

The coefficients plotted in Figure 8 come from a continuous-treatment version of equation (8). Specifically, functions of income for household i in state s and year t are modeled as determined by leads and lags of minimum wage *changes* according to

[25] Fernández-Val et al. (2024) introduces a related restriction on the CDF of $Y_{ist}(0)$ dubbed "no interactions." This approach has a TWFE-type linear function inside a nonlinear model like logit. The resulting model is invariant to monotonic transformations of dependent variables. Unlike parallel distributions, however, this no-interactions assumption doesn't deliver parallel trends in conditional mean outcomes. Abadie (2005) and Callaway, Goodman-Bacon and Sant'Anna (2021) extend parallel trends to cover multi-valued and continuous treatments by assuming that changes in untreated (e.g., no minimum wage change) potential outcomes are mean-independent of the level of treatment.

[26] Family income comes from the 1984-2013 CPS and excludes the elderly; state minimum wages are from the U.S. Department of Labor. As in Dube (2019), models estimated here include measures of family size and composition as additional controls. Also, as in Dube (2019), before taking logs, zeros are replaced with one-half times the smallest nonzero income observed in a given state-year cell. As noted above, this fudge can be consequential.

$$\tilde{Y}_{ist} = \sum_{j=-3}^{-2} \tau_j \Delta \ln W_{st-j} + \sum_{j=0}^{2} \tau_j \Delta \ln W_{st-j} \tag{25}$$
$$- \tau_{\leq -4} \ln W_{st+3} + \tau_{\geq 3} \ln W_{st-3} + \gamma_s + \lambda_t + \eta_{ist}.$$

Terms of the form $\Delta \ln W_{st-j}$ denote leads and lags of year-to-year changes in the log minimum wage. Coefficients $\tau_{\leq -4}, \tau_{-3}, \ldots, \tau_2, \tau_{\geq 3}$ capture the cumulative effects of minimum wage changes that occur at event time zero. The model is binned, with a single term ($\tau_{\leq -4}$) for effects of future minimum wage changes 4 or more years ahead and a single term ($\tau_{\geq 3}$) for effects of minimum wage changes 3 or more years ago. As before, τ_{-1} is normalized to zero.[27]

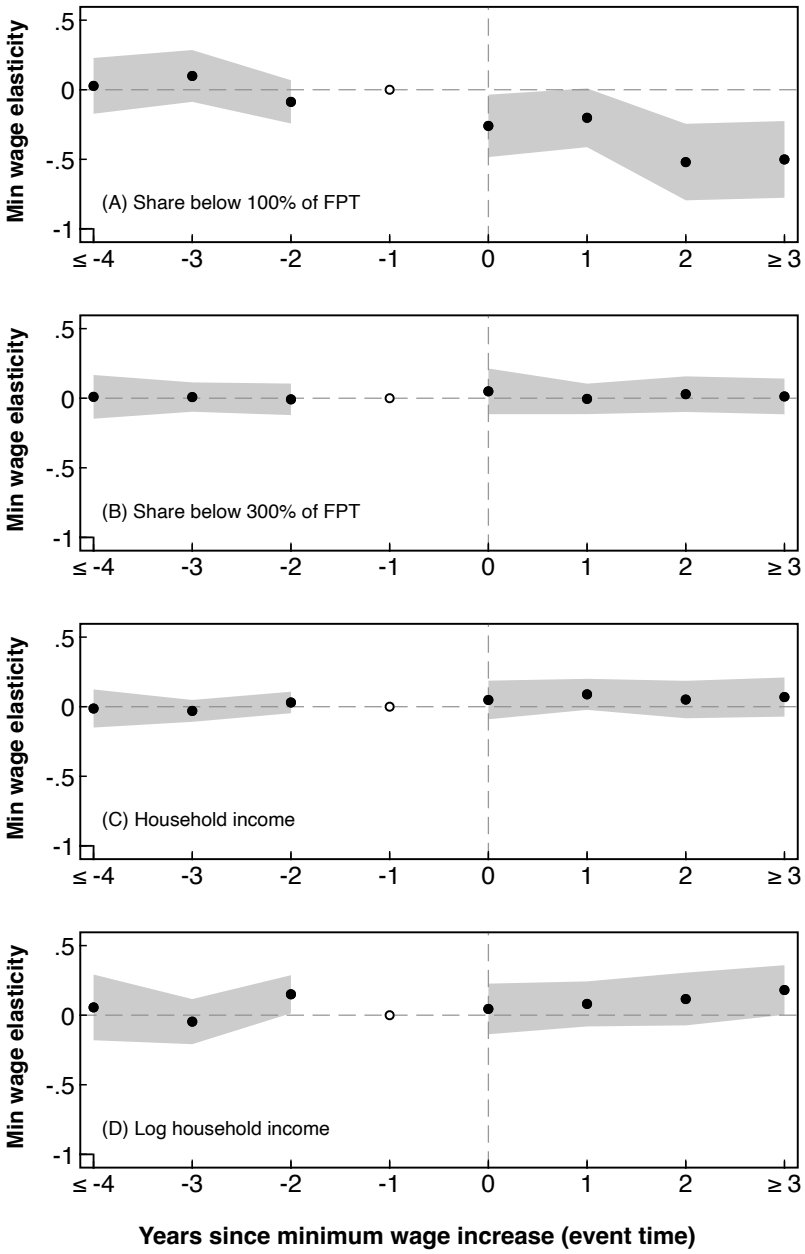
As in the divorce model, estimated leads in (25) serve as a check on the parallel trends assumption. Provided pre-change trends are parallel and outcomes do not change in anticipation of minimum wage changes, these should be zero. Coefficients multiplying $\ln W_{st+3}$ and $\ln W_{st-3}$ sum effects at horizons of 4 or higher for leads and 3 or higher for lags.[28] Finally, as in Dube (2019), estimated coefficients are divided by the dependent variable mean; this scaling makes the reported estimates elasticities.

The estimates plotted in the top panel of Figure 8 show that in years following a minimum wage increase, the share of households with income below FPT declines, an effect that grows over time. At the same time, the next panel down, which plots effects on the likelihood of income falling below 3 times FPT, suggests the minimum wage has no effect on the income distribution around this higher cutoff. The remaining two panels show little effect on average income, whether income is measured in levels or logs. The figure, therefore, supports the claim that minimum wage effects on income are concentrated at the bottom of the income distribution. Remarkably, Figure 8 shows little evidence of pretrends for any

[27] This setup mirrors the continuous-treatment event-study model described in Freyaldenhoven et al. (forthcoming).

[28] To see why this works, note that $\ln W_{st+3} = \ln W_{sT} - \sum_{j \leq -4} \Delta \ln W_{st-j}$, where T is the final period in the data, and $\ln W_{st-3} = \ln W_{s0} + \sum_{j \geq 3} \Delta \ln W_{st-j}$, where the first period in the data is $t = 0$. In other words, up to a constant for each state, $\ln W_{st-3}$ is equal to the sum of the first differences of $\ln W_{st}$ and $\ln W_{st+3}$ is equal to minus the binned sum, analogous to the binned terms in equation (8).

Figure 8
Estimated minimum wage effects on household income



Notes: This figure plots binned event-study estimates (reported as elasticities) and uniform 95-percent confidence intervals clustered by state for minimum wage effects on household income. The upper two panels show effects on indicators of family income below one and three times FPT, respectively. Effects one year before the minimum wage change are normalized to 0. Controls include state effects, Census division-by-year effects, state-specific indicators for each Great Recession year, state-level covariates (GDP per capita, EITC supplement, and unemployment rate), and individual demographic controls (a quartic in age as well as dummies for race, marital status, family size, number of children, education level, Hispanic status, and gender). Regressions are weighted by March CPS person weights. The confidence intervals are based on sup-t critical values, which ensure nominal simultaneous 95-percent coverage separately for leads and lags. Data are from the Dube (2019) replication package.

dependent variable. This would therefore appear to be a case of transformation-invariant parallel trends.

Although parallel distributions appears to hold in the Dube (2019) data, this robustness is a poor fit for the underlying theory. Roth and Sant'Anna (2023) shows that assumption (24) implies that the population of interest can be characterized as containing a mixture of two types of households, one for which the distribution of $Y_{ist}(0)$ is independent of state (but whose outcomes may be time-varying) and another for which the distribution of $Y_{ist}(0)$ is fixed over time (but might differ across states).[29] Yet, worker income everywhere is surely subject to time effects of the sort generated by business cycles, while states have persistently different income distributions. It's hard to imagine a world in which households experience income variation subject either to state effects or time effects, but not both.

What, then, explains the transformation-robustness apparent in Figure 8? Importantly, the models used to generate the estimates in Figure 8 (and in Dube, 2019) adjust for covariates beyond state and year effects. These covariates include time-varying controls such as GDP per capita, the aggregate unemployment rate, and changes in state-level EITC policy.

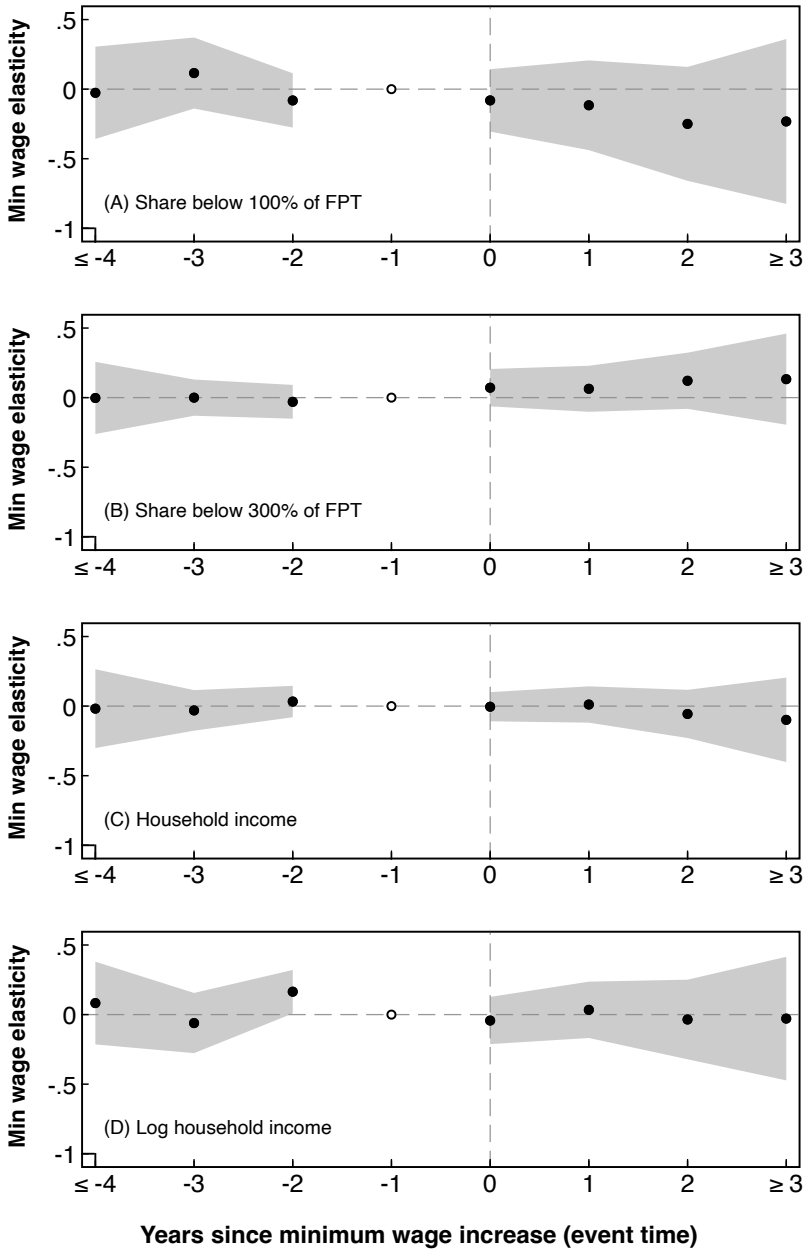
Figure 9 suggests time-varying controls matter. This figure plots event-study estimates constructed as in Figure 8, computed here without state effects. Although noisier than the estimates with state effects plotted in Figure 8, the estimated leads are reasonably precise zeros. This finding suggests that, conditional on controls, ahead of a minimum wage increase, average incomes in states that raise minimum wages are similar to average incomes in states without an increase. In DD applications where dependent variable *levels* are roughly comparable in treatment and control states, the parallel trends assumption is insensitive to functional form.[30]

The potential importance of time-varying controls in an event-study model raises two concerns. First, controls like state GDP and unemployment may be affected by minimum wages and, if so, are *bad controls* in the sense that they're really outcome variables (an ob-

[29] The appendix details this result.

[30] Meyer (1995) appears to be the first to make this point.

Figure 9
Estimated minimum wage effects on household income–
without state effects



Notes: This figure plots binned event-study estimates and simultaneous confidence intervals paralleling those reported in Figure 8, with the modification that state effects are omitted.

servation made in Burkhauser, McNichols and Sabia, 2023). As a rule, control for outcomes compromises rather than supports identification (Angrist and Pischke, 2008, 2015). Second, an empirical strategy that relies on covariates rather than state effects for identification is not really doing DD. Rather, this is old-fashioned regression conditioning.

These concerns lead us to conclude that while parallel trends *might* hold for a variety of transformations, in a setting where state and year effects are needed to identify causal effects, it's hard to see why we'd be so lucky. Within-state income distributions indeed appear to evolve in parallel, but only after conditioning on a rich set of time-varying controls. In our minimum wage analysis, these controls obviate the need for state effects, turning the two-way fixed effects setup into one with time effects only. Research designs with time-varying controls but no state effects inherit the robustness to functional form we expect of conventional regression estimators, but these are no longer models with *two-way* fixed effects. As we see it, TWFE identification strategies are inherently transformation-dependent.

References

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

Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature*, 59(2): 391–425.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica*, 66(2): 249–288.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*. Vol. 3, 1277–1366. Elsevier.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path From Cause to Effect.* Princeton University Press.

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

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review*, 103(6): 2121–2168.

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

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." *The Review of Economic Studies*, 91(6): 3253–3285.

Burkhauser, Richard V, Drew McNichols, and Joseph J Sabia. 2023. "Minimum Wages and Poverty: New Evidence from Dynamic Difference-in-Differences Estimates." NBER Working Paper No. 31182.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-Differences With Multiple Time Periods." *Journal of Econometrics*, 225(2): 200–230.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences With a Continuous Treatment." arXiv preprint arXiv:2107.02637.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference With Clustered Errors." *The Review of Economics and Statistics*, 90(3): 414–427.

Card, David. 1992. "Using Regional Variation in Wages to Measure the Effects of the Federal Minimum Wage." *Industrial and Labor Relations Review*, 46(1): 22–37.

Chamberlain, Gary. 1984. "Panel Data." *Handbook of Econometrics*, 2: 1247–1318.

Chen, Jiafeng, and Jonathan Roth. 2024. "Logs With Zeros? Some Problems and Solutions." *The Quarterly Journal of Economics*, 139(2): 891–936.

de Chaisemartin, Clément, and Xavier d'Haultfoeille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–2996.

DeGroot, Morris Herman. 1986. *Probability and Statistics.* Addison-Wesley Publishing Company.

Dube, Arindrajit. 2019. "Minimum Wages and the Distribution of Family Incomes." *American Economic Journal: Applied Economics*, 11(4): 268–304.

Fernández-Val, Iván, Jonas Meier, Aico van Vuuren, and Francis Vella. 2024. "Distribution Regression Difference-In-Differences." arXiv: 2409.02311.

Figlio, David N., and Umut Özek. 2025. "The Impact of Cellphone Bans in Schools on Student Outcomes: Evidence from Florida." NBER Working Paper No. 34388.

Finkelstein, Amy. 2007. "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." *The Quarterly Journal of Economics*, 122(1): 1–37.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. forthcoming. "Visualization, Identification, and Estimation in the Linear Panel Event-Study Design." *Advances in Economics and Econometrics: Twelfth World Congress.*

Friedberg, Leora. 1998. "Did Unilateral Divorce Raise Divorce Rates? Evidence from Panel Data." *The American Economic Review*, 88(3): 608–627.

Gannon, Mark Andrew, Carlos Alberto de Bragança Pereira, and Adriano Polpo. 2019. "Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels." *The American Statistician*, 73(sup1): 213–222.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2021. "Contamination Bias in Linear Regressions." arXiv: 2106.05024.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, 225(2): 254–277.

Gramlich, Edward M. 1976. "Impact of Minimum Wages on Other Wages, Employment, and Family Incomes." *Brookings Papers on Economic Activity*, 7(2): 409–462.

Leeb, Hannes, and Benedikt M. Pötscher. 2005. "Model Selection and Inference: Facts and Fiction." *Econometric Theory*, 21(1): 21–59.

Lee, Jin Young, and Gary Solon. 2011. "The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates." *The B.E. Journal of Economic Analysis & Policy*, 11(1).

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. 2023. "Cluster-Robust Inference: A Guide to Empirical Practice." *Journal of Econometrics*, 232(2): 272–299.

Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business & Economic Statistics*, 13(2): 151–161.

Miller, Douglas L. 2023. "An Introductory Guide to Event Study Models." *Journal of Economic Perspectives*, 37(2): 203–230.

Montiel Olea, José Luis, and Mikkel Plagborg-Møller. 2019. "Simultaneous Confidence Bands: Theory, Implementation, and an Application to SVARs." *Journal of Applied Econometrics*, 34(1): 1–17.

Nunn, Nathan, and Nancy Qian. 2011. "The Potato's Contribution to Population and Urbanization: Evidence from a Historical Experiment." *The Quarterly Journal of Economics*, 126(2): 593–650.

Pustejovsky, James E., and Elizabeth Tipton. 2018. "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." *Journal of Business & Economic Statistics*, 36(4): 672–683.

Roodman, D., J. MacKinnon, M. Nielsen, and M. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using boottest." *Stata Journal*, 19(1).

Roth, Jonathan. 2022. "Pretest With Caution: Event-Study Estimates After Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–322.

Roth, Jonathan. 2024. "Interpreting Event-Studies from Recent Difference-in-Differences Methods." arXiv: 2401.12309.

Roth, Jonathan, and Pedro HC Sant'Anna. 2023. "When Is Parallel Trends Sensitive to Functional Form?" *Econometrica*, 91(2): 737–747.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." arXiv preprint arXiv:2201.01194.

Sun, Liyang, and Jesse M. Shapiro. 2022. "A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure." *Journal of Economic Perspectives*, 36(4): 193–204.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, 225(2): 175–199.

Wolfers, Justin. 2006. "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results." *The American Economic Review*, 96(5): 1802–1820.

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics: A Modern Approach 6th ed.* South-Western Cengage Learning.

Wright, Philip G. 1928. *The Tariff on Animal and Vegetable Oils.* New York:Macmillan Company.

Appendix

Simulation Details

The simulations generating Figures 5 and 6 and Table 3 start with estimates of the event-study model (7) with a full set of leads and lags. The simulation data set begins in 1956 rather than in 1958. Predicted values are constructed using the state and time effects this regression generates, setting event-study coefficients to zero. Call these predicted values $\hat{Y}_{st}$ and the corresponding residuals e_{st}.

Simulated residuals follow an AR(2) with coefficients taken to be those obtained by estimating

$$e_{st} = \rho_1 e_{st-1} + \rho_2 e_{st-2} + u_{st}, \tag{26}$$

in the divorce data with t starting in 1958. Heteroskedasticity by state is calibrated by regressing squared residuals from the AR(2) model on state effects:

$$u_{st}^2 = \xi_s + v_{st}.$$

The simulation variance for each state, denoted σ_s^2, is set to ξ_s plus the variance of v_{st} computed over all states and years.

Simulations use a balanced panel with 41 periods (matching the 1958 to 1998 period used for estimation) and 48 states (Alaska, Louisiana, and Oklahoma are omitted in our analysis). Each simulation draw is a value of Y_{st} computed as follows:

$$Y_{st} = \hat{Y}_{st} + \psi \times BM_s t + \tilde{e}_{st},$$

where $\tilde{e}_{st}$ is obtained from the estimated AR coefficients in (26) and u_{st} is drawn from a normal distribution with mean zero and variance σ_s^2. ψ denotes the slope of a linear time trend in treated states indicated by M_s while B is a dummy with mean $\theta = 0.5$ that indicates draws from biased studies (i.e., with divergent trends).

Figure 5 displays simulation results from an event-study regression with standard errors that are either clustered (by state) or robust. For this figure, B is fixed at zero. Confidence limits in the figure are computed as +/-1.96 times the average estimated standard errors across simulation draws. The confidence interval for Monte Carlo sampling variance is +/-1.96 times the standard deviation of the estimated coefficients across draws.

To compute power and size in Table 3, a replication starts by estimating (8), binned at +/-15. We then test whether the absolute value of the largest t-statistic for estimated leads is significant, by criteria explained below. The table compares results for five estimation and testing strategies. The first four rows use TWFE event-study regressions. In the first row, standard errors are clustered and tests based on sup-t critical values, computed following the algorithm described in the paper. The second row uses robust standard errors and sup-t critical values. The third row uses clustered standard errors and a critical value of 1.96. The fourth row uses the wild block bootstrap based on the Stata `boottest` command by Roodman et al. (2019). Our implementation of the wild block bootstrap uses the maximum of the pretest t-statistics. We impose the null hypothesis (that all lead coefficients are zero), use `boottype(wild)`, Rademacher (equal) weights, and bootstrap t-statistics with

999 replications.

The fifth line in the table reports results using the BJS estimator, computed using software distributed with the BJS paper. The BJS pretest involves an event-study regression on untreated observations, with some lead coefficients set to zero while others are estimated. The pretest in this case is based on the first 14 lead coefficients and sup-t critical values.[31]

The first column in Table 3 reports the fraction of rejections in draws with $B = 0$; this is the Monte Carlo size of the pretrends pretest. The second column reports rejections in biased studies with $B = 1$, the Monte Carlo power of the pretest. The slope of the linear trend in treated states, ψ, is calibrated so that the baseline model has power of 0.5 using a TWFE regression with clustered standard errors and sup-t critical values; this turns out to be $\psi = 0.03519$. The third column is the difference between columns 2 and 1.

As for the event-study estimates in this chapter, simulation estimates are weighted by state population.

More on Robust Parallel Trends

Section 4 notes that TWFE holds for a set of transformations of a given dependent variable under a parallel distributions restriction. This appendix expands on the connection between parallel distributions cast in terms of distributions of individual outcomes and TWFE models for many state-year averages (thereby extending the Roth and Sant'Anna (2023) two-group analysis).

Parallel distributions is equivalent to a mixture model for individual outcomes. Let $F_{st}(y) = E[1\{Y_{ist}(0) < y\}]$ denote the CDF of $Y_{ist}(0)$ evaluated at y for state s and period t. Roth and Sant'Anna (2023) shows that parallel distributions holds if and only if $F_{st}(y)$ can be written as a mixture of a state-specific but time-invariant distribution $G_s(y)$ and a period-specific but state-invariant distribution $H_t(y)$. In other words,

$$F_{st}(y) = \omega G_s(y) + (1-\omega) H_t(y), \tag{27}$$

where $\omega \in [0,1]$ is a weight between zero and one.

Let $Y_{st}(0)$ denote the sample mean of $Y_{ist}(0)$ in cell (s,t). Then $Y_{st}(0) = \int y dF_{st}(y) + \eta_{st}$, where the integral $\int y dF_{st}(y)$ is the population expectation of $Y_{ist}(0)$ in cell (s,t) and η_{st} is a residual that reflects within-cell sampling variance. Using this to substitute for $F_{st}(y)$ in (27), we arrive at a version of the TWFE model for average outcomes specified by equation (23):

$$\begin{aligned} Y_{st}(0) &= \omega \int y dG_s(y) + (1-\omega) \int y dH_t(y) + \eta_{st} \\ &= \gamma_s + \lambda_t + \eta_{st}. \end{aligned}$$

The same argument applies if $Y_{st}(0)$ is not a sample mean but a population average for individuals living in state s at time t, as in the divorce example from Section 1. Let the

[31] Estimating the first 14 leads is equivalent to setting lead -1 to zero, estimating other leads, and binning at -15, up to normalization. This choice therefore mimics the pretest used for the other estimators and otherwise uses default settings in the BJS software.

population size in state s at time t be N_{st}. The total population at time t across all states is $N_t := \sum_s N_{st}$. Equation (27) implies that a fraction $(1-\omega)$ of individuals in each (s,t) have outcomes drawn from $H_t(y)$, which in this case is the exact finite-population cdf for the $(1-\omega) \times N_t$ individuals across all states at time t whose outcomes are drawn from a state-invariant distribution. The mean among the $(1-\omega) \times N_{st}$ individuals in state s whose $Y_{ist}(0)$ are drawn from $H_t(y)$ behaves like a sample mean, provided the N_{st} individuals in (s,t) are a random draw from the N_t individuals across all states at time t. A similar argument applies to the $\omega \times N_{st}$ individuals in state s at time t whose $Y_{ist}(0)$ are drawn from $G_s(y)$.

Now consider a transformation like $\ln Y_{ist}(0)$, denoted

$$\widetilde{Y}_{ist}(0) \equiv m(Y_{ist}(0)).$$

The state-year sample mean of $\widetilde{Y}_{ist}(0)$ is $\widetilde{Y}_{st}(0) = \int m(y)\, dF_{st}(y) + \tilde{\eta}_{st}$. As in equation (28), we have that

$$\begin{aligned} \widetilde{Y}_{st}(0) &= \omega \int m(y)\, dG_s(y) + (1-\omega) \int m(y)\, dH_t(y) + \tilde{\eta}_{st} \\ &= \tilde{\gamma}_s + \tilde{\lambda}_t + \tilde{\eta}_{st}, \end{aligned}$$

where $\tilde{\gamma}_s$ and $\tilde{\lambda}_t$ denote state and year effects in a model for the transformed outcome, with residual $\tilde{\eta}_{st}$.

Note that the theorem on mixture-model equivalence implies a valid TWFE model for, say, average log income but not for log average income. A related point is that TWFE models for something like the divorce rate are applied directly to the rates themselves. In this case, there's no underlying continuous microdata distribution to restrict because divorce rates are the mean of a divorce dummy. The latter has a distribution function with one parameter (its mean), so that parallel trends for divorce rates immediately imply parallel distributions for individual divorce status.