Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends

By Jonathan Roth[*]

This paper discusses two important limitations of the common practice of testing for pre-existing differences in trends ("pre-trends") when using difference-in-differences and related methods. First, conventional pre-trends tests may have low power. Second, conditioning the analysis on the result of a pre-test can distort estimation and inference, potentially exacerbating the bias of point estimates and undercoverage of confidence intervals. I analyze these issues both in theory and in simulations calibrated to a survey of recent papers in leading economics journals, which suggest that these limitations are important in practice. I conclude with practical recommendations for mitigating these issues. *Keywords: difference-in-differences, pre-trends, pre-testing*

When using difference-in-differences and related methods, researchers often test for pre-treatment differences in trends ("pre-trends") as a way of assessing the plausibility of the parallel trends assumption. These tests are remarkably common: my review of publications in three leading economics between 2014 and 2018 found 70 papers that use an "event-study plot" to visually test for pre-trends (see Section I for details).

This paper highlights two limitations with the practice of pre-testing for pre-trends. First, conventional pre-tests may have low power, meaning that pre-existing trends that produce meaningful bias in the treatment effects estimates may not be detected with substantial probability. Second, conditioning the analysis on the result of a pre-trends test induces distortions to estimation and inference from pre-testing. In other words, the draws of the data that survive a pre-test are a selected sample from the true data-generating process. Because of this selection, the bias caused by a violation of parallel trends can actually be worse conditional on passing the pre-test. Taken together, these results imply that pre-trends tests may be ineffective in avoiding biases from violations of parallel trends and can even exacerbate these biases.

I begin in Section I by illustrating the practical importance of these issues

[*] Brown University. jonathanroth@brown.edu. I am grateful to Isaiah Andrews, Kirill Borusyak, Gary Chamberlain, Peter Ganong, Ed Glaeser, Nathan Hendren, Peter Hull, Ariella Kahn-Lang, Max Kasy, Larry Katz, Sendhil Mullainathan, Andreas Pick, Ashesh Rambachan, Pedro Sant'Anna, Jesse Shapiro, Jann Spiess, Jim Stock, Sophie Sun, and Elie Tamer for helpful comments, as well as to seminar participants at Harvard University and Erasmus University. I also thank the editors, Amy Finkelstein and Rohini Pande, and three anonymous referees for useful suggestions that improved the paper. This material is based upon work supported by the NSF Graduate Research Fellowship under Grant DGE1144152. An earlier version of this paper circulated under the title "Should We Condition on the Test for Pre-trends in Difference-in-Difference Designs?" in April 2018.

in data-generating processes (DGPs) calibrated to a systematic survey of recent papers in three leading economics journals. I consider simulations in which the pre-trends test "passes" if no pre-treatment coefficient is individually statistically significant (and rejects otherwise). I evaluate the power of this pre-test when the true data-generating process has a linear violation of parallel trends. I find that linear violations of parallel trends that would be detected only 50 percent of the time can produce large biases in the treatment effects estimates and lead confidence intervals (CIs) to substantially undercover the true effect. In the most extreme case, the bias from a trend detected only half the time is larger than the estimated treatment effect and a nominal 95% CI contains the true parameter only 24% of the time. I also find that the bias in the selected draws of the data where no significant pre-trend is detected can be quite different from the average (unconditional) bias under the same DGP. The bias conditional on passing the pre-test is larger than the unconditional bias in most specifications – and can be over twice as large – indicating important additional distortions from pre-testing.

In Section II, I provide a theoretical treatment of the distribution of event-study estimates after surviving a pre-test for pre-trends. The analysis clarifies how pre-testing will affect the properties of estimates and CIs for the treatment effect under more general DGPs than in the simulations in Section I (e.g. with non-linear violations of parallel trends). It also clarifies the implications of using pre-tests as a screening device for publication. I begin by deriving the bias and variance of treatment effect estimates *conditional* on surviving the test for pre-existing trends. In general, the bias after surviving a pre-test can be larger or smaller than the unconditional bias. I show, however, that under homoskedasticity the bias will always be larger after surviving the pre-test whenever the difference in trends is monotonically increasing over time. My results also suggest that conditioning on the result of the pre-test can affect the coverage rates of CIs, although the direction of the impact is ambiguous. Finally, a stylized model of the publication process illustrates that screening papers based on pre-trends can either reduce or increase the bias in published work, with tradeoffs between the power of the pre-test to prevent biased estimates from being published and the distortions from pre-testing.

I conclude with practical recommendations for applied researchers in Section III. I describe simple diagnostics that researchers can conduct to evaluate when the limitations of pre-trends testing are likely to be severe, and provide software for their implementation. I also briefly highlight alternative approaches that avoid the pre-test altogether by exploiting economic knowledge about how parallel trends may be violated.

Related Literature. This paper highlights econometric issues with pre-testing for pre-trends. A large literature has considered issues arising from a pre-testing or model-selection step in a variety of other econometric and statistical settings. Early work on pre-testing includes Keynes (1939) and Friedman (1940); for more recent examples, see Giles and Giles (1993), Leeb and Pötscher (2005), Lee et al.

(2016), and Andrews (2018), among many others. Requiring an insignificant pre-trend to publish a paper can also be viewed as a form of publication bias, a topic which has been studied extensively (e.g., Rothstein, Sutton and Borenstein (2005), Christensen and Miguel (2016), Snyder and Zhuo (2018), Andrews and Kasy (2019)).

This paper also contributes to a large body of work on the econometrics of difference-in-differences and related research designs in particular. A topic of substantial recent interest has been the failure of standard two-way fixed effects (TWFE) models to recover a sensible causal estimand in settings with staggered treatment timing and heterogenous treatment effects, even under a suitable parallel trends assumption (Borusyak and Jaravel, 2016; Sun and Abraham, 2020; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2020; Athey and Imbens, 2021). This paper highlights a conceptually distinct issue: even if we were willing to rule out treatment effect heterogeneity (or use a method robust to it), conventional pre-tests may do a poor job detecting violations of the relevant parallel trends assumption. See Remark 1 for further connection to this literature.

Recent papers by Freyaldenhoven, Hansen and Shapiro (2019, FHS), Kahn-Lang and Lang (2020), and Bilinski and Hatfield (2020, BH) have cautioned that pre-trends tests may have low power to detect meaningful violations of parallel trends. I contribute to this work by providing empirical evidence on the power of pre-tests from a systematic review of recent papers. I also provide novel theoretical and empirical evidence on the additional statistical distortions from pre-testing. See Section III for discussion of the alterative approaches proposed by BH and FHS.

I. Survey of Recent Papers

A. *Selecting the sample of papers*

I searched on Google Scholar for occurrences of the phrase "event study" in papers published in the *American Economic Review*, *AEJ: Applied Economics*, and *AEJ: Economic Policy* between 2014 and June 2018.[1] I chose the phrase "event study" since researchers often evaluate pre-trends in an event-study plot.

The search returned 70 total papers that include a figure that the authors describe as an event-study plot. I exclude 43 papers for which data to replicate the main event-study plot were unavailable.[2] I further exclude 9 papers that do not report standard errors,[3] and 3 that do not normalize their estimates relative to a pre-treatment period.[4] Finally, I exclude 3 papers that do not attribute a

[1] I include papers that were forthcoming as of June 2018 if data was available on the AEA website.

[2] This includes one paper where the replication code produced different results from the published paper.

[3] Although standard errors could be estimated from the replication data, I wish to rely on the authors' choice of clustering method.

[4] This rules out financial event-studies examining the time series of returns of an asset.

causal interpretation to their estimates so that I can benchmark the magnitude of biases from differential trends relative to the estimated causal effects. This yields a final sample of 12 papers. For papers that present multiple event-study plots, I focus on the first plot meeting the criteria above, which I view as a reasonable proxy for the main specification.

B. *What pre-tests are researchers using?*

The most common formal criterion for evaluating pre-trends appears to be the individual significance of the pre-treatment coefficients, although this criterion does not appear to be universally applied. All 12 papers in my final sample show an event-plot with pointwise confidence intervals that allows for the evaluation of individual (but not joint) significance of the pre-treatment coefficients. Five of the 12 papers directly discuss individual significance.[5] Only one paper reports a test of joint significance (and it also discusses individual significance), and none of the papers discusses what magnitude of pre-trend can be rejected by the data. Several of the papers, however, appeal only to a visual inspection of the plot without specifying formal criteria. Further, Table 1 makes clear that a statistically significant pre-period coefficient does not necessarily preclude publication: there is at least one statistically significant pre-period coefficient in three of the 12 papers in my final sample, and in two papers the pre-period coefficients are also jointly significant.[6] Although this evidence suggests that not all papers use the individual significance of pre-treatment coefficients as their pre-testing criterion, I nevertheless focus my analysis on this criterion given its prominence in applied work.

C. *Evaluating power and pre-test bias in practice*

I now evaluate the power of conventional pre-tests and the distortions from pre-testing in data-generating processes calibrated to my survey of recent papers.

Data-generating processes. All of the papers in the survey plot a vector of coefficients $\hat{\beta}$, which has subvectors $\hat{\beta}_{pre} \in \mathbb{R}^K$ and $\hat{\beta}_{post} \in \mathbb{R}^M$ corresponding with the periods before and after a treatment occurs. In the simulations below, I consider calibrated data-generating processes (DGPs) in which

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \Sigma\right), \tag{1}$$

[5] Two other papers write that the plot shows “no significant” or “marginally significant” pre-trends, but it is not clear what type of significance they are referring to.

[6] In none of the papers is the slope of the best-fit line through the pre-period coefficients significant at the 5% level. However, no paper mentions this as a criterion of interest, and one case falls just short of significance with a t-statistic of 1.95.

where the mean β satisfies the causal decomposition

$$\beta = \underbrace{\begin{pmatrix} \delta_{pre} \\ \delta_{post} \end{pmatrix}}_{\delta} + \underbrace{\begin{pmatrix} 0 \\ \tau_{post} \end{pmatrix}}_{\tau}, \tag{2}$$

where τ is a vector of causal effects assumed to be zero in the pre-treatment period, and δ is the bias from a difference in trends. All of the papers report standard errors based on the asymptotic normal approximation (1). I impose that this normal approximation holds exactly in finite-sample so that any biases or coverage issues are the results of issues with violations of parallel trends and/or pre-testing rather than the asymptotic distribution providing a poor approximation in finite sample.

Calibrating the model. For each paper in my survey, I calibrate the finite-sample normal model (1) so that the number of pre-treatment and post-treatment periods matches that in the original paper. I set Σ to be the estimated variance-covariance matrix from the specification in the original paper, using whatever clustering method was specified by the authors. I set τ_{post} equal to the estimated coefficients $\hat{\beta}_{post}$, although this choice has no impact on the results.[7] The bias from the difference in trends δ is calibrated based on the power calculations described below.

Power calculations. For each study in my sample, I evaluate the power of pre-trends tests to detect linear violations of parallel trends. In light of the emphasis in published work on the individual statistical significance of the pre-period coefficients, I base my calculations on pre-tests that check this criterion for all pre-treatment coefficients (using 95% CIs). To be precise, I consider a pre-test that "passes" if there is no individually significant pre-treatment coefficient — that is, the test checks whether $\hat{\beta}_{pre} \in B_{NIS}(\Sigma)$, where $B_{NIS}(\Sigma) = \{\beta \in \mathbb{R}^K : |\beta_t| \leq 1.96\sigma_t, \text{ for all t}\}$ and σ_t is the standard error of $\hat{\beta}_{pre,t}$. I calculate the power of such tests against linear violations of parallel trends with a slope of γ, so that the element of δ corresponding with relative time t is $\delta_t = \gamma \cdot t$. I then compute the value of γ for which the probability of rejecting the pre-test is 50 or 80 percent, i.e. $P(\hat{\beta}_{pre} \notin B_{NIS}) = 0.5$ or 0.8. I choose 80 percent since this is often used as a benchmark for the minimum detectable effect in power analyses (Cohen, 1988). I refer to the resulting values, $\gamma_{0.5}$ and $\gamma_{0.8}$, as the slopes against which pre-tests have 50 or 80 percent power.[8]

Target Parameter and Estimator. For simplicity, I focus on estimation of

[7] Specifically, the distribution of $\hat{\beta}_{post}$ conditional on a pre-test of $\hat{\beta}_{pre}$ is equivariant with respect to τ_{post}, and thus has no impact on bias or coverage for τ_{post}.

[8] The power of the pre-test under a slope γ could easily be calculated via simulation. However, under the normality assumption, these probabilities can actually be calculated analytically using results from Cartinhour (1990) and Manjunath and Wilhelm (2012), which I implement using the R package `tmvtnorm`. A similar approach is applied for the bias and coverage calculations described below. I have verified that simulations yield similar results to the analytical approach.

a scalar estimand of the form $\tau_* = l'\tau_{post}$ ($l \in \mathbb{R}^M$). Researchers are often interested in the average treatment effect across all post-treatment periods, and so in the main text I focus on estimation of $\bar{\tau} = \frac{1}{M}(\tau_1 + ... + \tau_M)$. I also consider estimation of the effect for the first period after treatment, τ_1. I focus on the natural plug-in estimate of τ_* under parallel trends, $\hat{\tau} = l'\hat{\beta}_{post}$, and the associated CI, $CI_{\tau_*} = \hat{\tau} \pm 1.96\sigma_{\hat{\tau}}$, where $\sigma^2_{\hat{\tau}} = l'\Sigma l$.

Bias and size calculations. I evaluate the performance of these estimators and CIs under data-generating processes with linear violations of parallel trends with slopes $\gamma_{0.5}$ or $\gamma_{0.8}$. Specifically, I calculate the unconditional bias $\mathbb{E}\left[\hat{\tau} - \tau_*\right]$, and the bias conditional on passing the pre-test $\mathbb{E}\left[\hat{\tau} - \tau_* \mid \hat{\beta}_{pre} \in B_{NIS}(\Sigma)\right]$. Analogously, I calculate the size (i.e. null rejection probability) of $CI_{\tau*}$ both unconditionally and conditionally, $\mathbb{P}\left(\tau_* \notin CI_{\tau_*}\right)$ and $\mathbb{P}\left(\tau_* \notin CI_{\tau_*} \mid \beta_{pre} \in B_{NIS}(\Sigma)\right)$.

Results. My results indicate that pre-trends tests often have low power against violations of parallel trends that would produce meaningful bias in the treatment effects estimates. The green circles in Figure 1 show the bias for the average effect ($\bar{\tau}$) from a linear difference in trends which would be detected 80% of the time ($\gamma_{0.8}$). These biases are benchmarked relative to the magnitude of the treatment effect estimate in the original paper (plotted in blue). The bias from such a trend is often of a magnitude comparable to, and in some cases larger than, the estimated treatment effect! This implies that the slope of the differential trend needs to be quite large in order to achieve 80% power (and power will be even lower for smaller violations). As a result of these biases, traditional CIs exhibit substantial undercoverage under these violations of parallel trends, as shown in Table 2. Although the true parameter should nominally fall outside a 95% confidence interval no more than 5% of the time, in several specifications this occurs over 50% of the time. Results for the first period after treatment (τ_1) and using a 50% power threshold ($\gamma_{0.5}$) show qualitatively similar patterns, although somewhat less extreme, and are presented in the Online Appendix.

I also find substantial distortions from pre-testing. The red triangles in Figure 1 show the bias for $\bar{\tau}$ conditional on surviving the pre-test. As can be seen, the conditional bias can be different from, and in most cases worse than, the unconditional bias. Table 3 summarizes the additional bias from pre-testing as a fraction of the unconditional bias: for the trend against which pre-tests have 50 percent power, the pre-test bias can be as much as 103 percent of the unconditional bias for τ_1, and as much as 48 percent for $\bar{\tau}$.[9] Moreover, the pre-test bias and the bias from trend go in the same direction in all but two of the studies in the sample when the estimand is $\bar{\tau}$, and all but three of the studies when it is τ_1. Thus, in most cases the bias from pre-testing exacerbates the bias from the underlying trend. Similarly, Table 2 shows that the null rejection rates for 95% CIs conditional on passing the pre-test can differ substantially from the unconditional null rejection

[9] We expect the bias from pre-testing to be a larger fraction of the unconditional bias for periods closer to treatment, since the unconditional bias from the differential trend grows linearly in the number of periods after treatment, whereas the pre-test bias need not grow over time.

rates, and are worse in many cases.

Intuition. Some intuition for why the power of pre-trends tests may be low is as follows. Consider the case where we have one pre-treatment and one post-treatment coefficient ($M = K = 1$), the two coefficients have the same variance ($\Sigma_{11} = \Sigma_{22}$), and the true treatment effect is zero. Under a linear trend, $\delta_{pre} = -\delta_{post}$, and so by symmetry the probability that the CI for $\hat{\beta}_{pre}$ contains 0 is the same as the probability that the CI for $\hat{\beta}_{post}$ contains 0. Thus, if the pre-test rejects zero half the time, then the CI for $\hat{\beta}_{post}$ will reject zero half the time as well – that is, 10 times more often than a 95% CI is supposed to reject the true effect! This problem becomes even more severe when we have multiple post-treatment periods, since the bias from a linear trend grows over time. Likewise, it becomes less severe as we add more pre-treatment periods, which raises the probability of detecting a significant pre-trend.

It is worth highlighting, however, that these comparative statics with respect to the number of periods are somewhat particular to the assumed linear form for the pre-trend. Adding additional pre-treatment periods may not help the power of the pre-test if we expect treatment status to be determined only by events close to the time of treatment — in a study using COVID-19 cases as the outcome, for example, it would not be very informative to check for parallel pre-trends for years prior to 2019.

This two-period example can also provide some intuition for why pre-testing can exacerbate bias. If there is an upward-sloping trend so that $\beta_{pre} < 0$, then draws of the data where we the pre-test passes will tend to have $\hat{\beta}_{pre} > \beta_{pre}$. But if $\hat{\beta}_{post}$ and $\hat{\beta}_{pre}$ are positively correlated, then $\hat{\beta}_{post}$ will also tend to be above β_{post}, exacerbating the bias from the upward-sloping pre-existing trend.

D. Caveats and Discussion

An important caveat to these results is that by construction my sample only includes papers that made it through the publication process at leading economics journals and reported an event-study plot in the published manuscript. To the extent that papers with insignificant pre-trends are more likely to be published, or that analyses with significant pre-trends are not reported in the final manscript, the sample may be biased towards papers where the power of pre-tests is low.

A second important caveat is that these results only directly provide evidence about the power of pre-trends tests when there is a *linear* violation of parallel trends.[10] Assessing the power of pre-tests against linear violations of parallel trends is a natural benchmark given that researchers worried about differential trends often include parametric linear controls (e.g., Wolfers (2006); Dobkin et al. (2018); Goodman-Bacon (2018)), which suggests that authors perceive linear vi-

[10] In the Online Appendix, I conduct a similar power analysis in which there are stochastic shocks to the treated and control groups, and again find poor performance of standard pre-testing methods in controlling size distortions from the differential trends.

olations of parallel trends to be relevant in many cases. Nevertheless, one may be interested in the power of pre-tests against non-linear violations of parallel trends as well.[11] Heuristically, these issues will be even more severe if the difference in trends is becoming steeper over time. For instance, if the difference in trends is growing exponentially over time, then it will be small in the pre-treatment period (so rejecting the pre-test is unlikely), but the biases in the post-treatment period will be quite large. Conversely, if the difference in trends were becoming shallower over time (e.g., if it were logarithmic), then we would be more likely to detect the steep pre-trend even though it produces a relatively small post-treatment bias.

II. Theoretical Analysis

A. *Model*

I analyze the normal model introduced in equations (1) and (2) above. The main goal of our analysis will be to analyze the distribution of the post-treatment coefficients $\hat{\beta}_{post}$ conditional on passing a pre-test based on the pre-treatment estimates $\hat{\beta}_{pre}$, i.e. conditional on the event $\hat{\beta}_{pre} \in B(\Sigma)$ for some (measurable) set $B(\Sigma)$ potentially depending on the covariance matrix (e.g. individual or joint tests of significance). For ease of notation, I consider the case where there is one post-treatment period ($M = 1$) unless noted otherwise; all of the results for $M = 1$ will then apply to each individual post-period coefficient (or linear combinations thereof) in the case when $M > 1$.

Remark 1. The finite-sample normal model (1) can be be thought of as an asymptotic approximation to a variety of estimators which yield asymptotically normal coefficients, $\sqrt{N}(\hat{\beta}_n - \beta_n) \to_d \mathcal{N}(0, \Sigma)$. Estimators yielding event-study coefficients of this form (under suitable regularity conditions) include dynamic TWFE estimators, the GMM estimator of Freyaldenhoven, Hansen and Shapiro (2019), and methods for difference-in-differences conditional on covariates (Abadie, 2005; Heckman, Ichimura and Todd, 1997; Sant'Anna and Zhao, 2020). The recent proposals by Callaway and Sant'Anna (2020) and Sun and Abraham (2020) for constructing event-study estimates that have a sensible interpretation under staggered treatment timing and treatment effect heterogeneity also yield asymptotically normal coefficients. The results here are thus directly applicable to these estimators, which highlights that the issues surrounding pre-testing are distinct from those related to the interpretation of TWFE models under heterogeneity. ■

The Online Appendix shows that the results derived in the finite sample normal model hold uniformly over a wide range of data-generating processes under which the probability of passing the pre-test does not vanish asymptotically.[12] The

[11] Indeed, if linear violations of parallel trends were the only concern, one could include parametric controls and avoid the pre-test altogether.

[12] The condition that the probability of passing the pre-test does not vanish asymptotically requires that the pre-treatment trend δ_{pre} be shrinking with the sample size. This local-to-0 approximation

asymptotics also allow for the pre-test to depend on a consistently estimated covariance matrix, $\hat{\Sigma} \rightarrow_p \Sigma$.

B. Bias After Pre-testing

I begin by analyzing the bias of $\hat{\beta}_{post}$ for τ_{post} conditional on passing the pre-test. The following result provides a formula for the conditional bias.

Proposition 1. *For any conditioning set* $B(\Sigma)$,

$$\mathbb{E}\left[\hat{\beta}_{post} \mid \hat{\beta}_{pre} \in B(\Sigma)\right] = \tau_{post} + \delta_{post} + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbb{E}\left[\hat{\beta}_{pre} \mid \hat{\beta}_{pre} \in B(\Sigma)\right] - \beta_{pre}\right),$$

where $\mathbb{V}ar\left[\begin{pmatrix} \hat{\beta}_{post} \\ \hat{\beta}_{pre} \end{pmatrix}\right] = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.

Proof. Let $\tilde{\beta}_{post} = \hat{\beta}_{post} - \Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}$. By construction, $cov(\hat{\beta}_{pre}, \tilde{\beta}_{post}) = 0$, which by joint normality implies that $\hat{\beta}_{pre} \perp\!\!\!\perp \tilde{\beta}_{post}$. Hence $\mathbb{E}\left[\tilde{\beta}_{post} \mid \hat{\beta}_{pre} \in B(\Sigma)\right] = \mathbb{E}\left[\tilde{\beta}_{post}\right] = \beta_{post} - \Sigma_{12}\Sigma_{22}^{-1}\beta_{pre}$. The result then follows from taking conditional expectations on both sides of the equation $\hat{\beta}_{post} = \tilde{\beta}_{post} + \Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}$. □

The formula in Proposition 1 illustrates that the expectation of $\hat{\beta}_{post}$ conditional on passing the pre-test is the sum of i) the treatment effect of interest τ_{post}, ii) the unconditional bias δ_{post}, and iii) an additional "pre-test bias" term, which depends on the distortion to the mean of the pre-treatment coefficients from pre-testing and the covariance between the pre-treatment and post-treatment coefficients.

An immediate implication of Proposition 1 is that when parallel trends holds ($\delta = 0$), $\hat{\beta}_{post}$ remains unbiased for τ_{post} after pre-testing so long as the pre-test is such that $\mathbb{E}\left[\hat{\beta}_{pre} | \hat{\beta}_{pre} \in B\right] = 0$. It is straightforward to verify that this condition holds whenever the pre-test is symmetric in the sense that we reject the hypothesis of parallel pre-trends for $\hat{\beta}_{pre}$ if and only if we reject the hypothesis for $-\hat{\beta}_{pre}$, a property which holds for two-sided tests of significance.

Sufficient conditions for bias exacerbation

In the simulations in Section I, we saw that for most specifications, the bias of $\hat{\beta}_{post}$ for τ_{post} was worse conditional on passing the pre-test when there were linear violations of parallel trends. I now show that under homoskedasticity, the conditional bias will be worse than the unconditional bias whenever there is a monotone (possibly non-linear) difference in trends.

captures the fact that in finite samples the pre-trend may be of a similar order of magnitude as the sampling uncertainty in the data (as with $\gamma_{0.5}$ and $\gamma_{0.8}$). In a model with fixed δ_{pre}, the probability of rejecting the pre-test would be either 0 or 1 asymptotically, which does not capture the fact that in practice we are often uncertain whether the pre-trend is zero or not.

Assumption 1. *Σ has a common term σ^2 on the diagonal and a common term $\rho > 0$ off of the diagonal, with $\sigma^2 > \rho$.*[13]

When there is only one pre-treatment and one post-treatment coefficient, Assumption 1 merely imposes that the pre-treatment and post-treatment coefficients are positively correlated. In the more general case with multiple pre-treatment periods, Assumption 1 is implied by a suitable homoskedasticity assumption in the canonical two-way fixed effects difference-in-differences model with non-staggered timing. To see this, suppose that the data is generated from the model

$$y_{it} = \alpha_i + \phi_t + \sum_{s \neq 0} \underbrace{\beta_s}_{\tau_s + \delta_s} \times 1[t = s] \times D_i + \epsilon_{it}, \tag{3}$$

where D_i is an indicator for whether i is first treated at $t = 1$ or never treated. If the researcher estimates β_s via OLS, then the estimated coefficients will be given by

$$\hat{\beta}_s = \beta_s + \Delta\bar{\epsilon}_s - \Delta\bar{\epsilon}_0,$$

where $\Delta\bar{\epsilon}_t$ is the difference in the average residuals for the treatment and control groups in period t. It follows immediately that if the ϵ_{it} are homoskedastic, $\mathbb{V}\text{ar}\left[\hat{\beta}_k\right] = 2\mathbb{V}\text{ar}\left[\Delta\bar{\epsilon}_0\right] =: \sigma^2$ and $\text{Cov}(\hat{\beta}_k, \hat{\beta}_j) = \sigma^2/2 =: \rho$, so Assumption 1 holds.

We now show that under Assumption 1, the bias after testing for significant pre-treatment coefficients is worse than the unconditional bias under arbitrary monotone violations of parallel trends. This result complements the findings in Section I, since it allows for arbitrary non-linear violations of parallel trends but imposes stronger assumptions on the covariance matrix.

Proposition 2 (Sign of bias under monotone trend)**.** *Suppose that there is an upward pre-trend in the sense that $\delta_{pre} < 0$ (elementwise) and $\delta_{post} > 0$.*[14] *If Assumption 1 holds, then*

$$\mathbb{E}\left[\hat{\beta}_{post} \mid \hat{\beta}_{pre} \in B_{NIS}(\Sigma)\right] > \beta_{post} > \tau_{post}.$$

The analogous result holds replacing ">" with "<" and vice versa.

Proof. From Proposition 1, it suffices to show that $\Sigma_{12}\Sigma_{22}^{-1}(\mathbb{E}\left[\hat{\beta}_{pre} \mid \hat{\beta}_{pre} \in B(\Sigma)\right] - \beta_{pre}) > 0$. When $K = 1$, $\Sigma_{12}\Sigma_{22}^{-1} = \rho/\sigma^2$, which is positive by assumption. The formula for the mean of a univariate truncated normal implies that $\mathbb{E}\left[\hat{\beta}_{pre} \mid \hat{\beta}_{pre} \in B(\Sigma)\right] - \beta_{pre} \propto \phi(-1.96 - \beta_{pre}/\sigma_{-1}) - \phi(1.96 - \beta_{pre}/\sigma_{-1})$, which is positive since $\beta_{pre} < 0$

[13] If $K = 1$, it suffices to impose that $Cov(\hat{\beta}_{pre}, \hat{\beta}_{post}) > 0$.

[14] Technically, the restriction that $\delta_{pre} < 0$ and $\delta_{post} > 0$ is somewhat weaker than monotonicity. It allows, for instance, for $\delta_{-3} > \delta_{-2}$, so long as both are less than 0.

and $\phi(x)$ is decreasing in $|x|$. The argument for when $K > 1$ is similar but involves some rather tedious algebra since the mean of a truncated multivariate normal depends on the full covariance matrix. A full proof for the $K > 1$ case, which adapts arguments from Papadopoulos (2013) to this setting, is given in the Online Appendix. □

Remark 2. Monotonicity of δ is often implied in the discussion of violations of parallel trends in applied work. For instance, Lovenheim and Willen (2019) argue that violations of parallel trends cannot explain their results because "pre-[treatment] trends are either zero or in the wrong direction (i.e., opposite to the direction of the treatment effect)." Nonetheless, there are economic settings in which we do not expect monotonicity to hold, with the "Ashenfelter's dip" expected in job-training programs as a notable example (Ashenfelter, 1978). ■

Remark 3. The homoskedasticity assumption is of course strong and unlikely to hold exactly in practical applications. It may, however, be a reasonable approximation in many cases, as evidenced by the fact that the pre-test bias goes in the direction predicted by Proposition 2 in most of the simulations in Section I. Moreover, the fact that bias is exacerbated under homoskedasticity and arbitrary monotone violations of parallel trends suggests that these issues extend beyond the case of linear differences in trends considered in Section I. ■

C. *Variance after pre-testing*

We now consider the variance of $\hat{\beta}_{post}$ after pre-testing.

Proposition 3.

$$\begin{aligned}&\mathbb{V}ar\left[\hat{\beta}_{post}|\hat{\beta}_{pre} \in B(\Sigma)\right] = \\ &\mathbb{V}ar\left[\hat{\beta}_{post}\right] + (\Sigma_{12}\Sigma_{22}^{-1})\left(\mathbb{V}ar\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B(\Sigma)\right] - \mathbb{V}ar\left[\hat{\beta}_{pre}\right]\right)(\Sigma_{12}\Sigma_{22}^{-1})'.\end{aligned}$$

Proof. The proof is analogous to the derivation of the mean in Proposition 1. The result follows from taking conditional variances on both sides of the equation $\hat{\beta}_{post} = \tilde{\beta}_{post} + \Sigma_{12}\Sigma_{22}^{-1}\hat{\beta}_{pre}$ and using the fact that $\hat{\beta}_{pre} \perp\!\!\!\perp \tilde{\beta}_{post}$. □

Proposition 3 implies that the variance of $\hat{\beta}_{post}$ will typically be smaller after conditioning on the result of the pre-test. Indeed, this is the case when the acceptance region for the pre-test is convex, a property which holds for most tests of individual or joint significance.

Proposition 4 (Pre-testing reduces variance)**.** *Suppose that $B(\Sigma)$ is a convex set. Then* $\mathbb{V}ar\left[\hat{\beta}_{post} \,|\, \hat{\beta}_{pre} \in B(\Sigma)\right] \leq \mathbb{V}ar\left[\hat{\beta}_{post}\right]$.

Proof. From Proposition 3, it suffices to show that

$$\left(\mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre} \,|\, \hat{\beta}_{pre} \in B(\Sigma)\right] - \mathbb{V}\mathrm{ar}\left[\hat{\beta}_{pre}\right]\right) < 0.$$

Papadopoulos (2013) showed that that this was the case for the scalar case $K = 1$, exploiting the log-concavity of the normal distribution. This argument extends naturally to the multivariate case; see the Online Appendix for details. □

Since $\hat{\beta}_{post}$ is unbiased conditional on passing the pre-test under parallel trends (provided B is symmetric about 0), Proposition 4 suggests that typical confidence intervals will tend to over-cover conditional on passing the pre-test under parallel trends.[15] Intuitively, this is because standard errors are based on estimates of the unconditional variance, which is too large. When parallel trends is violated, however, $\hat{\beta}_{post}$ will be biased, and thus conventional CIs will tend to under-cover if the bias is sufficiently large, as shown in the simulations in Section I.

D. *Implications for Publication Rules*

What do the results above imply about the use of pre-trends tests as a screening device for publication? Our results so far imply that if all studies had the same "true" difference in trends, then only publishing studies without significant pre-trends would likely exacerbate the bias in published work owing to pre-test bias. However, in practice not all attempted studies will have the same true difference in trends. Requiring an insignificant pre-trend to publish may help to select studies in which the true difference in trends is small. Requiring an insignificant pre-trend to publish a paper thus has an ambiguous effect on average bias in published work, depending on which of these effects dominates.

The following simple model clarifies these tradeoffs. Suppose parallel trends holds ($\delta = 0$) in fraction $1 - \theta$ of studies and in the remaining θ fraction of studies $\delta = \bar{\delta} \neq 0$. If all studies were published, regardless of pre-trends, then the expected bias in published work would be

$$Bias^{Notest} = P(\delta = \bar{\delta})\bar{\delta}_{post} = \theta\bar{\delta}_{post}.$$

On the other hand, if we only published the studies without a significant pre-trend ($\hat{\beta}_{pre} \in B(\Sigma)$), the expected bias in published work would be

$$Bias^{Pre-test} = P(\delta = \bar{\delta} \mid \hat{\beta}_{pre} \in B(\Sigma))\mathbb{E}\left[\hat{\beta}_{post} - \tau_{post} \mid \hat{\beta}_{pre} \in B(\Sigma)\right].$$

Comparing the biases under the two publication regimes, we have

[15] This is not formally implied by the proposition, since the conditional distribution of $\hat{\beta}_{post}$ may be non-normal. It is, however, always the case in simulations based on the survey of papers in Section I; see Table 2.

$$\frac{Bias^{Test}}{Bias^{Notest}} = \underbrace{\frac{P(\delta = \bar{\delta} \mid \hat{\beta}_{pre} \in B(\Sigma))}{P(\delta = \bar{\delta})}}_{\text{Relative fraction of biased studies}} \cdot \underbrace{\frac{\mathbb{E}\left[\hat{\beta}_{post} - \tau_{post} \mid \delta = \bar{\delta}, \hat{\beta}_{pre} \in B(\Sigma)\right]}{\bar{\delta}_{post}}}_{\text{Ratio of bias when publish biased design}} . \tag{4}$$

The first term represents the relative fraction of published studies with a biased design ($\delta = \bar{\delta}$) across the two regimes. This will tend to be less than 1, since the pre-test will reject less frequently conditional on $\delta = 0$. By contrast, the second term is the ratio of the conditional and unconditional biases when $\delta = \bar{\delta}$, which will often be greater than 1 owing to pre-test bias (see Proposition 2).

The effect of requiring an insignificant on the bias in published work is thus ambiguous, and depends on the relative magnitude of these two factors. When is the pre-testing regime least effective (and potentially harmful)? It is straightforward to show that the first term in (4) converges to 1 if either i) $\theta \to 1$, so that nearly all studies have the same true trend, or ii) the Bayes Factor, $P(\hat{\beta}_{pre} \in B(\Sigma) \mid \delta = \bar{\delta}) / P(\hat{\beta}_{pre} \in B(\Sigma) \mid \delta = 0)$ converges to 1, so that the pre-test has no power to distinguish between a biased and unbiased design.

The pre-testing regime is thus ineffective at reducing bias when either the ex ante credibility of studies (as proxied by $1 - \theta$) is low, or the pre-test is underpowered (meaning the Bayes Factor is low). A similar analysis applies to the null rejection probability in published studies.

III. Practical Recommendations

In light of the results in Section I, researchers relying on pre-trends tests should assess whether their tests are likely to be well-powered against relevant violations of parallel trends that would produce meaningful biases in the treatment effect estimates. To facilitate such assessment, I provide the R package `pretrends` and an accompanying `Shiny application` to conduct power analyses analogous to those in Section I. The package can also assess the power of conventional pre-tests against hypothesized non-linear trends, allowing the user to do power analyses for the types of violations of parallel trends deemed to be most relevant in their context. Relatedly, Freyaldenhoven et al. (2021) provide tools for visualizing possible violations of parallel trends, and Bilinski and Hatfield (2018) propose alternative approaches to pre-testing that examine what magnitude of the pre-trend can be rejected. Paying careful attention to the power of pre-tests against economically relevant alternatives (and their magnitudes) would be a substantial improvement on the current practice of focusing on statistical significance. Nevertheless, doing so does not avoid the issues of statistical distortions from pre-testing, nor does it formally guarantee statistically valid inference on the treatment effect.

Researchers should therefore also consider alternative approaches that attempt to avoid the pre-testing problem altogether. Freyaldenhoven, Hansen and Shapiro

(2019) propose an approach that exploits a covariate assumed to be affected by the relevant confounding factors but not by the treatment itself. This covariate is then used to adjust for the counterfactual difference in trends, thus avoiding the need for non-zero pre-trends. Rambachan and Roth (2021) develop confidence sets for the treatment effect that are valid under the assumption that the counterfactual difference in trends in the post-treatment period cannot differ "too much" from the difference in trends in the pre-treatment period. Their confidence sets directly account for the uncertainty over the magnitude of the pre-treatment trend, and thus avoid the need to test whether the pre-trends are zero. Their approach also enables sensitivity analyses that show how much the post-treatment differences in trends would need to differ from the pre-trends for specific conclusions (e.g. a significant effect) to break down.

Regardless of the exact approach taken, I urge researchers to use context-specific economic knowledge to inform the discussion and analysis of possible violations of parallel trends. Bringing economic knowledge to bear on how parallel trends might plausibly be violated in a given context will yield stronger, more credible inferences than relying on the statistical significance of pre-trends tests alone.

REFERENCES

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

Andrews, Isaiah. 2018. "Valid Two-Step Identification-Robust Confidence Sets for GMM." *The Review of Economics and Statistics*, 100(2): 337–348.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review*, 109(8): 2766–2794.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics*, 60(1): 47–57.

Athey, Susan, and Guido W. Imbens. 2021. "Design-based analysis in Difference-In-Differences settings with staggered adoption." *Journal of Econometrics.*

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

Bailey, Martha J., and Andrew Goodman-Bacon. 2019. "Replication data for: The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Bilinski, Alyssa, and Laura A. Hatfield. 2018. "Seeking Evidence of Absence: Reconsidering Tests of Model Assumptions." *arXiv:1805.03273 [stat]*.

Bilinski, Alyssa, and Laura A. Hatfield. 2020. "Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions." *arXiv:1805.03273 [stat]*. arXiv: 1805.03273.

Borusyak, Kirill, and Xavier Jaravel. 2016. "Revisiting Event Study Designs." Social Science Research Network SSRN Scholarly Paper ID 2826228, Rochester, NY.

Bosch, Mariano, and Raymundo M. Campos-Vazquez. 2014. "The Trade-Offs of Welfare Policies in Labor Markets with Informal Jobs: The Case of the "Seguro Popular" Program in Mexico." *American Economic Journal: Economic Policy*, 6(4): 71–99.

Bosch, Mariano, and Raymundo M. Campos-Vazquez. 2019. "Replication data for: The Trade-Offs of Welfare Policies in Labor Markets with Informal Jobs: The Case of the." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2020. "Difference-in-Differences with multiple time periods." *Journal of Econometrics*.

Cartinhour, Jack. 1990. "One-Dimensional Marginal Density Functions of a Truncated Multivariate Normal Density Function." *Communications in Statistics-Theory and Methods*, 19: 197–203.

Christensen, Garret S., and Edward Miguel. 2016. "Transparency, Reproducibility, and the Credibility of Economics Research." National Bureau of Economic Research Working Paper 22989.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences.* Academic Press. Google-Books-ID: YleCAAAAIAAJ.

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

Deryugina, Tatyana. 2017. "The Fiscal Cost of Hurricanes: Disaster Aid versus Social Insurance." *American Economic Journal: Economic Policy*, 9(3): 168–198.

Deryugina, Tatyana. 2019. "Replication data for: The Fiscal Cost of Hurricanes: Disaster Aid versus Social Insurance." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Deschênes, Olivier, Michael Greenstone, and Joseph S. Shapiro. 2017. "Defensive Investments and the Demand for Air Quality: Evidence from the NOx Budget Program." *American Economic Review*, 107(10): 2958–2989.

Deschênes, Olivier, Michael Greenstone, and Joseph S. Shapiro. 2019. "Replication data for: Defensive Investments and the Demand for Air Quality: Evidence from the NOx Budget Program." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo. 2018. "The Economic Consequences of Hospital Admissions." *American Economic Review*, 108(2): 308–352.

Fitzpatrick, Maria D, and Michael F Lovenheim. 2014. "Early Retirement Incentives and Student Achievement." *American Economic Journal: Economic Policy*, 6(3): 120–154.

Fitzpatrick, Maria D., and Michael F. Lovenheim. 2019. "Replication data for: Early Retirement Incentives and Student Achievement." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2019. "Pre-event Trends in the Panel Event-Study Design." *American Economic Review*, 109(9): 3307–3338.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. 2021. "Visualization, Identification, and Estimation in the Linear Panel Event-Study Design." National Bureau of Economic Research Working Paper 29170. Series: Working Paper Series.

Friedman, Milton. 1940. "Review of Jan Tinbergen. Statistical testing of business cycle theories, II: Business cycles in the United States of America." *American Economic Review*, 30.

Gallagher, Justin. 2014. "Learning about an Infrequent Event: Evidence from Flood Insurance Take-Up in the United States." *American Economic Journal: Applied Economics*, 6(3): 206–233.

Gallagher, Justin. 2019. "Replication data for: Learning about an Infrequent Event: Evidence from Flood Insurance Take-Up in the United States." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Giles, Judith A., and David E. A. Giles. 1993. "Pre-Test Estimation and Testing in Econometrics: Recent Developments." *Journal of Economic Surveys*, 7(2): 145–197.

Goodman-Bacon, Andrew. 2018. "Public Insurance and Mortality: Evidence from Medicaid Implementation." *Journal of Public Economics*, 126(1): 216–262.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics.*

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies*, 64(4): 605–654. Publisher: Oxford Academic.

He, Guojun, and Shaoda Wang. 2017. "Do College Graduates Serving as Village Officials Help Rural China?" *American Economic Journal: Applied Economics*, 9(4): 186–215.

He, Guojun, and Shaoda Wang. 2019. "Replication data for: Do College Graduates Serving as Village Officials Help Rural China?" Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Kahn-Lang, Ariella, and Kevin Lang. 2020. "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications." *Journal of Business & Economic Statistics*, 38(3): 613–620. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07350015.2018.1546591.

Keynes, J. M. 1939. "Professor Tinbergen's Method." *The Economic Journal*, 49(195): 558–577.

Kuziemko, Ilyana, Katherine Meckel, and Maya Rossin-Slater. 2018. "Does Managed Care Widen Infant Health Disparities? Evidence from Texas Medicaid." *American Economic Journal: Economic Policy*, 10(3): 255–283.

Kuziemko, Ilyana, Katherine Meckel, and Maya Rossin-Slater. 2019. "Replication data for: Does Managed Care Widen Infant Health Disparities? Evidence from Texas Medicaid." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School Finance Reform and the Distribution of Student Achievement." *American Economic Journal: Applied Economics*, 10(2): 1–26.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2019. "Replication data for: School Finance Reform and the Distribution of Student Achievement." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

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

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. 2016. "Exact Post-Selection Inference, with Application to the Lasso." *The Annals of Statistics*, 44(3): 907–927.

Lovenheim, Michael F., and Alexander Willen. 2019. "The Long-Run Effects of Teacher Collective Bargaining." *American Economic Journal: Economic Policy*, 11(3): 292–324.

Manjunath, B.G, and Stefan Wilhelm. 2012. "Moments Calculation For the Doubly Truncated Multivariate Normal Density." *arXiv:1206.5387 [stat].*

Markevich, Andrei, and Ekaterina Zhuravskaya. 2018. "The Economic Effects of the Abolition of Serfdom: Evidence from the Russian Empire." *American Economic Review*, 108(4-5): 1074–1117.

Markevich, Andrei, and Ekaterina Zhuravskaya. 2019. "Replication data for: The Economic Effects of the Abolition of Serfdom: Evidence from the Russian Empire." Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Papadopoulos, Alecos. 2013. "Is the mean of the truncated normal distribution monotone in μ?" *Mathematics Stack Exchange*, URL:https://math.stackexchange.com/q/455809 (version: 2013-08-01).

Rambachan, Ashesh, and Jonathan Roth. 2021. "An Honest Approach to Parallel Trends." Working Paper.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein. 2005. "Publication Bias in Meta-Analysis." In *Publication Bias in Meta-Analysis.* , ed. Hannah R. Rothstein Co-Chair, Alexander J. Sutton Co-Author and Michael Borenstein Director Associateessor Lecturer PI, 1–7. John Wiley & Sons, Ltd.

Sant'Anna, Pedro H. C., and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of Econometrics*, 219(1): 101–122.

Snyder, Christopher, and Ran Zhuo. 2018. "Sniff Tests in Economics: Aggregate Distribution of Their Probability Values and Implications for Publication Bias." National Bureau of Economic Research Working Paper 25058.

Sun, Liyang, and Sarah Abraham. 2020. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics.*

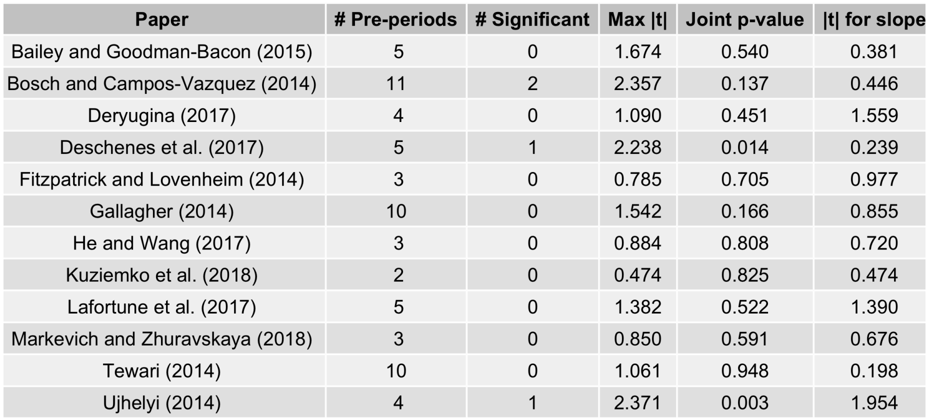
Tewari, Ishani. 2014. "The Distributive Impacts of Financial Development: Evidence from Mortgage Markets during US Bank Branch Deregulation." *American Economic Journal: Applied Economics*, 6(4): 175–196.

Tewari, Ishani. 2019. “Replication data for: The Distributive Impacts of Financial Development: Evidence from Mortgage Markets during US Bank Branch Deregulation.” Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Ujhelyi, Gergely. 2014. “Civil Service Rules and Policy Choices: Evidence from US State Governments.” *American Economic Journal: Economic Policy*, 6(2): 338–380.

Ujhelyi, Gergely. 2019. “Replication data for: Civil Service Rules and Policy Choices: Evidence from US State Governments.” Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.

Wolfers, Justin. 2006. “Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results.” *American Economic Review*, 96: 1802–1820.



Paper	# Pre-periods	# Significant	Max |t|	Joint p-value	|t| for slope
Bailey and Goodman-Bacon (2015)	5	0	1.674	0.540	0.381
Bosch and Campos-Vazquez (2014)	11	2	2.357	0.137	0.446
Deryugina (2017)	4	0	1.090	0.451	1.559
Deschenes et al. (2017)	5	1	2.238	0.014	0.239
Fitzpatrick and Lovenheim (2014)	3	0	0.785	0.705	0.977
Gallagher (2014)	10	0	1.542	0.166	0.855
He and Wang (2017)	3	0	0.884	0.808	0.720
Kuziemko et al. (2018)	2	0	0.474	0.825	0.474
Lafortune et al. (2017)	5	0	1.382	0.522	1.390
Markevich and Zhuravskaya (2018)	3	0	0.850	0.591	0.676
Tewari (2014)	10	0	1.061	0.948	0.198
Ujhelyi (2014)	4	1	2.371	0.003	1.954

Table 1—Summary of Pre-period Event-Study Coefficients

Note: This table provides information about the pre-period event-study coefficients in the papers reviewed. The table shows the number of pre-periods in the event-study, the number of the pre-period coefficients that are significant at the 95% level, the maximum t-stat among those coefficients, the p-value for a chi-squared test of joint significance, and the t-stat for the slope of the linear trend through the pre-period coefficients. See Section I for more detail on the sample of papers reviewed.

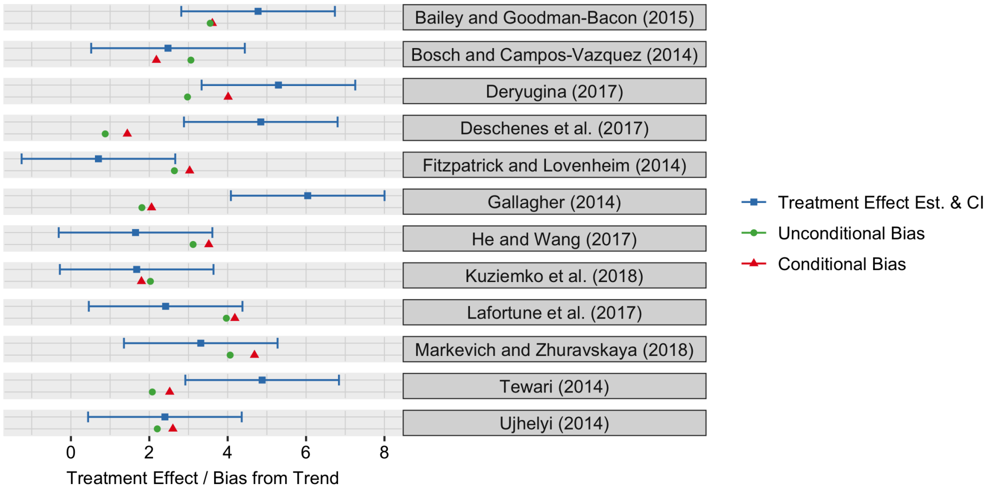


Figure 1. Original Estimates and Bias from Linear Trends for Which Pre-tests Have 80 Percent Power – Average Treatment Effect

Note: I calculate the linear trend against which conventional pre-tests would reject 80 percent of the time ($\gamma_{0.8}$). The red triangles show the bias that would result from such a trend conditional on passing the pre-test ($\mathbb{E}\left[\hat{\tau} - \tau_* \,|\, \hat{\beta}_{pre} \in B_{NIS}(\Sigma)\right]$); the green circles show the unconditional bias from such a trend ($\mathbb{E}\left[\hat{\tau} - \tau_*\right]$). As a benchmark, I plot in blue the original OLS estimates and 95% CIs from the paper. All values are normalized by the standard error of the estimated treatment effect and so the OLS treatment effect estimate is positive. The estimand is the average of the treatment effects in all periods after treatment began, $\tau_* = \bar{\tau}$.

	Unconditional			Conditional on Passing Pre-test		
	Slope of differential trend:					
	0	$\gamma_{0.5}$	$\gamma_{0.8}$	0	$\gamma_{0.5}$	$\gamma_{0.8}$
Bailey and Goodman-Bacon (2015)	0.05	0.61	0.94	0.05	0.62	0.95
Bosch and Campos-Vazquez (2014)	0.05	0.49	0.86	0.03	0.28	0.61
Deryugina (2017)	0.05	0.49	0.84	0.01	0.75	1.00
Deschenes et al. (2017)	0.05	0.09	0.14	0.03	0.09	0.25
Fitzpatrick and Lovenheim (2014)	0.05	0.41	0.75	0.05	0.50	0.87
Gallagher (2014)	0.05	0.19	0.44	0.04	0.22	0.54
He and Wang (2017)	0.05	0.54	0.88	0.05	0.63	0.95
Kuziemko et al. (2018)	0.05	0.28	0.53	0.04	0.20	0.42
Lafortune et al. (2017)	0.05	0.71	0.98	0.05	0.75	0.99
Markevich and Zhuravskaya (2018)	0.05	0.76	0.98	0.04	0.87	1.00
Tewari (2014)	0.05	0.20	0.55	0.04	0.25	0.72
Ujhelyi (2014)	0.05	0.29	0.60	0.04	0.36	0.76

Table 2—Null Rejection Probabilities for Nominal 5% Test of Average Treatment Effect Under Linear Trends Against Which Pre-tests Have 50 or 80% Power

Note: This table shows null rejection probabilities, i.e. the probability that the true parameter falls outside a nominal 95% confidence interval, under data-generating processes in which parallel trends holds (slope of differential trend = 0) or in which there are linear violations of parallel trends that conventional pre-tests would detect 50 or 80% of the time ($\gamma_{0.5}$ and $\gamma_{0.8}$). The first three columns show unconditional null rejection probabilities, whereas the latter three columns condition on passing the pre-test. The estimand is the average of the post-treatment causal effects, $\bar{\tau}$.

	Estimand:			
	τ_1		$\bar{\tau}$	
	Slope of differential trend:			
Paper	$\gamma_{0.5}$	$\gamma_{0.8}$	$\gamma_{0.5}$	$\gamma_{0.8}$
Bailey and Goodman-Bacon (2015)	51	56	1	2
Bosch and Campos-Vazquez (2014)	-29	-34	-25	-29
Deryugina (2017)	103	120	30	35
Deschenes et al. (2017)	88	119	48	64
Fitzpatrick and Lovenheim (2014)	25	30	12	15
Gallagher (2014)	57	62	11	14
He and Wang (2017)	29	34	11	13
Kuziemko et al. (2018)	-16	-20	-9	-11
Lafortune et al. (2017)	-9	-10	5	5
Markevich and Zhuravskaya (2018)	52	62	13	15
Tewari (2014)	90	102	19	21
Ujhelyi (2014)	51	59	15	18

Table 3—Percent Additional Bias Conditional on Passing Pre-test

Note: This table shows the additional bias from conditioning on none of the pre-period coefficients being statistically significant as a percentage of the unconditional bias, i.e. $100 \cdot$ (Conditional Bias $-$ Unconditional Bias)/(Unconditional Bias). Biases are calculated under linear violations of parallel trends with slopes $\gamma_{0.5}$ and $\gamma_{0.8}$, against which conventional pre-tests have 50 or 80% power. The estimand in the first two columns is the treatment effect in the first period (τ_1), and in the last two columns it is the average effect across all post-treatment periods ($\bar{\tau}$).